Decrypting soil microbial communities using metagenomic approaches

Tom Delmont

Thèse

D'octobre 2008 à Octobre 2011

Présentée devant

L'Ecole Centrale de Lyon

Pour obtenir le grade de

DOCTEUR

Préparée au sein de l'école doctorale

Electronique, électrotechnique, automatique de Lyon

Spécialité : Microbiologie environnementale

Par

Tom O. DELMONT

Description des communautés microbiennes du sol par une approche métagénomique

Soutenue le 19 décembre 2011 devant la commission d'examen

JURY

Robert Duran	Professeur - Université de Pau et des pays de l'Adour	**Rapporteur**
George A. Kowalchuk	Professeur - University of Amsterdam	**Rapporteur**
Penelope R. Hirsch	Docteur - Rothamsted research, Harpenden	**Examinateur**
Jed A. Fuhrman	Professeur - University of Southern California	**Examinateur**
Pascal Simonet	Directeur de recherche - Ecole centrale de Lyon	**Examinateur**
Timothy M. Vogel	Professeur – Université Claude Bernard Lyon 1	**Directeur de thèse**

Decrypting soil microbial communities and mining its genetic richness using metagenomic approaches

Tom O. DELMONT

Advisor: Pr. Timothy M. Vogel

Table of contents:

Introduction:

1. **Preamble (French part)**
2. **Preamble**
3. **Context and synthesis of the project (French part)**
4. **Context and History of the project**

Préambule (en français):

En tant que jeune scientifique et du haut tout relatif de trois années d'expérimentations et de réflexions personnelles et collectives que l'on nomme thèse, je me permets de présenter d'un point de vue très personnel un domaine de recherche dans lequel j'ai baigné et que je considère des plus prometteurs. En effet, par la force du nombre des êtres qu'il a pour but de décrire, ce domaine a le potentiel de fournir une quantité de données et découvertes des plus conséquentes. Les intimes l'appellent metagenomique car il transcende l'étude génomique de microorganismes. Dans une de ses définitions, il comprend le séquençage et l'analyse de fragments génétiques extraits directement de l'environnent et correspondant à un pool biologique.

Cette aventure singulière n'en est qu'à ses premières lignes et pourrait bel et bien altérer certains aspects de l'Humanité. Il ne faut cependant pas oublier l'histoire qui a permis sa création et qui en fait sa force, son héritage scientifique. Comme l'a dit Isaac Newton afin de justifier de l'importance de ses travaux de la plus humble des manières, il n'est possible d'user de cet outil de recherche que parce nous sommes juchés sur les épaules de géants. En effet, c'est grâce à un effort multi-centenaire de compréhension de la forme de vie la plus répandue mais aussi la plus discrète de notre planète qu'une annotation metagenomique toujours imparfaite mais probablement cohérente est possible. J'ai nommé les êtres unicellulaires : ils se font appeler bactéries, archées et eucaryotes inférieurs pour les spécialistes, microbes par le reste de la population, et pullulent à la surface de la troisième planète de notre système solaire.

Fort de connaissances cumulées au fil des ans, des pionniers sont en train de décrire de nouvelles formes de vies microbiennes présentes sur notre planète usant d'outils de plus en plus sensibles et sophistiqués. A l'instar des astrophysiciens fouillant les fin fonds de cette partie du cosmos que nous appelons univers mais a de toute autre échelles spatio-temporelles, nous pourrions les appeler explorateurs des temps modernes. Leur base de travail et source d'inspiration se nomme communément ADN, une molécule nécessaire à toutes les formes de vie que nous connaissons et qui aura permis 3.4 milliards d'années d'évolution continue sur Terre.

Cependant, et parce qu'ils sont pionniers d'un nouveaux domaine de recherche, leurs perceptions sont limitées, voire dans une certaine mesure erronées. En effet, la liste des étapes méthodologiques induisant des erreurs dans les résultats générés ne cesse de s'allonger. Une des difficultés majeures étant que la rigueur scientifique n'est pas suffisante pour avoir la certitude de leur exactitude, poussant logiquement certains scientifiques à se détourner de la metagenomique pour revenir à des recherches certes plus classiques mais aussi moins aléatoires. Ainsi, pour les pessimistes ou pragmatiques en fonction du point de vue, la métagénomique pourrait être définie comme une somme toujours croissante d'inexactitudes calculées scientifiquement, possédant pourtant une statistique robuste, et

de ce fait présentée d'une manière erronée comme un domaine incontournable de l'écologie microbienne contemporaine.

De ce fait, un fossé entre ceux qui doutent et ceux qui se basent entièrement sur les données générées est en train de se former au sein de la communauté scientifique. Pourtant, « douter de tout ou tout croire sont deux solutions également commodes, qui l'une et l'autre nous dispensent de réfléchir » (Henri Poincaré). De ce fait, la persévérance d'une partie de la communauté scientifique est nécessaire et devrait permettre de limiter si ce n'est de supprimer une grande partie de ces biais, ceux-là même qui empêchent actuellement de fournir une image précise des communautés microbiennes dans leur environnement naturel. En conséquence, les études métagénomiques effectuées depuis quelques années afin de décrire la vie microbienne présente dans des habitats naturels bien particuliers ou des plus communs doivent être vues avec une humilité toute particulière, et ce malgré la présence de lacunes et incohérences scientifiques dans certains de ces travaux.

Dans une moindre mesure, les études qui seront décrites dans ce manuscrit de thèse représentent une pierre ajoutée à un édifice toujours grandissant. Cet édifice symbolise les données métagénomiques permettant de décrire le plus précisément possible ces êtres unicellulaires *in situ*, mais aussi de comparer les différents styles de vies microbiennes présentes sur Terre. Tout comme ses prédécesseurs, ces études ne sont pas dénuées d'erreurs. Cependant, un effort tout particulier a été fait pour limiter ces limites méthodologiques pour le mieux afin d'être plus confortable dans l'analyse des données générées. Je vous laisse juges de ces travaux. Leur critique ne pourra qu'améliorer ma vision de cette science.

Je me permet de finir ce préambule par une autre citation de Henri Poincaré qui pourrait stimuler une certaine réflexion sur un aspect insidieux de la metagenomique: « On fait la science avec des faits, comme on fait une maison avec des pierres : mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison ». En conclusion, la metagenomique ne doit pas être considérée comme une science en soit, mais doit être utilisée à bon escient afin d'extraire des données qui pourraient éventuellement faire avancer notre connaissance de cette majorité invisible que l'on dénomme microorganismes.

Preambule:

As a young scientist and from the lofty pinnacle of three years of experiments and personal and collective reflections that constitute the thesis, I would like to present from a personal point of view a research area in which I have been immersed and which I consider to be extremely promising. In fact, by the number of living organisms that it aims to describe, this field has the potential to provide substantial amounts of data and discoveries. Specialists in this area of research call it “metagenomics” because it transcends the genomic study of microorganisms. From one of its definitions, it includes the sequencing and analysis of genetic fragments directly extracted from the environment.

This singular adventure is in its infancy and might change certain perceptions of humanity. It is important, however, to bear in mind the history behind it that gives it its strength, its scientific heritage. As Isaac Newton said to justify the importance of his work in the most apparently insignificant areas, it is possible to use this research tool because we are standing on the shoulders of giants. It is due to research over several centuries to understand the most widespread but also most subtle form of life on earth that we can achieve a metagenomic annotation, an annotation that is not yet perfected but which is feasible. I am referring to unicellular organisms: they are called Bacteria, Archaea or inferior Eukaryotes by specialists and microbes by the rest of the population and they populate at the surface of the third planet of our solar system.

With knowledge accumulated over the years, some pioneers are currently describing new forms of microbial life present on our planet, using more and more sophisticated and sensitive tools. In the manner of astrophysicians who delve into the most distant part of the cosmos that we call the universe, but with totally different space and time scales, we can call them modern-days explorers. The basis and driving force of their work is commonly called DNA, a molecule essential to all known forms of life, vital to the 3 to 4 billion years of continuous evolution on earth.

Precisely because they are pioneers of a new research field, however, their perceptions are limited and, to a certain extent, incorrect. Indeed, the list of methodological steps inducing errors in the generated data constantly becomes longer and longer. One of the main difficulties is that the scientific rigor is not adequate to allow us to be certain of the exactness of those steps, and this causes some scientists to revert to more traditional, less hazardous research. So, for the pessimist or pragmatic scientist, depending on one’s point of view, metagenomics could be defined as a constantly increasing sum of scientifically calculated inexactitudes, which nonetheless provide robust statistics. For this group of the scientific community, metagenomics should not be presented as an essential part of contemporary microbial ecology.

A gap between those who have doubts and the others who rely exclusively on metagenomic data is developing within the scientific community. However, in the words of Henri Poincarré,

“To doubt everything or to believe everything are two equally convenient solutions; both dispense with the necessity of reflection.” The perseverance of a part of the scientific community, is therefore necessary and should limit, or better, eliminate an important part of these biases, the ones that actually prevent us from supplying an accurate and sensitive picture of microbial communities in their natural environment. Consequently, metagenomic studies performed in recent years that aimed to describe microbial life, in particular natural habitats or common habitats, have to be seen with a certain humility, with the presence of gaps and scientific incoherencies in some of these works.

To a lesser extent, studies that are described in this PhD thesis are adding a stone to a building that is constantly becoming larger and larger. This building symbolizes the metagenomic data that allow us to describe in the best possible way these unicellular organisms in situ, and additionally allows us to compare the different microbial ways of life on earth. Like its predecessors, this study is not free from difficulties and incertitudes. Nonetheless, special care has been taken to limit some well defined methodological inexactitudes in order to be more comfortable with the data analysis.. Constructive criticism can only improve my perception of this science.

I would like to finish this introduction with another comment of Henri Poincarré citation that may stimulate a measure of thought on a difficult aspect of metagenomics: « Science is an accumulation of facts, in the same way that a house is an accumulation of stones. But a collection of facts is no more a science than a heap of stones is a house. » To conclude, metagenomic should never be considered as a science on its own, but should be used punctiliously to extract data that could eventually improve our knowledge of this invisible majority that we call microorganisms.

Contexte et synthèse du projet de recherche

Introduction :

Le triplet de concepts de « thèse-antithèse-synthèse » a pour but de représenter le processus de transformation de l'abstrait vers le concret en deux étapes distinctes (Hegel, G. W. F., Phénoménologie de l'Esprit). En effet, la « thèse » étant définie comme une unité abstraite, l'« antithèse » représente quant à elle la transformation de cette unité en une multiplicité concrète. Enfin, la « synthèse » a pour objectif de transformer la multiplicité concrète en une unité concrète.

De par mon domaine d'étude et d'une manière générale, il est possible de définir la « thèse » comme la perception que détient l'humanité, par le biais de la communauté scientifique, des communautés microbiennes pullulant à la surface de la troisième planète du système solaire. Cette « thèse » est hautement abstraite puisqu'elle varie en fonction des scientifiques, des idées qui leurs passent par la tête, du temps, du transfert d'information entre scientifiques et la population aussi, et ne peut donc être quantifiée. L'« antithèse » serait donc par définition l'ensemble des travaux de recherches, expérimentations, découvertes et hypothèses générées au fil des ans sur ces communautés microbiennes : cette multiplicité concrète représentant la perception que possède l'homme sur la vie microscopique. Face à la complexité et la richesse scientifique de cette antithèse, il n'est pas humainement possible (il l'était autrefois, mais ne l'est plus) de relier les différentes facettes de cette perception dans leur globalité, et donc d'en faire une unique « synthèse ». En conséquence de quoi, les efforts de synthèse proposés par la communauté scientifique ne survolent dorénavant qu'un domaine bien précis, une facette ne pouvant représenter l'ensemble.

Cependant, dans le cadre de mon doctorat et de ces trois années de recherche, je me propose de simplifier l'équation et de définir la « thèse » comme ma perception personnelle des communautés microbiennes, et l'« antithèse » comme l'ensemble des travaux de recherches, expérimentations, découvertes et hypothèses générées dans l'unique cadre de ce doctorat. La majeure fraction des données produites durant ces années de recherche est accessible au lecteur dans les différents chapitres de ce manuscrit de thèse. Ces multiplicités concrètes représentent ma perception de cette vie microscopique, en faire la « synthèse » à pour but d'aider le lecteur à appréhender à la fois mes connaissances et ma vision d'un domaine de recherche précis (communément appelée métagénomique), mais aussi et surtout les avancées concrètes qui m'ont été données de réaliser dans le cadre de mes recherches durant ces trois années de doctorat.

Contexte du projet de recherche :

Après cinq années d'études académiques au sein de l'université de Pau et des pays de l'Adour (2003-2008) et mon diplôme de Master recherche (spécialité microbiologie et biotechnologies) en poche, j'ai eu l'opportunité d'intégrer l'équipe de génomique microbienne environnementale, hébergée par le laboratoire Ampère (UMR 5005) au sein de l'Ecole Centrale de Lyon, dans le cadre d'un doctorat dirigé par le professeur Timothy M. Vogel.

Ce doctorat, financé par la région Rhône-Alpes, avait pour principal objectif de soutenir expérimentalement un projet de recherche financé par l'agence nationale de recherche (ANR) et édifié par Timothy M. Vogel et Pascal Simonet. Ce projet, appelé Metasoil (http://metasoil.univ-lyon1.fr/), avait (avait ou a ?) pour objectif scientifique d'étudier les communautés microbiennes présentes dans un sol de référence. Cependant, le rôle de ce projet n'était pas seulement scientifique puisqu'il s'intégrait dans un calendrier de recherche international bien particulier.

En effet, durant les années 2006-2008, la communauté scientifique spécialisée dans la microbiologie de l'environnement sol a décidé de se focaliser sur l'étude d'un unique site afin de palier aux problèmes majeurs de ce domaine de recherche. Ces problèmes sont majoritairement dus aux limites de culture (moins de 1% des espèces sont actuellement cultivables), à la complexité microbiologique de cet environnement et enfin à ses caractéristiques physico-chimiques très hétérogènes qui créent des micro-niches de communautés microbiennes. De ce fait, cet environnement reste une boite noire pour les microbiologistes (par exemple, nous n'avons à l'heure actuelle, aucune idée du nombre d'espèces présentes dans un gramme de sol, les estimations allant de 10^4 à 10^7), et la stratégie consistant à soutenir les recherches en un unique site expérimental avait pour but honorable de stimuler la perception scientifique de cet environnement.

En parallèle de ces réflexions de la communauté scientifique, de nouvelles technologies de séquençage massif d'ADN ont émergé. Ces outils ont représenté une réelle révolution en microbiologie, permettant la génération de millions de séquences de quelques centaines de bases à partir d'échantillons d'ADN extraits directement de l'environnement. Ces séquences correspondent donc à des fragments de structures génétiques provenant de nombreuses espèces microbiennes. Leur annotation à partir de bases de données de références provenant des connaissances cumulées au fil des siècles à partir de microorganismes cultivées en laboratoire permet entre autres avantages de percevoir le potentiel fonctionnel d'un environnement.

Cette approche d'étude de communautés microbiennes plus ou moins complexes par un séquençage direct d'ADN (sans étape de culture) se nomme la métagénomique. Ce domaine a été défini en 1998 mais n'a vraiment émergé qu'àa partir de 2004 avec les études pionnières de séquençage des océans et d'un environnent extrême, respectivement. Ainsi,

les données générées à partir d'un échantillon d'ADN sont appelées des métagénomiques, même si, paradoxalement, ils ne représentent, en général, qu'une infime partie de la micro-biodiversité de l'environnement qu'ils représentent. La taille de ces métagénomiques, et donc, la fraction de la diversité nucléique qu'ils représentent, est directement corrélée aux technologies de séquençage émergentes et à leur sensibilité. Ces technologies évoluent actuellement à une vitesse étonnante, rendant la métagénomique un domaine en perpétuelle évolution et offrant des opportunités uniques à ceux qui savent les saisir.

De ce fait, fort de nombreuses expériences dans le domaine, et intégrant les réflexions de la communauté scientifique spécialisée dans la microbiologie du sol et les technologies émergentes de séquençage d'ADN, Timothy M. Vogel et Pascal Simonet ont élaboré un project ANR d'envergure, Metasoil. Ce projet avait pour but d'aider à la formation d'un consortium international visant à étudier, par une approche métagénomique, les communautés microbiennes présentes dans un sol de référence, ce sol restant à définir. Le projet Metasoil avait donc pour objectif de fournir une quantité considérable de séquences d'ADN à partir de l'environnement sol (jusque là très peu séquencé) usant de nouvelles technologies de séquençage (collaboration avec le centre national de séquençage, le Genoscope), mais aussi de construire une banque fosmidique considérable (2 millions de clones, inserts de 40 000 bases, collaboration avec Libragen) accessible à tous.

Et c'est dans ce cadre que j'ai accepté de m'intégrer dans ce projet de caractérisation de la micro-biodiversité d'un sol de référence afin de repousser les frontières de la science dans ce domaine bien précis. Trois mois seulement après le début de mon doctorat, une conférence a eu lieu à Lyon (Metasted, 13-14 décembre 2008, http://www.ampere-lab.fr/spip.php?article308) afin de créer officiellement une entité scientifique d'étude d'un sol de référence. Ce congrès a regroupé 80 scientifiques provenant de 14 pays, et a permis l'émergence du consortium international nommé Terragenome. Le sol de référence a quant à lui aussi été défini. Il s'agit d'une prairie témoin (vierge d'expérimentations humaines) localisée dans la plus vieille station expérimentale au monde (Park Grass, Rothamsted, Angleterre, Royaume-Uni). L'objectif de Terragenome était à la fois simple et très ambitieux : le séquençage et assemblage intégral du métagénome de ce sol de référence. Les détails sont présentés dan le chapitre 2, section 2 (« Terragenome : a consortium for the sequencing of a soil metagenome »).

Ainsi a vraiment commencé mon doctorat, dont les objectifs principaux ont été d'accéder au mieux à la diversité nucléique de ce sol de référence (tache rendu difficile par la présence de nombreuses limites physiques et méthodologiques), puis de caractériser les communautés microbiennes présentes usant des dernières technologies de séquençage et de plateformes d'annotation performantes. Ces objectifs peuvent paraître simples, mais comportent de nombreuses difficultés, dont je n'avais bien entendu aucune idée à cette époque. Eviter les pièges et avancer dans l'étude de ces communautés reflètent les trois années qui ont suivi.

Leur synthèse a pour but d'aider le lecteur à appréhender ma perception de ce domaine de recherche, son passé mais aussi ses futurs prometteurs possibles.

Synthèse :

1. **Elaboration d'une stratégie visant à accéder au mieux à la diversité nucléique du sol**

Ma première année de doctorat a été principalement focalisée sur l'étude de méthodologies existantes permettant d'extraire et de purifier des échantillons d'ADN représentant les communautés microbiennes d'un sol dans leur globalité. Après avoir fait le tour des articles scientifiques publiés dans le domaine, il m'est vite apparu que la majorité des équipes de recherches avaient leur propre protocole pour accéder à la diversité nucléique du sol. Ne voyant pas comment en choisir un de manière objective (le sol représentant une boite noire), j'ai décidé d'en tester le plus possible, puis d'étudier les communautés microbiennes au travers de chacun d'entre eux.

L'objectif premier de mon doctorat était d'accéder au mieux à la diversité nucléique de ce soil. Usant tout d'abord d'une technique sommaire permettant de différencier la structure de communautés basée sur une région particulière des génomes de Bactéries et Archae (l'analyse de l'espace intergénique ribosomal, produisant des électrophérogrammes nommés profils RISA et pouvant être comparés les uns aux autres), j'ai rapidement observé des différences considérables entre protocoles d'extraction/purification d'ADN à partir d'un même échantillon de sol. Cependant, la plupart de ces protocoles étaient très reproductibles. Il fallait donc garder en tête que ce n'était pas la représentation des communautés qui était reproductible mais bel et bien les biais liés à chaque protocole.

J'ai ensuite utilisé une technologie plus avancée, nommée puce taxonomique, afin de détecter et quantifier (même si la technologie n'est que semi-quantitative) des empreintes de diversité basées sur la séquence nucléique de 16S rRNA (un gène largement utilisé par la communauté scientifique pour définir et différencier les espèces bactériennes et archaebactériennes). Bien entendu, j'ai là aussi observé d'importantes différences entre protocoles. Cependant, chaque protocole permettait apparemment d'accéder à de nouvelles diversités. Donc, lorsqu'on utilise deux protocoles par exemple, on accède à plus de diversité que lorsqu'on n'utilise qu'une seule approche. Cette réflexion a été la base de la plupart des expérimentations qui ont suivi. En effet, à l'opposé de pratiquement toutes les études passées et actuelles qui sont basées sur l'utilisation d'un protocole hautement standardisé (et donc reproductible) afin d'étudier et surtout de comparer des échantillons, je me suis atteler à la création d'une approche globale visant à varier au maximum les approches pour stimuler au mieux la détection d'espèces dans ce sol de référence.

J'ai testé de nombreuses variables afin de fractionner la diversité de ce sol. Certains n'ont pas fonctionné. C'est le cas notamment de la séparation physique du sol en fonction de la

taille de ses particules. J'ai utilisé une colonne d'eau, puis déposé une centaine de grammes de sol, et attendu la sédimentation de toutes les particules. Les grosses se sont déposées très rapidement, les dernières après quelques heures en une fine couche blanche. J'ai enfin réussi à récupérer des échantillons de chacune des couches, puis à analyser les différents échantillons d'ADN extraits usant d'un même protocole. Aucune différence n'a pu être observée sur les profiles RISA, l'idée était pourtant bonne.

Cependant, d'autres approches ont été plus fructueuses. C'est bien sûr le cas de la lyse cellulaire (la base des biais induits lors des protocoles d'extraction d'ADN). Il était donc possible de jouer avec la stringence de cette lyse afin d'accéder a des diversités différentes. L'échantillonnage du sol permettait aussi d'accéder, dans une moindre mesure, à d'autres diversités (par exemple en échantillonnant le sol à différentes profondeurs).

Cependant, il était possible d'aller bien plus loin afin de fractionner artificiellement cette diversité. En effet, j'ai utilisé une technique qui permet de concentrer des cellules microbiennes dans un anneau basé sur la densité des cellules. Cette technique est habituellement utilisée pour concentrer les cellules et les récupérer dans une unique fraction. Cependant, j'ai essayé de récupérer les cellules présentes au dessus et en dessous de cet anneau cellulaire après stabilisation de la colonne en fonction de la densité cellulaire. En j'ai pu accéder aux cellules dans chaque fraction, même si leur nombre était bien plus faible que dans l'anneau. Et la diversité présente était très différente de celle de l'anneau. En fractionnant la communauté totale en fonction du paramètre de densité cellulaire, il était possible d'accéder à des espèces normalement trop faiblement représentées pour être détectées.

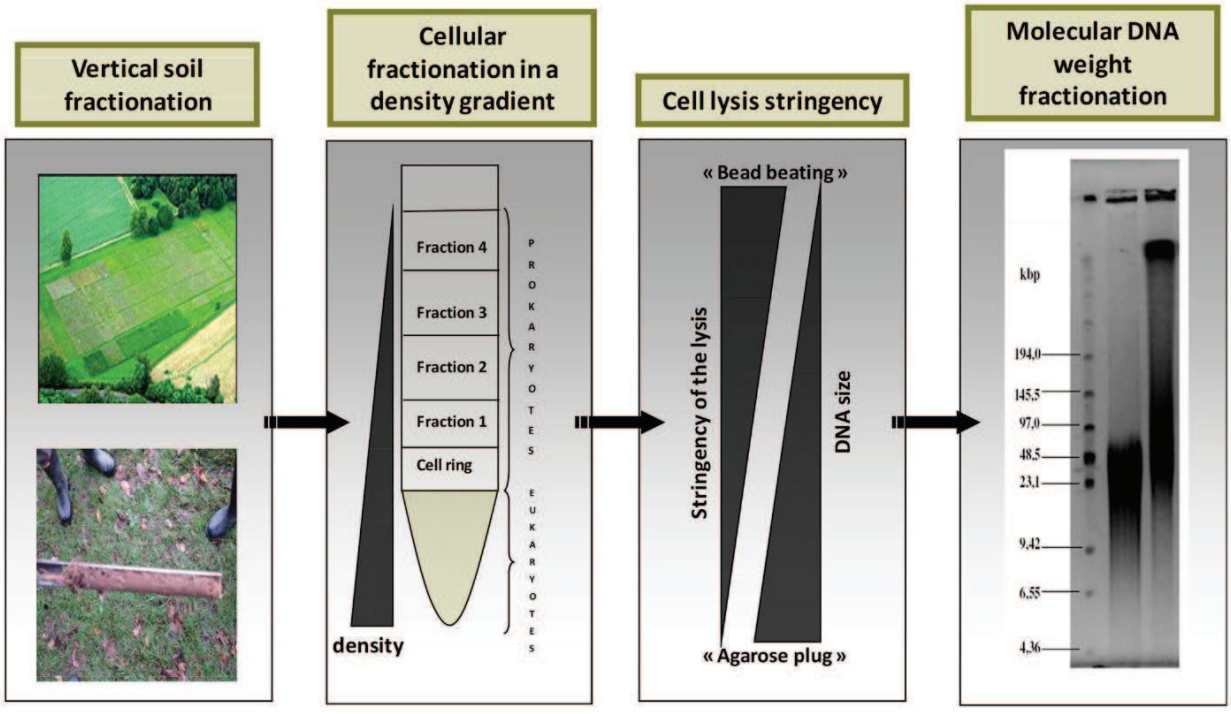


Figure 1 : Schémas représentant la stratégie de fractionnement de la diversité d'un sol.

Finalement, j'ai aussi fractionné non pas les cellules mais l'ADN en fonction de la taille des fragments extraits et purifiés utilisant une électrophorèse (l'ADN étant chargé). Assez étonnamment, la diversité nucléique pouvait aussi être fractionnée en utilisant le paramètre correspondant à taille de l'ADN (une des hypothèses étant que le taux de GC par exemple avait une influence sur la solidité des fragments d'ADN ; ainsi les longs fragments auraient une haute proportion en GC, qui proviennent de communautés bien particulières, les Actinobacteries par exemple).

Ainsi, j'ai construit une stratégie de fractionnement de la diversité du sol de Rothamsted en me basant sur l'échantillonnage, la séparation des cellules en fonction de leur densité, la stringence de la lyse cellulaire, et enfin la taille de l'ADN extrait. La figure 1 est un schéma représentant cette approche. J'ai ensuite validé cette stratégie avec la puce taxonomique. Utilisant la meilleure approche, environ 40% des espèces possiblement détectées avec cette technologie l'ont été. En utilisant 15 échantillons représentant les diversités les plus originales d'un même sol (basées sur les profils RISA à partir du sol de référence de Terragenome), 99.76% des espèces sont détectées. Ainsi, cette approche permet de stimuler considérablement le nombre d'espèces détectées à partir d'un même sol. Une autre conclusion importante étant que toutes les estimations de diversité du sol, s'étant basées sur une seule approche, sont largement sous estimées. Cette étude est décrite en détail dans le chapitre 2, section 3 (« Accessing the soil metagenome for studies of microbial diversity »).

2. Caractérisation des communautés microbiennes du sol de Rothamsted

Mon second objectif étant de caractériser au mieux les communautés microbiennes du sol de Rothamsted par une approche métagénomique, j'ai ensuite construis une expérimentation permettant de fournir une image globale de ces communautés, intégrant à la fois les fluctuations naturelles (comme la profondeur du sol, et les effets saisonniers) et méthodologiques (la stringence de lyse cellulaire). Deux profondeurs (0-10cm et 1-20cm), deux saisons (Mars et Juillet), deux années (2009 et 2010) ainsi que 6 protocoles d'extraction d'ADN ont été utilisés pour générer 13 jeux de données métagénomiques. Un million de séquences de 350 bases ont été ainsi générées dans chacun de ces jeux de données. J'ai ensuite utilisé une plateforme d'annotation des séquences afin de définir le potentiel fonctionnel de ces communautés. Bien sûr, pour chaque fonction détectée, il m'était possible de définir une variation correspondant à la fois aux fluctuations naturelles (comme généralement fait dans les études métagénomiques) et méthodologiques (jamais fait avant).

Plusieurs points importants sont ressortis de cette étude. Tout d'abord, nous n'avons observé que très peu de variations de profondeur et de saison, ni même entre les deux années. Par contre, la stringence de lyse à produit bien plus de variations, générant d'important problèmes pour les études de comparaison métagénomiques intra-

environnementales. En effet, les fluctuations méthodologiques apparaissent comme étant plus importantes que les variations naturelles. Nous avons ainsi observé plus de différences entre deux échantillons correspondant au sol de Rothamsted qu'entre ce sol et un sol de Puerto Rico dont le métagénome était accessible.

Ensuite, nous nous sommes rendu compte que la majorité des espèces du sol (plus de 99%) n'était pas présente dans les bases de données permettant l'annotation de métagénomiques. Ainsi, un effort considérable doit être fait pour accéder aux génomes des communautés du sol, et notamment les prédominantes.

Nous avons aussi essayé d'assembler ces jeux de donner afin de reconstruire des structures génétiques complexes. Cependant, les efforts d'assemblages n'ont pas été fructueux (le plus long contig étant de 13 000 bases en utilisant les 13 jeux de données simultanément).

Finalement, nous avons comparé le potentiel fonctionnel des communautés du sol (de Rothamsted et d'autres) à des métagénomiques correspondant à d'autres environnements, et ce, afin de définir les particularités fonctionnelles des communautés présentes dans ces différents habitats. La figure 2 représente certaines de ces particularités, sous la forme d'une analyse en composante principale basée sur la distribution relative de sous-systèmes fonctionnelles des métagénomiques sélectionnés.

Figure 2 : analyse en composante principale basée sur la distribution relative de systèmes fonctionnels des métagénomiques correspondant à trois environnements distincts : océans, sols et fèces humains. .

Cette étude de caractérisation du sol de Rothamsted par une approche métagénomique est décrite plus en détail dans le chapitre 2, section 4 (« Structure, fluctuation and magnitude of a grassland soil metagenome »).

Une des conclusions de cette étude a aussi été que le séquençage massif d'ADN extrait directement du sol n'est pas suffisante pour accéder à la diversité présente, ni pour reconstruire des structures complexes par des approches d'assemblage de séquences générées. Or l'objectif de Terragenome était clair : le séquençage et assemblage total d'un métagénome de sol.

Ainsi, d'autres stratégies doivent émerger pour stimuler l'accessibilité d'un métagénome de sol, le séquencer puis l'assembler.

3. Etude théorique de la richesse de diversité microbienne du sol

Il est actuellement très difficile d'estimer la diversité d'un sol, et de savoir quelle fraction est commune à tous les sols. Certaines théories considèrent même que tout pourrait être partout, et que l'environnement ne ferait que sélectionner la distribution des espèces, et donc celles qui seront représentées dans les jeux de données. Connaître l'amplitude d'un sol est crucial, notamment quand l'objectif est d'étudier un métagénome dans sa globalité.

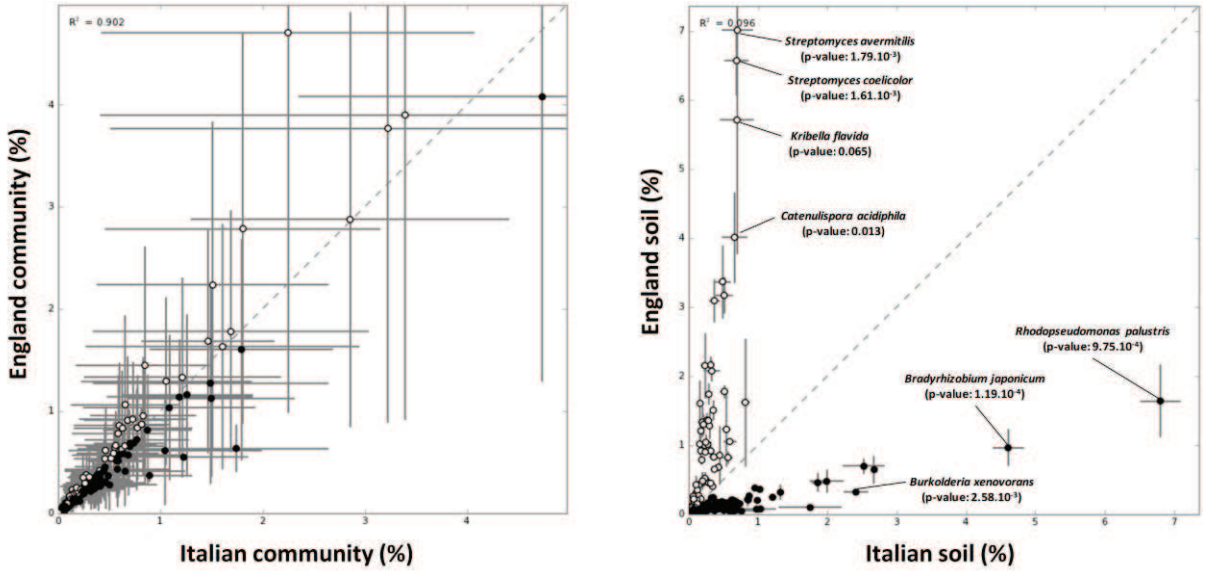


Figure 3 : Comparaison de la distribution relative de sous-systèmes taxonomiques entre groupes de métagénomiques, basée sur la communauté inoculée, ou le sol stérilisé utilisé lors de l'inoculation.

Afin d'étudier le core microbien (ce qui est commun) de deux sols distincts (celui de Rothamsted et un sol forestier localisé en Italie), des expériences d'inoculations de sols stérilisées ont été réalisées utilisant les deux communautés microbiennes. L'idée était de confronter l'effet de la communauté et des caractéristiques physico-chimiques de chaque sol lors de l'inoculation des deux sols stérilisées.

Apres deux mois d'inoculation, les deux communautés sont apparues très similaires lorsque inoculées dans le même sol, la structure étant très différente en fonction du sol utilisé pour l'inoculation (figure 3). En d'autres termes, les caractéristiques physico-chimiques apparaissent comme étant le facteur définissant la structure de la communauté. Ainsi, les deux communautés, d'après les résultats obtenus, ont probablement un core microbien considérable. Même si les données générées ne peuvent rien conclure définitivement, il est possible que la plupart des sols possèdent la même diversité, et une distribution différente due aux caractéristiques physico-chimiques qui leurs sont propres. Cette étude est décrite plus en détail dans le chapitre 3, section 1 (« Following the colonization effect of distinct communities in sterilized environments : a soil microbial richness study »).

Ainsi, le consortium Terragenome prendrait tout son sens, car se focaliser sur un unique sol permettrait d'accéder à presque, si ce n'est la totalité, des espèces natives du sol. Cependant, des stratégies doivent émerger pour accéder a la vaste majorité d'espèces faiblement représentées, et donc non accessible usant d'outils métagénomiques classiques.

4. Atteindre l'objectif de Terragenome : séquencer et assembler un métagénome de sol

Le séquençage massif d'ADN extrait directement du sol en variant les méthodes est apparu comme étant une bonne approche pour caractériser les communautés microbiennes prédominantes, et pour définir le potentiel fonctionnel de l'environnement étudié. Cependant, la distribution des espèces est très inégale dans le sol, et ainsi la diversité génétique de la vaste majorité de microorganismes, faiblement représentés, n'a aucune chance d'être représentée dans les jeux de données générés.

Fractionner les cellules en fonction de leur densité est apparu comme étant une très bonne approche pour accéder à certaines espèces faiblement représentées. Malheureusement, le rendement d'ADN s'est avéré trop faible pour convenir aux appétits des technologies de séquençage (plusieurs microgrammes d'ADN purifié). Ne souhaitant pas user de protocoles permettant d'amplifier une faible quantité d'ADN métagénomique (une étape qui modifie quantitativement le jeu de données généré), je me suis attelé à une toute autre approche, utilisée depuis très longtemps par les microbiologistes, mais pas encore lors d'études métagénomiques. J'ai nommé les microcosmes, qui permettent de modifier de manière contrôlée la structure de communautés microbiennes.

Ainsi il est potentiellement réalisable de modifier une communauté n fois usant de n conditions particulières, chacune stimulant une partie des microorganismes normalement faiblement représentés, puis de générer n métagénomiques (en plusieurs exemplaires lorsqu'on intègre des réplicas dans l'étude).

Théoriquement, lorsque n se rapproche de l'infini, la proportion du métagénome du sol étudié par cette approche tant vers 1. Ainsi il serait possible usant de cette stratégie

(schématisée dans la figure 4) d'atteindre l'objectif de Terragenome, et ainsi de séquencer et d'assembler un métagénome de sol.

A ce moment là, il sera possible de connaître en détail les espèces présentes. Il sera aussi possible d'utiliser cette mine d'or génétique afin de produire des biomolécules d'intérêt. Les opportunités scientifiques et appliquées seront considérables.

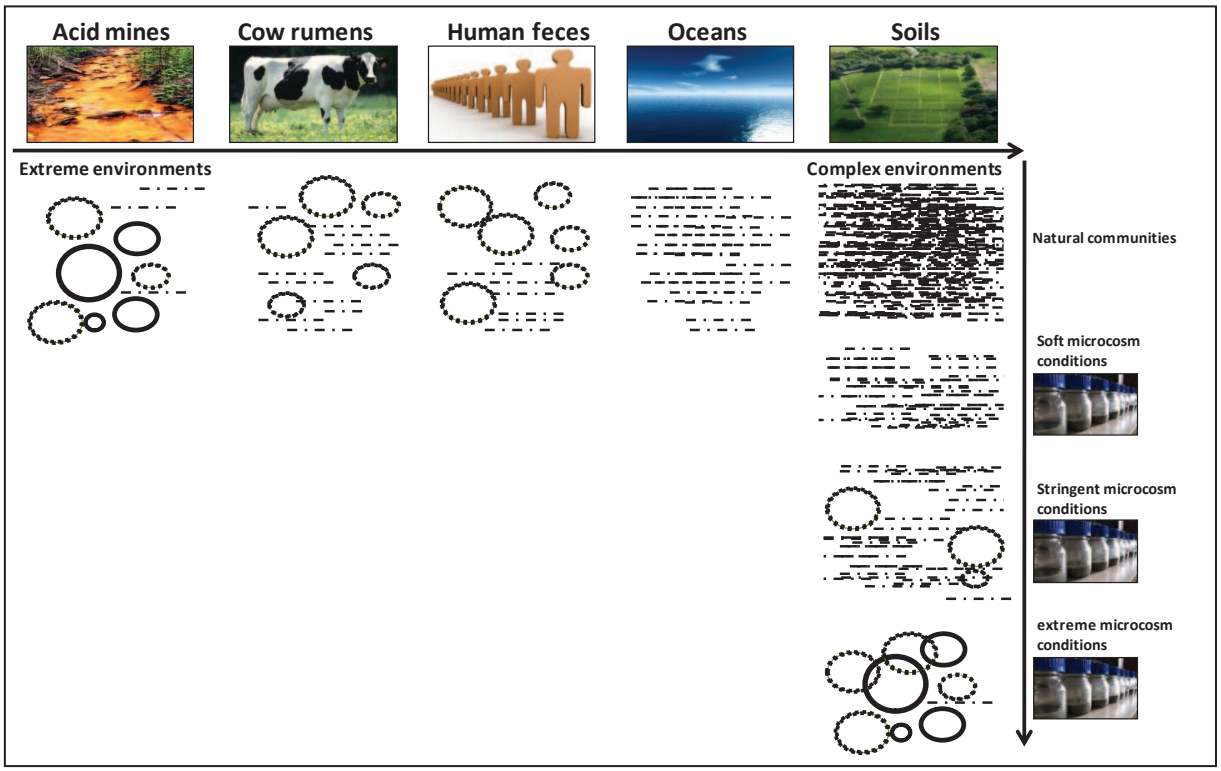


Figure 4 : Schémas représentant la stratégie expérimentale visant à assembler un métagénome de sol.

Afin de tester cette stratégie, j'ai mis au point une expérience en microcosmes, utilisant une condition contrôle et dix conditions particulières (par exemple un apport considérable de mercure, de métaux lourds, de sel ou d'éthanol respectivement) dont le but était d'accéder à de nouvelles diversités fonctionnelles et taxonomiques.

Après quatre mois d'incubation, les communautés ont étaient séquencées en *duplicata* biologiques pour chacune de ces conditions, puis les jeux de données annotés et assemblés. Comme attendue, certaines de ces conditions ont permis d'accéder à des diversités très inhabituelles, allant jusqu'à pratiquement représenter une culture pure *in situ* dans un cas (une première). Ainsi, dans le cas de diversités très particulières, l'assemblage des jeux de données à été très efficace. La majorité des séquences ont été assemblées, et des contigs de plusieurs centaines de milliers de bases ont été générés. Certains des génomes enrichis et séquencés représentaient des espèces déjà séquencées, d'autres des espèces connues pour être récalcitrantes à la culture. Cette étude est décrite plus en détails dans le chapitre 3, section 2 (« Stressing complex microbial communities for metagenomic discoveries : one designed evenness at the time »).

Les opportunités de cette stratégie sont considérables. Il est dorénavant possible d'accéder à la richesse génétique de la majorité d'espèces faiblement représentées usant de modifications environnementales contrôlées. La communauté scientifique spécialisée dans la microbiologie du sol devrait maintenant appliquer une large variété de conditions à différent sols, puis générer des jeux de données permettant la reconstruction de structures génétiques complexes (comme les génomes, bacteriophages ou plasmides).

5. Comparaisons métagénomiques inter-environnementales :

En parallèle de mes travaux de recherche sur le sol de Rothamsted, j'ai créé une compilation de métagénomiques accessibles dans les plateformes d'annotations et représentant 15 environnements (dont le sol largement représenté par les données générées dans le cadre de Metasoil), puis j'ai comparé le potentiel fonctionnel de ces jeux de données regroupés par environnement.

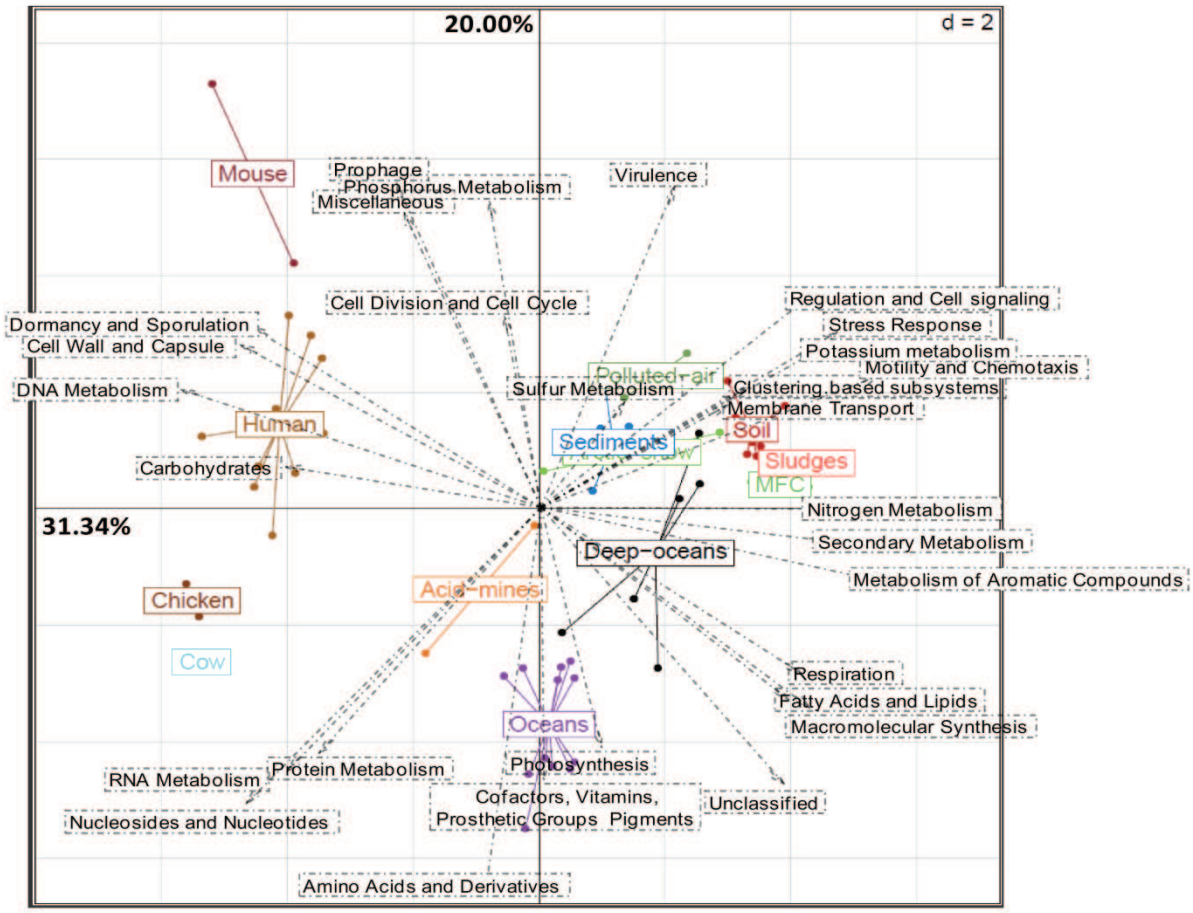


Figure 5 : Analyse en composante principale basée sur la distribution relative de systèmes fonctionnels des métagénomiques correspondant à 15 environnements distincts.

Cette approche a permis de définir les particularités de chacun de ces environnements utilisant les autres comme référence. Alors que la grande majorité des métagénomiques

sélectionnés étaient déjà publiée, un grand nombre de résultats non présentés dans les articles (car non basés sur une telle approche) ont été générées. Un exemple général est présenté dans la figure 5. Cependant, je me suis efforcé de définir les distributions fonctionnelles inhabituelles environnement par environnement. Cette étude, possédant aussi une large introduction présentant toute l'histoire de la métagénomique, est présentée dans le chapitre 1, section 2 (« Decrypting global metagenomic comparisons of the microbial word »).

Même si elle a été réalisée en amont, cette étude s'inscrit dans le cadre du projet microbien mondial (Earth microbiome project, http://www.earthmicrobiome.org/) et pourrait être définie comme un pilote de ce projet ambitieux.

De plus, cette approche de comparaison multiple d'environnements représente une opportunité considérable pour tous les microbiologistes, qui n'ont pas forcement accès aux plateformes de séquençage mais qui peuvent accéder aux nombreux métagénomiques accessibles dans les plateformes d'assemblage. Ils peuvent utiliser ces données pour étudier leur environnement de prédilection, ou encore pour confirmer des hypothèses particulières (par exemple la corrélation d'une fonction et d'un groupe bactérien particulier, qui peut être positive dans certains environnements, négative ou neutre dans d'autres). Ces intérêts de la metagenomique pour la communauté scientifique ont été décrits dans la conclusion, section 1 (« Metagenomic mining for microbiologists »).

6. Perspectives de la métagénomique

Durant l'analyse de mes données de séquençage, et utilisant aussi l'expérience de ma première année d'étude, très méthodologique, j'ai défini un certain nombre de limites et biais qui potentiellement peuvent influencer les conclusions scientifiques de toute étude métagénomique.

Notamment, la stringence de lyse cellulaire et d'annotation des séquences apparaissent comme étant deux étapes cruciales, à manipuler avec beaucoup d'attention. Le problème étant que chaque approche propose une diversité différente pour représenter une boite noire. De ce fait il n'est pas possible de choisir objectivement un protocole d'extraction, une stringence d'annotation. Pourtant, toutes les études en cours sont basées sur l'utilisation d'une seule méthode, très standardisée encore une fois afin de limiter au mieux les fluctuations méthodologiques. Cependant, je considère plus judicieux de ne pas limiter, mais bel et bien maximiser les fluctuations méthodologiques afin de définir une image globale d'un environnement, même si elle est moins sensible. Il serait ensuite possible de comparer ce groupe de jeux de données à d'autres, représentant quant à eux d'autres environnements, caractérisés de la même manière. J'ai proposé de nommer ces fluctuations méthodologiques des « flexicats », qui au contraire des réplicats actuellement largement utilisés, permettent de maximiser les fluctuations de notre perception d'un metagenome,

basée sur les limites de nos outils actuels. Un exemple de caractérisation de métagénome utilisant des flexicats est présenté dans la figure 6.

Basée sur les comparaisons de trois environnements, les flexicats n'empêchent pas les comparaisons inter-environnementales, même si elles rendent encore plus difficile les comparaisons intra-environnementales. Cette étude est présentée plus en détail dans le chapitre 4, section 1 (« How to avoid pitfalls in the metagenomic jungle »). De plus, le concept de flexicat est proposé sous la forme d'un commentaire dans la conclusion, section 2 (« Decrypting microbial communities and performing global comparisons in the 'omic era : replicates vs flexicates »).

L'objectif principal de ce commentaire étant de stimuler un débat au sein de la communauté scientifique sur les alternatives permettant de limiter les biais liés à la métagénomique.

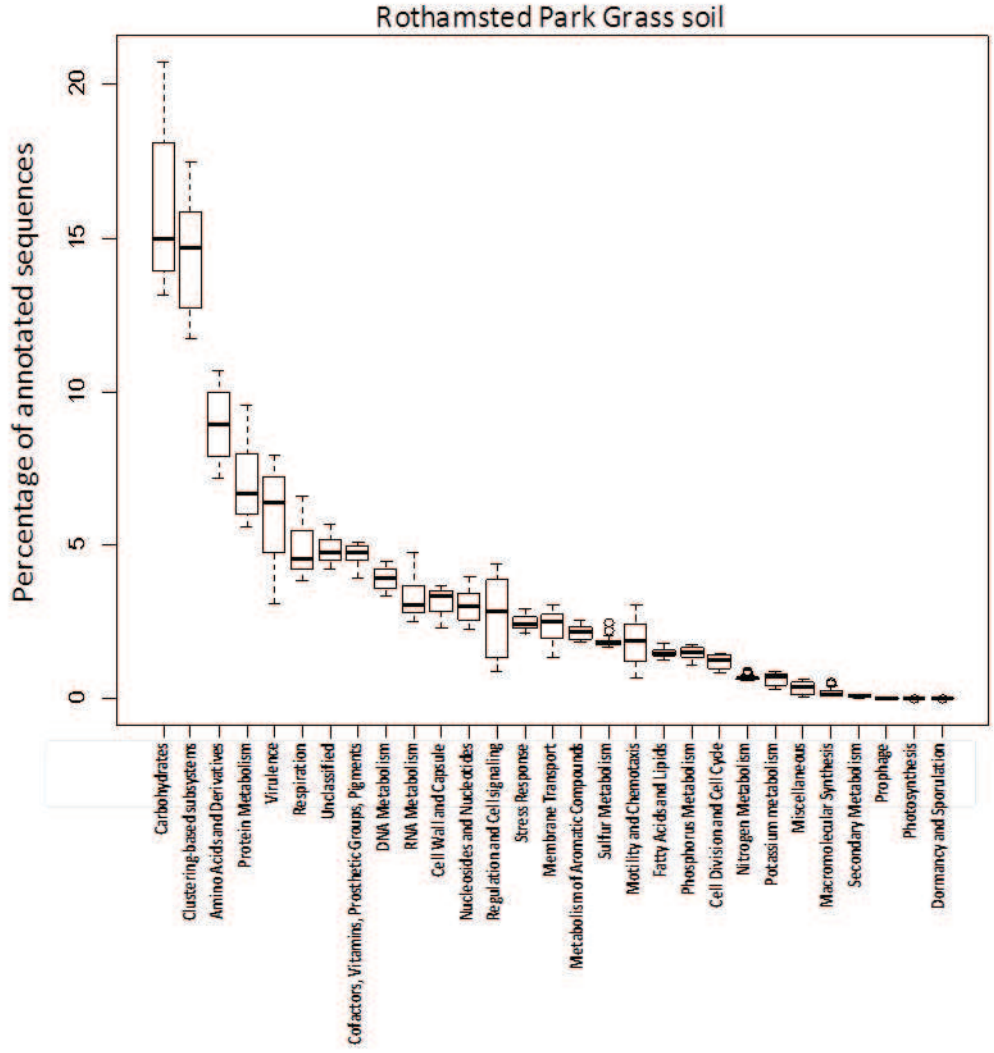


Figure 6 : Distribution relative de sous-systèmes fonctionnels du métagénome de Rothamsted en intégrant des flexicats (stringence de lyse cellulaire et d'annotation des séquences en faisant varier l'E-value).

Afin d'aller un peu plus loin qu'un débat sur la méthodologie à appliquer en métagénomique (même si ce débat me paraît important pour définir de meilleurs expérimentations dans le futur proche), je me suis essayer à imaginer quelques applications de ce domaine pour un futur plus lointain. En effet, Lorsque l'on regarde la métagénomique dans sa globalité, les perspectives apparaissent clairement : usant d'outils de plus en plus performants (séquençage, annotation, assemblage, *etc.*), les microbiologistes n'ont dorénavant plus qu'à se baisser pour récolter le fruit de plus de trois milliards d'années d'évolution et d'adaptation microbienne. Le tout étant maintenant de savoir ou se trouvent les vrais découvertes, et faire la moisson le premier… La course est lancée, les équipe de recherche sur la ligne de départ, ou bien juste quelques longueurs plus loin, mais rien n'est jouer. L'imagination, la subtilité, et bien sur le facteur chance aura aussi son mot à dire.

Cependant, dans quelques décennies, la majorité des fonctions, connues ou non, seront déjà dans les bases de données numériques. L'intérêt des études de séquençage environnemental devraient décroitre en conséquence. La communauté se concentrera alors probablement dans la biologie de synthèse, utilisant les données cumulées au fil des ans pour créer chimiquement des microorganismes capable de réaliser des procédés d'intérêt. Ma vision de ce futur possible est décrit dans le chapitre 4, section 2 (« Digitizing genetic structures prior to synthesizing new microorganisms: from knowledge to evolution v2.0, but for whom? »).

Pour allez un peu plus loin dans les perspectives de ce domaine de recherche, j'ai aussi proposé une approche expérimentale permettant, basée sur des données métagénomiques, de synthétiser des microorganismes capables de survivre et de se développer à la surface de Mars, la planète rouge, afin d'optimiser des efforts futurs de terraformation de cette planète. Cette perspective est présentée dans le chapitre 4, section 3 (« Synthetizing microbial life to optimize Martian terraformation labors : from the terrestrial evolution to in lab extraterestrial adaptation experiments »).

Ces deux perspectives sont indépendantes de mes travaux de recherche, mais m'ont fournis un certain recul sur mes propres données. Elles permettront aussi au lecteur de se faire une idée succincte sur des futurs scientifiques possibles de mon point de vue.

7. Conclusion :

Cette « synthèse » de quelques pages a pour but d'aider le lecteur à découvrir ce manuscrit de thèse en comprenant les liens qui unissent les différentes sections, les différents chapitres présentés. Au travers de cette « synthèse », il est possible de visualiser une unité concrète, représentant ma perception des communautés microbiennes basée sur mes propres travaux, mais aussi bien sûr de ceux que j'ai eu l'occasion d'étudier, à travers les articles ou conférences internationales.

J'ai en effet eu la chance d'assister à certaines conférences internationales, et d'y présenter mes travaux :

-Le chapitre 2, section 3 a été oralement présenté durant la conférence ISME, Seattle, USA, en aout 2010. Titre: "Soil metagenomics: can we access the diversity?".

-Le chapitre 1, section 2 a été oralement présenté durant une conférence sur la métagénomique du sol, a Braunschweig, Allemagne, en décembre 2010. Titre: "Global metagenomic comparisons: a functional confrontation of pieces of the microbial word puzzle".

-La conclusion, section 2 a été oralement présentée durant la première conférence du Earth Microbiome Project, a Shenzen, Chine, en juin 2011. Titre: "From the Terragenome project to the global metagenomic comparisons: implications for the Earth microbiome project".

-Le chapitre 3, section 2 a été oralement présentée durant le troisième workshop annuel de métagénomique du sol à Chicago, USA, en Octobre 2011. Titre: "Perturbing complexe microbial communities for metagenomic discoveries: one design evenness at the time".

Présenter mes travaux avant leur publication m'a permis de les améliorer considérablement, de part l'effort intellectuelle nécessaire a la construction d'un oral de qualité, mais aussi de part les questions du public. Ils ont forgé une partie de ce manuscrit de thèse. Je les en remercie donc.

Pour finir, je souhaite juste expliquer au lecteur que les parties présentes dans les annexes ne sont pas dans le manuscrit afin d'aider à la cohérence de ce dernier. Ce choix est donc indépendant de la qualité scientifique de ces sections.

Context:

As recent technologies allowed the sequencing of millions of DNA fragments, Timothy M. Vogel (my PhD supervisor), Pascal Simonet and the French national research agency (ANR) decided to lead an international soil metagenomic consortium (named Terragenome, http://www.terragenome.org/). The objective was to combine the experiences of the environmental microbial genomics group, the Libragen Company (Renaud Nalin and Patrick Robe) and the Genoscope sequencing platform (Denis Lepaslier and Eric Pelletier) to provide a considerable quantity of soil nucleic diversity data to the international community.

Thanks to Tim and Pascal, I had the opportunity to take a major role in this effort during three years of studies entirely dedicated to this effort. The main objective of my thesis was to characterize as deeply as possible the microbial communities present in the soil selected by Terragenome: grassland (Park Grass plot 3d) from an experimental station located in United Kingdom and named Rothamsted. The global strategy was to use an environmental DNA sequencing approach instead of a cultural method, which is highly limited by the number of species that can grow easily. This approach is commonly called metagenomics and aims to study the genomic structure of microorganisms in their environment.

The ANR project was called Metasoil (http://metasoil.univ-lyon1.fr/) and helped start the Terragenome consortium during the Metasted meeting (held on December 13-14, 2008 in Lyon, France; http://www.ampere-lab.fr/spip.php?article308). The Terragenome objective was clear: to sequence and assemble entirely the Rothamsted soil metagenome (see Chapter II.1). In fact, due to the complexity of soil microbial diversity, scientists decided to focus on one unique soil and selected the Rothamsted experimental station due to the information scientists possess about this soil for more than 150 years.

This meeting took place only three months after the beginning of my PhD, and provided me with an important vision of the topic (soil metagenomics), by integrating limits, objectives and perspectives presented by the international community. It was also a unique opportunity to meet researchers leading microbial ecology, and to be heard in the framework of Metasoil.

The first year of my PhD was mainly methodological. I tried to increase the accessibility of the Rothamsted soil metagenome, a major objective to reach the Terragenome objective. In fact, several studies emphasized the difficulty to access soil genetic diversity due to important DNA extraction limits. Thus, tens of protocols are published and can be selected for metagenomic surveys. None was proved to be unbiased. However, the goal of Terragenome was to sequence entirely this specific soil metagenome. It appeared to me that a single DNA extraction approach was insufficient to access all the diversity.

Thus, I designed and proposed a metagenomic fractionation strategy to stimulate the detection of soil microbial diversity signatures (see Chapter II.2). Using this approach, the number of detected taxa increased considerably (>80%) in comparison to what could be

detected with one single method. In addition, each method provided a different microbial community evenness, making quantitative comparisons between datasets difficult. In fact, how can scientists select one method to represent an uncharacterized environment when all methods access different parts of the total genetic diversity?

Because it was novel, a major difficulty was then to explain to the international community that it is probably more reasonable to apply several approaches than one for metagenomic surveys. We presented this strategy during several international conferences, and finally proposed a commentary to describe in more detail the possible alternative (see conclusion part, section 2). The word “Flexicate” was proposed to represent the different methods that can be used to study microbial communities using ‘omic approaches.

Then, we decided to apply this strategy to the Rothamsted soil. 13 pyrosequencing runs (a run generates generally one million reads of 350bp) were done by varying the season, depth and DNA extraction protocols from the same Rothamsted soil. This approach aimed to provide a global vision of this microbial community (the main objective of my PhD) by creating a standard deviation representing both natural variations (usually used during metagenomic surveys) and methodological fluctuations (never done before). These datasets were used to characterize this community and to compare it to those present in other already sequenced environments (see Chapter II.3). This global metagenomic comparison effort aimed to observe unusual microbial life styles from across the planet. Due to the interest of this approach for microbiologists, we proposed a perspective to democratize the mining of metagenomic datasets (see conclusion part, section I).

I used 14 other environmental metagenome to help study by comparison the soil habitat, I decided to look at the peculiarities of all of them even if soil was, of course, my prime objective. This descriptive effort was considerable but quenched my curiosity and is presented as the bibliography part of the manuscript (see Chapter I).

The 13 million reads generated from the Rothamsted soil where unfortunately (or fortunately when we think about what we still have to discovered on this environment) largely insufficient to access all the genetic diversity we can extract from soil. In fact, only short genomic structures were reconstructed from these datasets. However, we were able to estimate a certain sequencing effort necessary to begin stimulating soil metagenomic assembly efficiency (see Chapter II.3). Interestingly, an important American project named Great Prairie and focusing on the same environment was able to generate billions of soil metagenomic reads using another sequencing technology. This project was launched few months after Metasoil. Based on the partial information I have about this project, even a deep sequencing effort was unable to reconstruct long genomic structure from a soil metagenome.

The sequencing of DNA extracted directly from soil was a good approach to study predominant microorganisms and to characterize these communities (relative distribution of

major functions and taxa). However, only a tiny part of a soil metagenome can currently be sequenced mainly due to sequencing technology limits. In addition, only predominant microorganisms can be targeted. Thus, the vast majority of soil microorganisms (between 10^4 and 10^7 species estimated in one gram of soil) cannot be studied using this approach. In addition, due to the complexity of the sequenced DNA, it is not possible to reconstruct genomic structures from a soil metagenome (based on Metasoil and Great Prairie current results).

Then, to access other diversities and stimulate assembly efficiency by limiting the accessible genetic diversity, I proposed to apply extreme environmental conditions to the Rothamsted soil during months prior to extracting and sequencing DNA (see Chapter III.2). Interestingly, the approach succeeded and provided access to several mainly reconstructed genomes corresponding to microorganisms capable of resisting high concentrations of mercury, heavy metals and ethanol, respectively. But more important, this work provided a considerable opportunity: the possibility to access the genomes of the vast majority of lowly represented soil microorganisms using controlled environmental modifications.

Thus the Terragenome goal was finally partially possible, by applying an alternative strategy. Microbiologists focusing on soil metagenomics should now apply a large variety of conditions to soil microbial communities. Everything has still to be done but these three years focusing on one habitat provided a novel approach for metagenomic surveys (flexicates instead of replicates only) and new possibilities to sequence and assemble sub-diversities of a soil metagenome.

Since the prime objective of the Metasoil project was done only after few months of work; I had a considerable flexibility to propose additional experiments. My supervisor always let me pursue my ideas and helped me design new experimental approaches, proposing perspectives. Thus, I had the opportunity to study the richness aspect of soil microbial communities by transferring communities between soils based on an original idea of Pascal Simonet (see Chapter III.1), to discuss metagenomic pitfalls with a part of the community (see Chapter IV.1), and even to reflect on the potential of metagenomic efforts to help a Martian Terraformation (see Chapter IV.3). Additional approaches were done (*e.g.*, GC content fractionation of a soil metagenome prior sequencing, study of old samples dried since 1876). Since these last projects are ongoing, they will be only partially presented in the manuscript (annexes sections).

In parallel to my research in the laboratory, I had the opportunity to present a majority of my results during international conferences:

-The chapter 2, section 3 was an oral presentation during the **ISME conference, Seattle, USA, August 2010**. Title: “Soil metagenomics: can we access the diversity?”

-The chapter 1, section 2 was an oral presentation during the **soil metagenomics conference, Braunschweig, Germany, December 2010**. Title: “Global metagenomic comparisons: a functional confrontation of pieces of the microbial word puzzle”.

-The conclusion part, section 2 was an oral presentation during the the **First International Earth Microbiome Project Conference, Shenzhen, China, June 2011**. Title: “From the Terragenome project to the global metagenomic comparisons: implications for the Earth microbiome project”.

-The chapter 3, section 2 was an oral presentation during the **3rd Annual Argonne Soil Metagenomics Workshop, Chicago, USA, October 2011**. Title: “Perturbing complexe microbial communities for metagenomic discoveries: one design evenness at the time”.

Chapter 1. Bibliography:

1. **Bibliography introduction (French part)**
2. **Decrypting global metagenomic comparisons of the microbial world**

Introduction bibliographique:

Bien que les scientifiques ne sachent pas avec précision quand, où, et surtout comment la vie a émergé sur Terre, il est généralement accepté que les microorganismes étaient déjà largement repandus et évolués dans l'Archéen, il y a 2.5 à 3.5 milliards d'années (Altermann and Kazmierczak, 2003). Due à leur taille microscopique et leurs considérables capacités d'adaptation (Siefert et al., 2009), ils ont colonisé l'immense majorité de la surface de la planète et sont dorénavant une population estimé a 10^{30} cellules (Whitman et al, 1998). Par ailleurs, ils sont indispensables à notre vie (Turnbaugh et al., 2006) et à la santé de notre planète (Falkowski et al., 2001), ce qui les rend cruciaux à nos yeux.

Cette diverse et complexe forme de vie à d'abord été décrite par Anton van Leeuwenhoek en 1676 (pour plus de détails, se reporter à Porter et al., 1976) et est devenue avec l'émergence de la microbiologie le sujet d'étude des écologistes pour comprendre leurs rôles et fonctions mais aussi afin de découvrir de nouveaux gènes.

Cependant, due à de nombreuses difficultés liées à l'étude des microorganismes et en particulier aux limites de culture cellulaires (Amann et al., 1995), seulement une minorité d'espèces (moins de 1%) ont été étudiées durant des siècles. Ainsi, malgré de nombreuses études sur les procaryotes, notre connaissance de ces 10^{30} cellules a été considérablement limitée.

En 1998, Handelsman , Rondon , Brady , Clardy et Goodman proposèrent une nouvelles approche afin d'accéder à la diversité génétique du sol en clonant des séquences d'ADN extraites de l'environnement dans des banques de clones (Handelsman et al., 1998). Cette méthode avait pour but de stimuler la découverte de produits naturels d'intérêt en détournant les limites de culture.

Cette équipe de recherche lançait avec cette publication un nouveau domaine d'étude pour les écologistes microbiens et appelé la metagenomique (Meta: *μετά* en grec).

La metagenomique a été dans un premier temps concentrée sur le criblage de librairies environnementales afin de détecter de nouvelles activités biologiques, et le même laboratoire démontra par la suite que de l' « ADN extrait directement du sol pouvait représenter une source valable de nouvelles informations génétiques et était accessible en utilisant des banques BAC » (Rondon et al, 2000), mais aussi que de nouveaux antibiotiques pouvaient être découvert en utilisant cette approche (Gillespie et al., 2002).

Durant les années qui suivirent, cette méthode fut utilisée et améliorée dans d'autres laboratoires. Comme exemple, et pour stimuler le nombre de clones positifs après criblage, les microorganismes d'intérêt ont été stimulés avant la construction de banques (e.g., incubation en présence de glycérol et 1,2-propanediol afin de détecter de nouvelle activités alcool oxydoreductase; Knietsch et al., 2003). Pour améliorer la caractérisation de clones, Sebat et ses collègues (2003) ont proposé d'hybrider les banques avec des puces à ADN

(microarray) correspondant à une banque cosmidique déjà connue. Enfin, la méthodologie a elle aussi été améliorée afin d'accéder à des fragments d'ADN de haut poids moléculaire (e.g., Bertrand et al., 2005).

En parallèle de la construction et du criblage de librairies de clones, des études se sont basées sur le séquençage direct (« shotgun ») d'échantillons d'ADN environnementaux. Avec cette approche, les séquences d'ADN environnementales sont extraites, clonées dans Escherichia coli puis séquencées sans aucune étape de criblage. Cette méthode elle aussi détourne les limites de culture cellulaire d'espèces procaryotiques, et les séquences ainsi générées ont pour but de refléter en termes de distribution relative la population microbienne vivant dans un environnement d'étude et sont communément appelées des metagenomes. En utilisant cette technique, des metagenomes ont dans un premier temps été générés à partir d'ADN extrait de la mer des Sargasses (Venter et al., 2004), de biofilms d'un drainage minier acide (Tyson et al, 2004), de sol et du fond des océans (Tringe et al., 2005).

Avec la présence de deux approches distinctes pour étudier un metagenome (expression contre séquençage direct), la description de la metagenomique a évolué. Basé sur notre propre vision du domaine, la définition de Schloss et Handelsman (2005) pour qui « la metagenomique est l'analyse d'un mélange de génomes microbiens (appelé le metagenome) en utilisant une approche basée sur l'expression ou le séquençage » reflète ce domaine de recherche et accepte le fait que le « mélange » ne représente pas nécessairement les communautés microbiennes dans leur état naturel. En effet, de récentes études ont soulignées d'important biais durant les approches d'extraction d'ADN (e.g., dans le sol ; Delmont et al., 2011) qui impliquent une modification de la distribution de séquences à la base de toute étude de metagenomique, générant ainsi de possibles erreurs scientifiques importantes.

Cependant, cette définition exclue les technologies de puce à ADN bien que les chercheurs utilisant cette technologie considèrent qu'ils exploitent des approches métagénomiques. Cette définition peut donc évoluer. Pourquoi pas : « Les approches métagénomiques englobent l'étude d'ADN extraits directement de l'environnement et correspondant à un mélange de génomes microbiens appelé metagenome. » Mais dans ce cas la, les techniques d'empreintes moléculaires sont elles aussi incluses. Pour les chercheurs considérant que ces techniques ne font pas partie du domaine de la metagenomique alors que les puces à ADN le sont, ce pourrait être un compromis de dire que « les approches métagénomiques englobent l'expression, le séquençage ou l'hybridation d'ADN extraist directement de l'environnement et correspondant à un mélange de génomes microbiens appelé metagenome. »

Dans un futur proche, les technologies permettront plus facilement d'assembler un nombre considérable de génomes directement de l'environnement, modifiant ainsi en profondeur la microbiologie environnementale. En conséquence, la metagenomique est probablement juste une période entre l'étude de microorganismes cultivables et l'étude de l'immense

majorité non cultivable, et fera un jour partie de l'histoire de la microbiologie. La définition sera alors probablement fixée définitivement.

Après 2005, la metagenomique a été appliquée à une large gamme d'applications et est devenue un domaine de recherche d'importance. Comme exemples, cette approche a été utilisée pour séquencer le code génétique d'un mammouth ((Poinar et al., 2006), l'ADN génomique d'un homme de Neandertal (Noonan et al., 2006), du microbiome de l'intestin de l'homme (Gill et al., 2006; Qin et al., 2010) et de souris obèses ou non (Turnbaugh et al., 2006) ou encore afin de construire des gènes de dégradation de xénobiotiques (Boubakri et al., 2006). Mais la principale utilisation d'approches métagénomiques a probablement été de séquencer de l'ADN environnemental, et des metagenomes ont ainsi été générés a partir d'ADN extraits d'océans (avec le projet de surveillance globale des océans ; Yooseph et al., 2007, Williamson et al., 2008), du fond de la Méditerranée (Martin-Cuadrato et al., 2007), d'atolls coralliens (Dinsdale et al, 2008), de boues actives (Garcia et al, 2006), de sédiments (Kunin et al., 2008), de neiges arctiques (Larose et al., 2010), ou encore d'air pollué (Tringe et al., 2008). Chaque metagenome fournit de nouvelles informations sur les communautés microbiennes de l'environnement séquencé et améliore la connaissance globale des microorganismes.

Dans certains cas, le rendement d'ADN extraits n'est pas compatible avec l'appétit des technologies de séquençage, et les chercheurs peuvent alors utiliser une étape de PCR (Polymerase Chain Reaction) appelée « Whole Genome Amplification » (WGA) à partir d'une amplification dites en déplacement multiple (MDA pour Multiple Displacement Amplification) afin d'augmenter la concentration d'ADN. L'originalité de cette amplification est l'utilisation de petites amorces qui ont ainsi le potentiel de fixer à peu près chaque partie des génomes extraits et donc d'amplifier aléatoirement la totalité d'un échantillon d'ADN. Cette technologie a été utilisée afin de produire des banques pour séquencage direct (Rohwer et al, 2001), de générer des metagenomes représentant des communautés virales marines ((Breitbar et al., 2002) ou encore d'accéder à la population microbienne de sédiments contaminés (Abulencia et al., 2006) ou d'atolls coralliens (Dinsdale et al., 2008). Cette étape « magique » qui peu aussi créer de l'ADN à partir de rien (expériences personnelles) représentait un outil puissant pour les écologistes microbiens (Binga et al., 2008). Cependant, une partie de la communauté scientifique a toujours été suspicieuse et a préféré ne pas l'utiliser même si c'était la seule méthode permettant d'étudier des échantillons d'ADN ou d'ADN complémentaire. Finalement, les kits de MDA ont été décrits comme compromettant l'analyse quantitative de metagenomes (Yilmaz et al., 2010) et ces résultats devraient limiter l'utilisation de cette technique dans le futur.

Plus récemment, les kits de MDA ont aussi été utilisés afin d'amplifier des cellules seules extraites de l'environnement sans aucune étape de culture cellulaire (Woyke et al, 2009; Woyke et al, 2010). Cette approche a pour but de stimuler l'assemblage de parties de metagenomes, qui est considérablement limité pour les communautés procaryotes

complexes lorsque des fragments d'ADN extraits et purifiés sont séquencés aléatoirement. Cependant, les biais liés au MDA doivent être définis avec précision pour comprendre leur impact lors de l'assemblage des séquences d'une cellule seule.

Depuis l'utilisation du séquençage dit en « shotgun » et plus récemment des technologies de séquençage à haut débit (Shendure and Ji, 2008; Kahvejian et al., 2008) afin de générer des metagenomes de plus en plus gros, des outils de bioinformatique ont été développés pour aider les microbiologistes a analyser leurs propres jeux de données. Avec plus ou moins d'originalités, ces outils ont été nommés MetaGene (Noguchi et al, 2006), Uniprot (Uniprot consortium, 2007, 2008, 2009, 2010), MEGAN (Huson et al, 2007, 2009; Mitra et al, 2009), CAMERA (Seshadri et al, 2007), IMG/M (Markowitz et al, 2006, 2008) et MG-RAST (Meyer et al, 2008) dont les tables peuvent être exporter dans STAMP (Parks and Beiko, 2010) pour des analyses statistiques, MetaTISA (Hu et al, 2009), Orphelia (Hoff et al, 2009), ShotgunFunctionalizeR (Kristiansson et al, 2009), WebCARMA (Gerlach et al, 2009), ou encore METAREP (Goll et al, 2010).

Malgré des avancées considérables dans le domaine depuis 1998, il y a actuellement d'importantes limites dans l'étude exhaustive de metagenomes environnementaux. En premier lieu, du a la présence d'importants biais d'extraction d'ADN, l'échantillon d'ADN qui est la base de toute étude metagenomique est limité et ne représente pas forcement la diversité nucléique présente quelque soit l'approche choisie. Cette limite peut être allégée en variant les protocoles permettant d'accéder à la richesse génétique présente dans un endroit donné mais ne peut malheureusement pas être résolue entièrement pour le moment (Delmont et al., 2011).

En second lieu, malgré d'importante avancées en terme de technologie de séquençage ((Shendure and Ji, 2008; Kahvejian et al., 2008) et des travaux ambitieux (e.g. Nealson and Venter, 2007; Vogel et al, 2009, Qin et al, 2010), la majorité des études n'a pas été capable pour le moment de séquencer la totalité de la diversité génétique présente dans un échantillon d'ADN, prévenant ainsi l'assemblage complet de metagenomes.

Afin de passer outre cette considérable limite, des courbes de raréfaction sont souvent utilisées afin d'estimer à la fois les diversités fonctionnelle et taxinomique présentes dans un échantillon d'ADN (e.g. Tringe et al., 2005; Roesch et al., 2007). Malheureusement, la distribution inégale des espèces ainsi que le manque d'information à propos de l'amplitude de la « rare biosphere » (espèces peu représentées), les courbes générées peuvent être biaisées et sous estimer la vraie diversité environnementale.

Finalement, après avoir extrait et séquencé une partie (qui peut être minuscule pour les environnements hautement biodiverses comme le sol et les sédiments) d'un metagenome, des processus d'annotations sont nécessaires afin de proposer une fonction et une bactérie hôte a la séquence générée. Cependant, les bases de donné utilisées pour cette annotation sont toujours limitées et possèdent des informations environnementales inégales (A

sequence of changes, 2010), impliquant ainsi une possible vision alternée de la diversité d'un metagenome.

En dehors de ces restrictions communes, certains environnements possèdent des difficultés additionnelles dans l'étude complète de ses communautés. En fait, la diversité nucléique varie fortement entre niches écologiques et est connue pour être limitée dans le corps des animaux par exemple (e.g., Qin et al, 2010) et considérable dans les sols et sédiments qui représentent toujours des boites noires pour les scientifiques (Sleator et al., 2008). D'autres paramètres, comme la distribution des espèces (Morales et al., 2009), la stabilité du metagenome au cours du temps (Alonso-Sáez et al., 2008) et l'hétérogénéité environnementale (e.g. dans le sol) peut augmenter la difficulté lors de l'étude d'un metagenome complet.

Cependant, avec l'émergence d'importante avancées technologiques et de nombreux outils pour étudier des échantillons d'ADN complexes (Rajendhran and Gunasekaran, 2008), des travaux ambitieux ont récemment été effectués afin d'étudier des metagenomes environnementaux et des données non négligeables ont ainsi été générées.

De plus, grâce aux banques de données publiques accessible sur internet (e.g., http://metagenomics.nmpdr.org/), la comparaison de séquences annotées provenant de divers metagenomes peut être effectuée par tous les microbiologistes intéressés.

En fait, en utilisant des outils de bioinformatique et des metagenomes correspondant à des environnements variés, il est possible de comparer des potentiels fonctionnels de différentes communautés procaryotiques et ainsi de mieux comprendre leur style de vie. Ce type de comparaison a été réalisé dans un premier temps par Tringe et ses collaborateurs en 2005, suivi par Dinsdale et al. (2008, et Willner et al (2009). Cependant, par rapport a la quantité de metagenomes générés (e.g., plus de 10 000 metagenomes privés dans MG RAST en 2011), le nombre de publications traitant de comparaisons globales de metagenomes est considérablement limité. Cette observation peut facilement être expliquée par le fait que tous les projets métagénomiques ont été réalisés avec des approches spécifiques, sans standards généraux entre projets. Ainsi, la communauté internationale a des doutes quant a la possibilité de comparer une compilation de metagenomes quand différentes approches d'extraction d'ADN et différentes technologies de séquençage sont utilisées.

Cependant, grâce à de considérables différences entre environnements, les fluctuations méthodologiques ne limitent pas la comparaison globale de metagenomes (Delmont et al., perspective) et chaque écologiste environnemental peut utiliser ces données pour ses propres recherches (souligner des particularités métagénomiques, rechercher une fonction ou une espèce parmi des environnements, trouver des corrélations entre espèces et fonctions, sélectionner les séquences correspondant a un gène d'intérêt dans divers environnements, etc.).

Il est important de noter que la technologie de séquençage utilisée impacte considérablement le pourcentage de séquences annotées (Delmont et al., perspective). En fait, plus les séquences générées sont longues, meilleur sera le processus d'annotation. En addition, quand la proportion de séquences eucaryotes augmente, le pourcentage de séquences annotées décroit fortement (données non fournis).

Cependant, standardiser la distribution fonctionnelle et taxinomique en fonction de ce pourcentage parmi les différents metagenomes étudiés allège ces problèmes et permet la comparaison de metagenomes correspondant à de nombreux projets même si les standards ne sont pas présents. Bien sur, différents protocole d'extraction d'ADN sont aussi utilisés en fonction du projet, mais nous assumons que d'imposer des standards à la communauté internationale peut être dangereux pour la science car nous ne pouvons savoir quelle approche fournit un échantillon d'ADN reflétant au mieux la réalité.

De plus, la présence de biais lors des processus d'annotation phylogénétique a pour effet la génération de résultats considérablement différents en fonction de la base de donnée utilisée (e.g., RDP, Greengenes, SEED). En conséquence, comparer la distribution taxinomique a l'intérieur d'un même metagenome est critique et hautement influencé par le jeu de données de référence utilisé. Cependant, lorsque l'on compare la distribution de la même espèce parmi différents metagenomes, bien sûr cette distribution est influencée par le jeu de données utilisé, mais les différences de distributions sont réelles (si les biais d'extraction ne sont pas tenus en compte). Ainsi les comparaisons globales contournent une source importante d'erreurs présentes quand la distribution d'espèces est comparée dans un metagenome. En fait, il est actuellement hasardeux de dire l'espèce « a » est plus représentée que l'espèce « b » dans le metagenome « x », mais en comparant les metagenomes « x » et « y », il est possible d'observer que les espèces « a » et « b » sont plus représentée dans un environnement que dans un autre. Cette exemple est aussi applicable pour les fonctions, et représente un considérable avantage des comparaisons globales de metagenomes.

Comparer des metagenomes provenant de points spécifiques de part le monde devrait permettre de comprendre la distribution des 10^{30} procaryotes présents sur Terre et ainsi d'apprécier leur évolution depuis presque 4 milliards d'années. C'est aussi une approche unique afin de visualiser le potentiel génétique de différent environnements de notre planète dans l'adaptation des futures changements climatiques et pollutions crées par l'Homme.

Nous assumons que les comparaisons globales peuvent aider les scientifiques à observer sans aucun *a priori* les particularités fonctionnelles et taxinomiques des metagenomes et environnements qu'ils étudient en comparant leurs propre jeux de donner à d'autres. Dans le but de stimuler l'étude de données séquencées et de visualiser la répartition et spécificité de gènes en fonction des environnements déjà séquencés, nous avons comparé a différents niveaux taxinomiques et fonctionnels 77 metagenomes correspondant a des échantillons

d'ADN extraits d'océans, d'atolls coralliens, du fond des océans, d'environnements aquatiques d'Antarctique, de neige arctique, d'environnements terrestres (sédiments hyper salins, sols, boues actives, biofilms, drainages miniers acides), d'air pollué, et d'animaux (fèces humains, caecum de souris et de poulet, et rumen de vache). Les résultats montrent d'importantes spécificités fonctionnelles et taxinomiques pour chaque environnement, soulignant ainsi les capacités d'adaptation des procaryotes au niveau de la planète.

Avec la révolution actuelle de séquençage, la rapide augmentation de projets de séquençage d'environnements et l'émergence de nouveaux outils de bioinformatique, nous somme probablement entrés dans un nouvel âge d'or de la microbiologie avec les comparaisons globales de metagenomes comme symbole de réussite. Cependant et à cause de considérable limites et biais durant l'extraction d'ADN et les processus d'annotation des séquences, la science metagenomique pourrait représenter une boite de Pandore qui attirera les scientifiques dans une connaissance erronée durant des années. Mais qui veut étudier seulement 1% d'un domaine de recherche ? Ainsi les scientifiques étudient activement l'image générale et déformée des microorganismes en utilisant des approches métagénomiques. Nous devons améliorer les méthodes métagénomiques mais aussi optimiser la culture des 10^{30} procaryotes afin d'affiner notre perception de cette forme de vie, leur interaction et impact sur le climat et les activités humaines*.

*Regarder la section suivante pour voir les références.

Decrypting global metagenomic comparisons of the microbial world

Tom O. Delmont and Timothy M. Vogel

Environmental Microbial Genomics Group, Laboratoire AMPERE,
Ecole Centrale de Lyon, Université de Lyon,
36 avenue Guy de Collongue, 69134 Ecully, France.

Abstract:

Microbial ecology is beginning to interact with metagenomics and many microbiologists are attracted to metagenomics in the hope of discovering novel relationships between microorganisms and/or confirming that work done on isolates applies to the remaining uncultured members of the different ecosystems. With a growing number of available metagenomic datasets, metagenomes can be intensively mined by microbial ecologists in search of previously undetected correlations (both structural and functional). Here, we provide a preliminary exploration of 77 publically available metagenomes corresponding to DNA samples extracted from oceans, atoll corals, deep oceans, Antarctic aquatic environments, Arctic snows, terrestrial environments (sediments, soils, sludges, microbial fuel cell anode biofilms, acid mine drainage biofilms), polluted air, and animal and human microbiomes (human feces, mouse and chicken cecum, and cow rumen). Results show well-defined environmental specificities that emphasize microbial adaptation and evolution capabilities. Unexpected observations were also made for several ecosystems, thus providing new hypotheses about the life style of their microbial communities. Available metagenomes are a gold mine of underexploited information that could be used to explore specific microbial structural and functional relationships. The statistical analysis provided here depends in part on replicates from the different ecosystems. With the continued emphasis on metagenomic sequencing, future analyses should support rigorous statistical treatment. This preliminary metagenomic decryption could represent a pilot-scale test for a future Earth microbiome global comparison.

Key words: Global metagenomic comparisons, microbial communities, adaptation, evolution.

Summary

Introduction:

Even if scientists do not know with precision when, where and especially how life emerged on Earth, it is generally accepted that microorganisms were already relatively widespread and advanced in the Achaean between 3.5 and 2.5 billion years ago [1]. Due to their considerable adaptation capacities [2], they colonized the entire planet with current population estimated to be 10^{30} cells [3] and are considered indispensable for our life (e.g., [4]) and the health of the planet [5]. This diverse and complex form of life was first described by Anton van Leeuwenhoek in 1676 (For more details, see Porter and collaborators, [6]) and has today become a critical area of study for microbial ecologists. Considerable knowledge was first generated from cultivated microorganisms. They were found to be ubiquitous, and were studied from various habitats, like in bread [7], milk [8], intestinal tract of animals [9], or in the environment (e.g., hot springs [10]). Cultivable microorganisms were also studied for their bioremediation capacities (e.g., hydrocarbon degradation, [11], or the discovery of bio-molecules of interest (e.g., new antibiotics [12]. However, because of important difficulties to study microorganisms due to cell culture limitations (e.g., [13]), only a minority of species (less than 1%) were studied for centuries. As a consequence and in spite of numerous studies on microorganisms, the knowledge about all of these 10^{30} cells was limited. Technological advances have had important influences on increasing our knowledge of the critical role of microorganisms. For example, with the emergence of fluorescent microscopy technologies and the possibility of counting cells in the environment, uncultured bacteria cells were unexpectedly found to be highly abundant in the ocean [14]. As a consequence, they were found to be responsible for up to 98% of marine primary productivity [15] and to play a major role in the carbon cycle.

During the last several decades, scientists focused on various experiments to improve knowledge about microbial communities and to study the impact of environmental factors, time and space on their biodiversity (see [16] for a review of articles from 1975 to 1999). Thus, markers were defined to facilitate their identification (e.g., rRNA gene, [17]) and new methods emerge to study complex communities without any culture step. In particular, molecular biology based on fingerprint methods were applied to environmental DNA samples during the 1990s (e.g., denaturing gel electrophoresis [18]) and are still largely used to study microbial diversity in environmental samples (e.g., using terminal restriction fragment length polymorphism [19]). At the same time, microbial ecologists began sequencing parts of 16S rRNA genes amplified from the DNA pools [20]. Their annotation based on a reference databases provided phylogenetic analyses about microbial populations. The sequencing of 16S rRNA amplicons is now largely used by microbial ecologists to study various complex environments, like oceans [21], soils [22,23] and extreme environments (e.g., Antarctic soils, [24]. However and as a major limit, molecular techniques do not provide access to entire genomic structures like chromosomes and plasmids.

In 1998, Handelsman and collaborators proposed a new approach to access soil genetic diversity by cloning DNA sequences extracted and purified from the environment into a clone library [25]. By-passing culture limitations, this method stimulated the discovery of natural-products of interest. This began the explicit field of “metagenomics” (Meta: μετά in Greek). Initially, metagenomics focused on the screening of environmental genetic libraries to detect new biological activities [26] and new antibiotic activities were discovered using this approach (e.g., turbomycin A and turbomycin B genes discovery in a metagenomic library, [27]). Since then, this method has been both used and improved. As an example and to stimulate the number of positive clones, microorganisms of interest were enriched prior library construction (e.g., the sample incubation in the presence of glycerol and 1,2-propanediol to stimulate new alcohol oxidoreductase activities; [28]). To improve clone characterization, Sebat and colleagues hybridized libraries with a microarray corresponding to an already characterized cosmid library [29]. Methodology was also improved to access high molecular weight DNA fragments (e.g., [30]). As expected, the construction of genetic libraries allowed the detection of several genes of interest and is still largely used and improved to increase metagenomic discoveries (e.g., Uchiyama and Miyazaki, 2009, 616-22). As examples, novel lipolytic genes [31,32], antibiotic resistance genes [33], amylase [34] and alkali-thermostable lipase [35] were identified using this approach. In parallel to library construction and screening, some studies were based on the direct shotgun sequencing of environmental DNA pools. With this approach, environmental DNA sequences are extracted, cloned into Escherichia coli and then sequenced without any screening step. The sequences so generated provide in principle relative distributions of the microbial populations living in the studied environments and are also commonly called metagenomes.

Using this strategy, the first metagenomes were generated from Sargasso Sea [36], acid mine drainage biofilms [37], soil and sunken whale skeletons [38]. Considerable information was extracted from these pioneer studies, making it a promising new field of research. For example, more than one million of previously unknown genes including hundreds of new rhodopsin-like photoreceptors were sequenced from the Sargasso Sea. This study emphasized also an unexpected high diversity in this environment that included new phylotypes. In contRAST, the partial genomic assembly of predominant microorganisms from the acid mine drainage biofilms emphasized the interest of metagenomic approaches to reconstruct genomes when the diversity is highly limited. In addition, the reconstruction of genomes from this environment provided new insights on resistance strategies in an extreme environment. Finally, Tringe and colleagues highlighted the interest of comparing the relative distribution of genes between metagenomes for interpreting and diagnosing environments. These three articles could represent the base of a new field that aims to decrypt, assemble and compare microbial communities using direct sequencing strategies.

After 2005, metagenomic approaches were applied to a broad range of applications and became a considerable field of research. As examples, this approach was used to sequence the genetic code of a mammoth [39], a Neanderthal genomic DNA [40], the human distal gut

microbiome ([41,42]), obese and lean mouse gut microbiomes [4] and to construct a xenobiotic gene [43]. But the principal utilization of metagenomic approaches was probably to sequence environmental DNA, and metagenomes were generated from oceans (in particular the global ocean survey project, [44,45]), deep Mediterranean [46], coral atolls [47], the Peru Margin subseafloor biosphere [48], north pacific deep ocean [49], hypersaline sediments [50], arctic snows [51], and polluted air [52]. Each metagenome provides new information about specific microbial communities and increases the global knowledge about microbial functional potential in the environment. One major advantage of the metagenomic approach is the ability to explore a new environment without detailed hypotheses. Thus, unexpected results can emerge from metagenomic surveys (e.g., important presence of rhodopsin-like photoreceptors in oceans). Another recent example is the proposed and unanticipated characterization of three distinct enterotypes in human faecal metagenomes [53].

In some cases, DNA extraction yield is not compatible with the appetite of sequencing technologies. As an alternative, some researchers use a PCR step called whole genome amplification (WGA) via multiple displacement amplification (MDA) to increase their DNA concentration. The originality of this PCR is the utilization of small primers that possess the potential to fix almost any part of genomes and so to amplify the entire DNA pool randomly. This technology was used to produce shotgun libraries [54 114-6, 118], to generate metagenomes from marine viral communities [55] and to access microbial populations in contaminated sediments [56] or from coral atolls [47]. This “magical” step which can also create DNA from nothing (personal experiments) was presented as a power tool for microbial ecologists [57]. MDA kits have been shown to compromise quantitative analysis of metagenomes [58] and these results might limit the use of this technique in the future if the inherent errors cannot be corrected. More recently, MDA kits were also used to amplify single cells selected from the environment without any culture step [59,60]. This approach aids the assembly of parts of a metagenome that is considerably limited in complex microbial communities when sequencing randomly DNA fragments extracted and purified.

Due to two distinct approaches for studying metagenomes (clone library and functional screening versus direct sequencing), the description of metagenomics has evolved. The definition of Schloss and Handelsman [61]: “metagenomics is the culture-independent analysis of a mixture of microbial genomes (termed the metagenome) using an approach based either on expression or on sequencing” reflects this domain of research and accepts that the mixture does not necessarily represent in situ communities quantitatively. Recent studies have demonstrated considerable biases during DNA extraction approaches (e.g., in soils; [62]), which modify the sequence distribution of metagenomic studies. As a consequence, conclusions might be misleading due to the presence of false positives. However, this definition of metagenomics excludes microarray technologies while a majority of researchers using this technique believe that microarrays exploit metagenomic approaches. So this definition could evolve. Another possibility would be “Metagenomic

approaches are the study of DNA pools extracted directly from the environment and corresponding to a mixture of microbial genomes termed the metagenome". But in this case, fingerprint techniques are also included in this definition. If fingerprint approaches are not part of metagenomics, "metagenomic approaches are the expression, sequencing or hybridization of DNA pools extracted directly from the environment and corresponding to a mixture of microbial genomes termed the metagenome". In the near future, technologies will probably improve the assembly of a considerable number of genomes directly from the environment. At that time, thousands of genomes will be assembled and their evolution followed (e.g., gene duplication or deletion, acquisition of mobile genetic elements) during controlled environmental modifications (in situ or in microcosms). As a consequence, the currently defined metagenomic field represents just the beginning of the evolution from the study of cultivated to the study of uncultivated microorganisms.

Since the utilization of shotgun sequencing and, more recently, high throughput sequencing technologies [63,64] to generate more and more high-sequence-read metagenomes, bioinformatics tools were developed to help microbiologists studying these datasets. Some of these tools include MetaGene [65], Uniprot [66-69], MEGAN [70-72], CAMERA [73], IMG/M [74] and MG-RAST [75], where tables can be exported to STAMP [76] for statistical analyses, MetaTISA [77], Orphelia [78], ShotgunFunctionalizeR [79], WebCARMA [80], and METAREP [81]. While still not perfect (e.g., impact of the e-value threshold on the annotation, Delmont et al., in press), these tools are now largely used by microbial ecologists to decrypt environmental microbial populations.

In spite of considerable advances in metagenomics since 1998, there are important limits to studying a complete environmental metagenome. First and because of important DNA extraction biases, the metagenomic DNA sample that is the base of any metagenomic study is biased. This limit can be alleviated in part by varying DNA extraction protocols to create a global picture of a given environment, although a "true" picture is not yet possible [62]. Secondly, despite important sequencing technology advances [63,64] and productive studies (e.g. as described in [42,82-84]), the majority of published studies were unable to sequence the entire nucleic diversity presents in a given DNA sample. Rarefaction curves are often used to estimate both taxonomical and functional diversities present in a DNA sample (e.g. [22,38]); unfortunately, because of the unequal species proportion and the lack of information about biosphere richness, magnitude curves are possibly biased and could underestimate true environmental microbial diversity. Finally, after extracting and sequencing a part (which can be relatively small for biodiverse environments like soils and sediments) of a metagenome, annotation processes are used to assign a function and a potential bacterium to the sequences generated. However, databases used for this annotation are still limited and appear to possess unequal environmental information [85]. Besides these common limitations, some environments possess additional difficulties for a complete study of the communities. Nucleic diversity varies strongly between ecological niches and is known to be relatively limited in animal bodies [42] and quite diverse in soils

and sediments [86]. Others parameters like species evenness [87], metagenomic stability as a function of time [88] and environmental heterogeneity (e.g. in soil) can increase the difficulties to study a complete metagenome.

Nevertheless, with the emergence of important technological advances and various tools to study complex DNA samples [89], productive studies were recently performed and relatively large metagenomic datasets were generated. Moreover, due to public databases availability on the web (e.g., http://metagenomics.nmpdr.org/), the comparison of annotated sequences from various metagenomes can be performed. Using bioinformatic tools and metagenomes corresponding to various already sequenced environments, functional potentials of the different microbial communities can be compared to determine their specific life styles. This type of comparison was done first by Tringe and colleagues in 2005 [38], followed by Dinsdale and colleages [90] and Willner and colleages [91] with larger datasets. However, in contRAST to the quantity of generated metagenomes (e.g., more than 30 000 private metagenomes on MG RAST in 2011), the number of publications treating global metagenomic comparisons is noticeably limited. This report can easily be explain by the fact that all metagenomic projects possess specific approaches without general inter-projects standards and the compilation of metagenomic comparisons when different DNA extraction approaches and sequencing technologies are used had not been statistically demonstrated until recently [92]. These data can be used for specific objectives such asvemphasizing peculiarities of different metagenomes, tracking a function or a species among environments, finding correlations between functions or species, selecting all sequences related to a gene of interest in several environments, etc.

The sequencing technology used influences the percentage of annotated sequences [92]. The longer the sequences generated, the larger the percentage of sequences annotated. In addition, when the proportion of eukaryotic sequences increases, the percentage of annotated sequences decreases considerably. However, when the functional and taxonomical distributions as a function of this percentage among the different metagenomes are normalized, these problems are alleviated and the comparison of metagenomes corresponding to various projects even if standards are not present can provide some useful information. Different DNA extraction protocols are used in different projects, but standards cannot yet be imposed as the approach that provides the DNA pool closest to the reality is currently unknown. As a perspective to limit metagenomic biases, methodological fluctuations could be integrated in metagenomic surveys to limit the proportion of false positive results {Delmont et al., commentary submitted in ISMEj}.

In addition, biases in phylogenetic annotation systems (e.g. RDP, Greengenes, SEED) used to comparing phylotype distributions in a metagenome influences the final conclusions. When comparing the same species among different metagenomes, the distribution is also influenced by the data used, but the differences in phylogenetic distribution observed are less biased (if DNA extraction biases are not taken into account). So global comparisons by-

pass an important source of error present when species distribution is compared within one metagenome. So while whether the phylogenetic group “a” is more or less represented than the phylogenetic group “b” in one metagenome “X” is dependent on the annotation system, the relative proportion of the two groups between two metagenomes X and Y can be made. This represents a considerable advantage in comparison to 16S rRNA genes amplification and sequencing where quantitative comparisons are not possible due to amplification biases [93]. This example is also applicable for functions and emphasizes the interest of global metagenomic comparisons. Comparing metagenomes from specific points around the word provides information concerning the distribution of the 10^{30} bacteria and archaea present on earth.

With the goal of stimulating the study of sequencing data and to visualize microbial and gene distributions and specificities as a function of different environments, we compared 77 metagenomes corresponding to DNA samples extracted from oceans, coral atolls, deep oceans, Antarctic aquatic environments, Arctic snows, terrestrial environments (hypersaline sediments (corresponding to different depth horizons in millimeters and represented in the graphs from the left (top) to the right [94]), soils, sludges, microbial fuel cells, acid mine drainage biofilms), polluted air and animal microbiomes (human feces, mouse and chicken cecum, and cow rumens) at different taxonomical and functional levels. The metagenomic datasets used were generated at different times, by different groups and using different methods. However, they represent the major metagenomic datasets made public between 2004 and 2011 and for which at least two datasets are available for each environment. Up to 293 genera, 599 species or strains, and 487 functional subsystems vary in their relative distribution between the selected environments (based on the ANOVA test and Bonferroni correction) and provide information about microbial life styles.

With the current sequencing revolution, the rapid increase in ecosystem sequencing projects and new bioinformatic tools, we are probably entering in a new gold age of microbiology with global metagenomic comparisons as a symbol of success (observations without any a priori before hypotheses elaborations). However, microbial ecologists need to apply ecological concepts to metagenomics in order to refine our perception their role in different ecosystems, their interactions and impact on climate and human activities.

Methods summary:

For all the public and private metagenomes compared in this study (see [92] for more information about these datasets), data tables corresponding to sequences showing similarities to known species (phylum, genera and species levels) and functions (Subsystem Hierarchy 1 and level 3) were exported from the Metagenomics SEED Viewer version 2.0 (http://metagenomics.nmpdr.org/). Only hits with an E-value < 10^{-5} were considered to be significant. Moreover, for the phylogenetic profiles, only the SEED annotation system which

provides taxonomical diversity tables from all the sequences annotated was used and the limited 16S rRNA gene sequences in metagenomic sequences were not used due to statistical limitations. This approach is influenced by the phylogeny and the low number of already sequenced genomes. In addition, metagenomic data matching these genomes can correspond both to their actual sequences and/or to genomes possessing the same genes with a high degree of similarity to the genes in the completely sequences genomes. However, this approach is significantly more sensitive than when using only sequences related to 16S rRNA genes. In addition, we assume that it is of greater interest to study relations between metagenomic data and already sequenced environments than to study sequences related to a single amplified gene even if this gene has been extensively characterized and sequenced (e.g. RDP project, http://rdp.cme.msu.edu/). The metagenomic comparisons (taxonomical and functional distributions and principal component analyses) were performed on relative proportion (percentage) of target sequences compared to the total annotated sequences [92]. For taxonomical comparisons at the species level, distribution graphs correspond to the sum of sequences from the completely sequenced genomes as well as from all genomes possessing similar genes or genomic regions. Statistical calculations (ANOVA and Bonferroni correction) of probability values for the null hypotheses were performed using STAMP [76]. Finally, principal component analyses were done with the r software and ade4 package [95].

Results:

GC content and relative distribution of both taxa and functions were compared at different levels for 77 metagenomes corresponding to 15 distinct environments.

GC content distribution:

The distribution of G+C content was compared in the 77 metagenomes. The G+C content can represent a fingerprint of microbial communities and provides information about environmental specificities. Microbial communities from ocean, humans, and soil, the three biomes the most common in sequence databases, appear to possess respectively a low, a medium and a high G+C content (Figure 1, panel A). These observations are due in part to considerable community structure differences (described below).

Interestingly, when comparing the distribution of sequences related to specific subsystems (here the subsystem shown corresponds to ATP synthase), the G+C content distribution is similar to that for the entire datasets (Figure 1, panel B). So these observations are not only due to functional distribution peculiarities as the G+C content of related genes varied considerably between environments for the same function. As a consequence, annotation

processes can be biased if databases do not represent a range of sequences with different G+C content for each gene or subsystem.

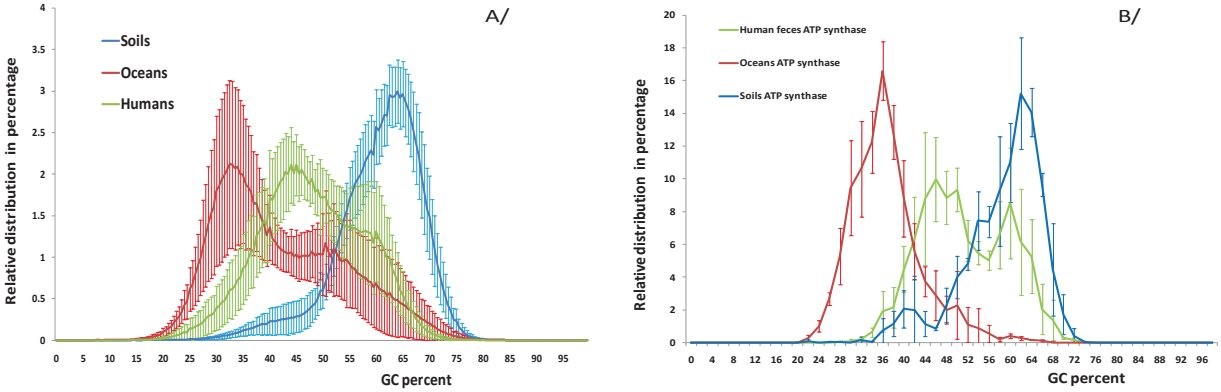


Figure 1. Plot of relative metagenomic sequence read abundance as a function of G+C content for all the metagenomic data pooled for each microbial ecosystem, oceans, humans, and soil. Error bars are due to variance among the different datasets for the same ecosystem. Panel A is the distribution for the entire dataset and Panel B is the distribution for the reads assigned to the metabolic subsystem ATP synthase.

When comparing the G+C content distribution of the 77 metagenomes, a graph such as found in figure 1 becomes a jumble of overlapping data. However, the different datasets can be analyzed by principal component analysis (PCA) where the G+C content is compared between different ecosystems (Figure 2). For example, metagenomic sequences from Bacteria and Archaea in soil, polluted indoor air, sediments, microbial fuel cells (MFCs) and sludges have a high G+C content, and therefore, are grouped separately (to the right) on the PCA (Figure 2). In contRAST, human, mouse and acid mine drainage biofilm microbial populations possess a medium G+C content and are found grouped in the lower part of the PCA. These differences in G+C content profiles reflect differences in community structure between ecosystems. However, the primary cause of these differences is still a mystery and to understand why a high GC percent would be an advantage for soil microbial populations (e.g., temperature effects?) would aid in future microbial community studies.

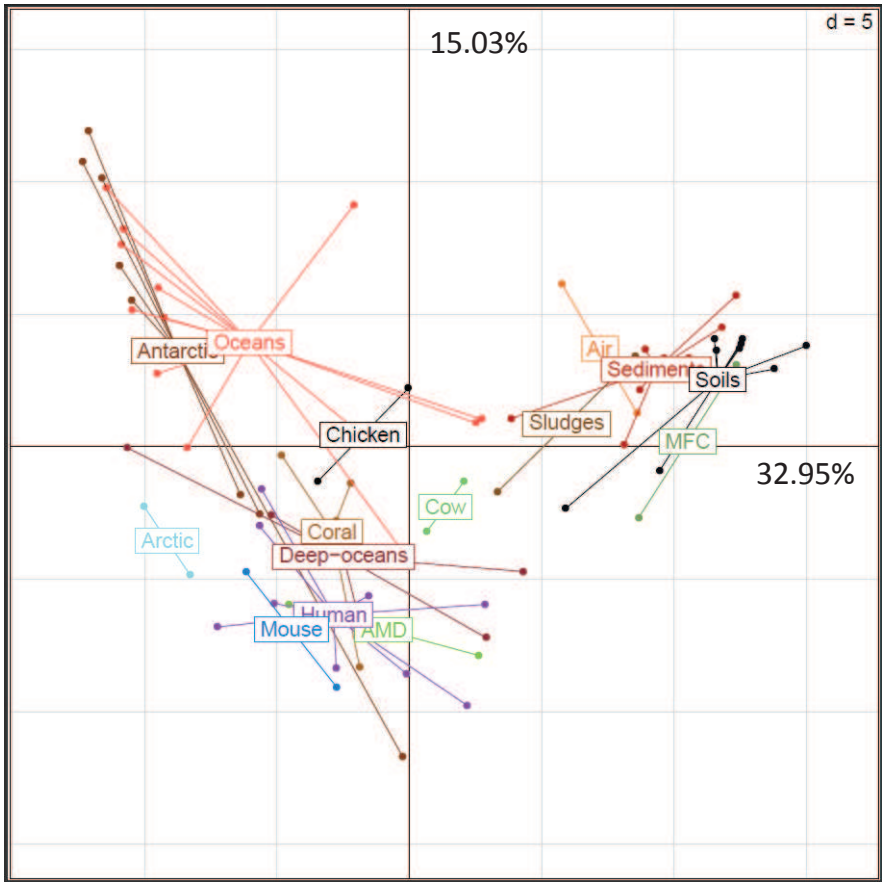


Figure 2. Principal component analysis of the G+C content profile of the microbial metagenomic reads from different ecosystems. AMD: acid mine drainage, MFC: microbial fuel cell.

Taxonomical comparison:

The 10^{30} Bacteria and Archaea estimated to exist on earth [3] are taxonomically grouped in 22 bacterial and 5 archaeal phyla, which are then subdivided in class, order, family, genus, species and finally strains. The metagenomic data from these different ecosystems can be used to describe potential differences in community structure (and function) between them. Here we compared the distribution of microbial communities at different levels (from phylum to species) for the 77 metagenomes selected and corresponding to 15 distinct environments without any specific 16S rRNA gene amplification.

Principal phyla

To highlight the disparity in microbial communities in the different ecosystems at a general level, the distribution of principal bacterial and archaeal phyla was compared (Figure 3 and 4). These phyla are clearly unevenly represented in the 77 metagenomes, but tend to be relatively stable within a given environment. In other words, the variation of their distribution between different published metagenomic data from different researchers for the same ecosystem is less than the differences between ecosystems.

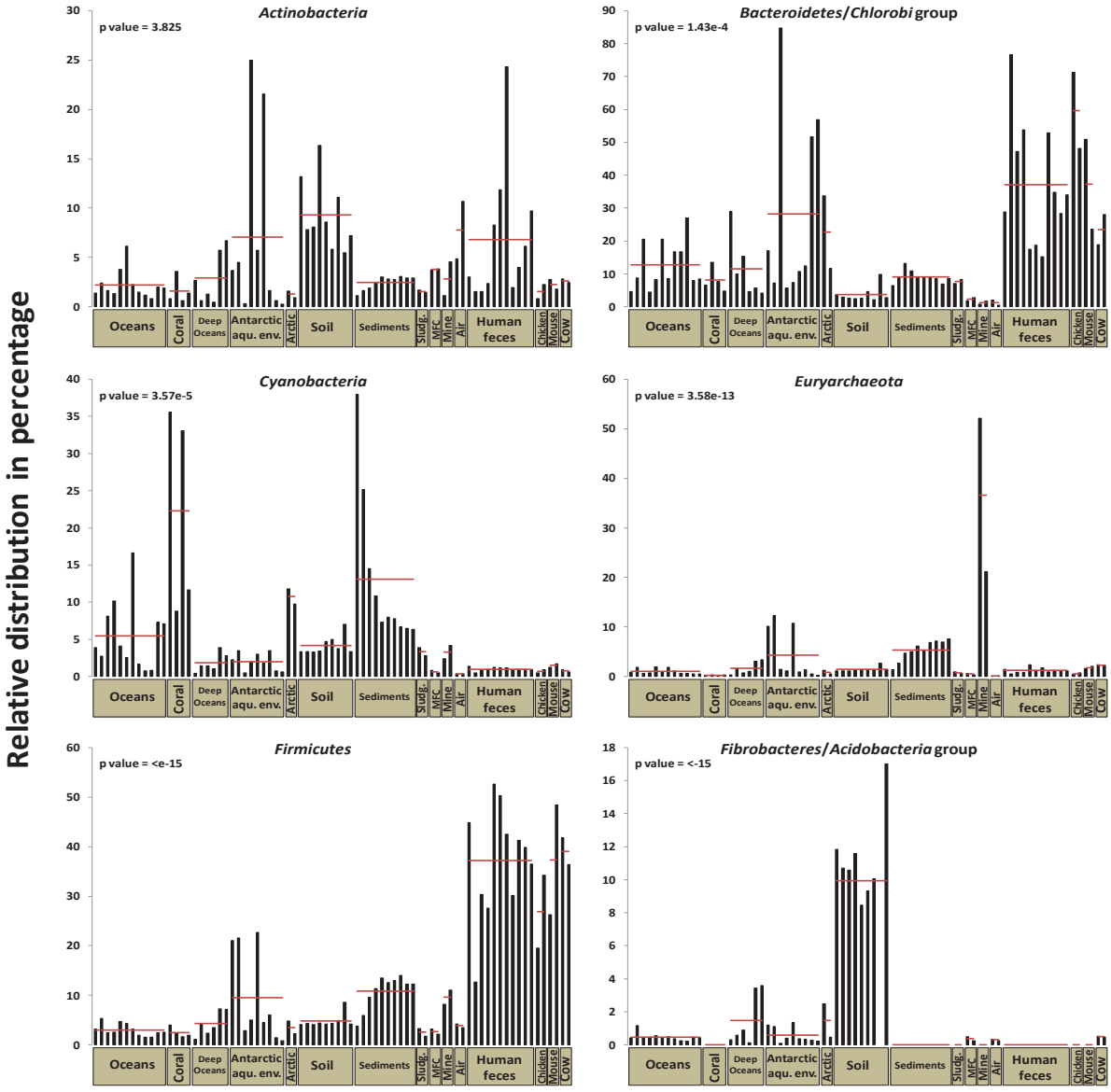


Figure 3. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

As an example, the Fibrobacteres/Acidobacteria group is found more commonly in soils (more than 10%) in comparison to the other environments (undetected in sediments and human feces, and a maximum of 3.6% in deep oceans) even if its distribution is relatively low in one soil (From Waseca farm, [38]). On the other hand, Firmicutes and the group Bacteroidetes/Chlorobi are present more in animals in contRAST to oceans. Cyanobacteria are important in the hypersaline sediments (but decreasingly so with depth) and atoll corals, but also in Arctic snows and some metagenomes extracted from oceans. Interestingly, the phylum Euryarchaeota is largely presents in acid mine drainage biofilms in comparison to the

other environments. Finally, the phylum Actinobacteria, known to possess important secondary metabolites and bio-actives molecules like antibiotics (e.g., [96]) is highly present in soil (up to 15% of the species present in some cases) but is also commonly detected in Antarctic aquatic environments and human feces. The biodiversity of Antarctic aquatic environments, which are actually underexplored and as a consequence underexploited, could be an important source of new antibiotic discovery in the future.

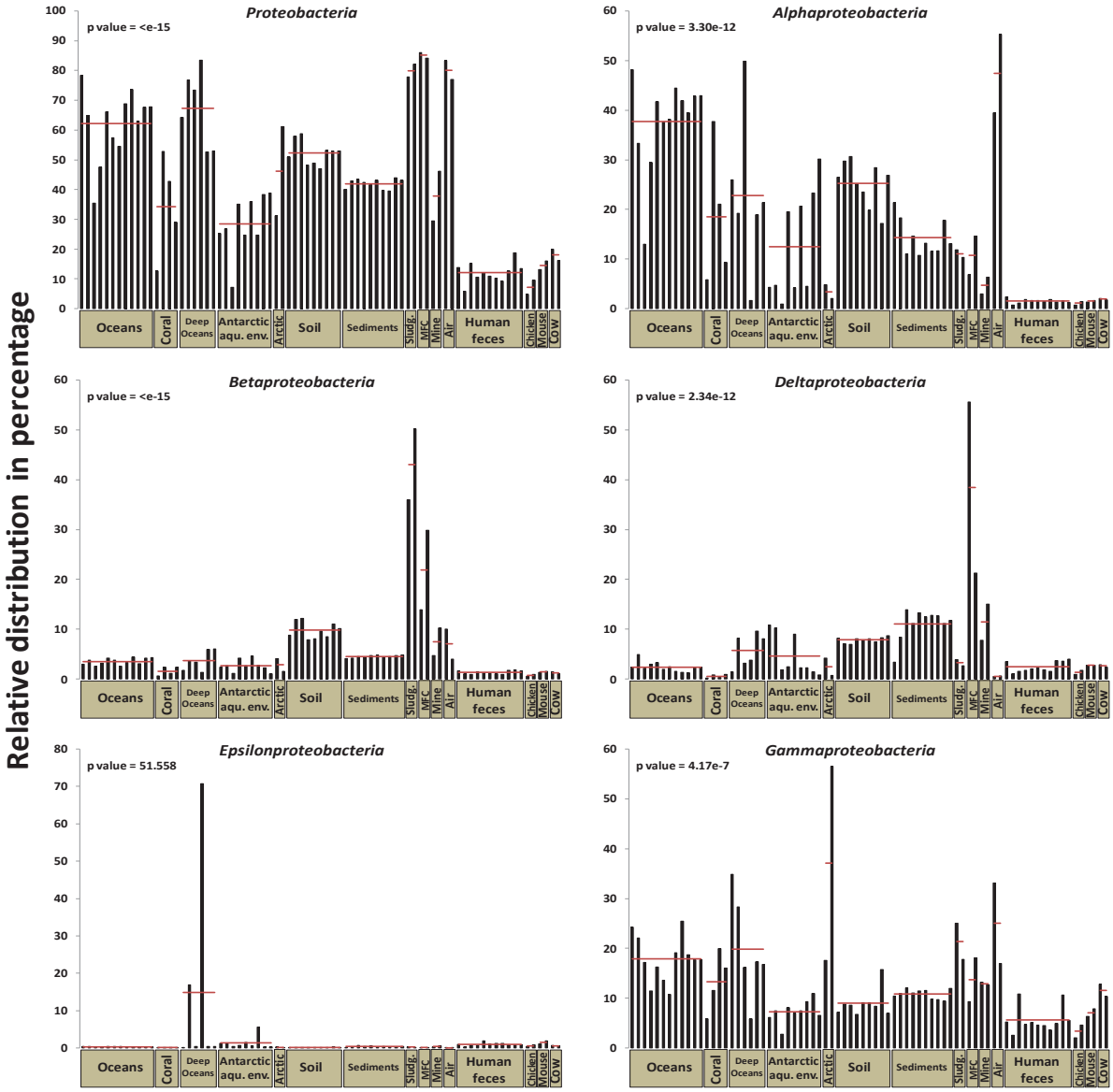


Figure 4. Relative distribution (in percentage of annotated reads) of Proteobacteria (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the differences observed are random.

The phylum Proteobacteria is present in all the metagenomes, but its distribution varies as a function of the environments (Figure 4). Indeed, it is prevalent in oceans, deep oceans, sludges, microbial fuel cells and polluted air, but not in animal or human microbiomes. Moreover, the different classes and sub-classes within Proteobacteria are unevenly represented. Alphaproteobacteria are present more frequently in oceans (an average of 38%) and polluted air, are quite variable in deep oceans (between 1.7 and 50%) and are relatively rare in Arctic snows. Betaproteobacteria are prevalent in sludges (between 36 and 51%) in comparison to the other environments (e.g. average of 3.6% in ocean). On the other hand, Deltaproteobacteria are present more in sediments, acid mine drainage biofilms, soil and in microbial fuel cells (Figure 4). Moreover, Gammaproteobacteria and Epsilonproteobacteria dominate two metagenomes extracted from Arctic snow (56.5%) and a hydrothermal vent from a deep ocean (70.1%, Grzymski et al., 2008).

In addition, the distribution of less represented phyla was compared (supplement data). Fusobacteria is more present in animal microbial populations. Korarchaeota is relatively more represented in aquatic environments and especially in Antarctic aquatic environments than in others. Planctomycetes is more highly represented in soil, sediments, and some deep oceans than other environments. Finally, the distribution of Aquificae is relatively similar in the different metagenomes, excepted in the acid mine drainage biofilms where it is relatively more highly represented.

Principal genera

The total phylogenetic diversity detected in the 77 metagenomes includes 1590 sequenced species and 687 genera. Interestingly, some of these species and genus appear to be strongly unequally represented in the different ecosystems. For example, the genera Bacteroides is mostly present in animals (From 13.66 to 72.39% as a function of the individuals) and in low proportion in the other environments (e.g. 0.1% in soil) (Table 1). Moreover, Ferroplasma, Prosthecochloris and Sulfurovum are detected in relatively high proportions in acid mine drainage biofilms, an Antarctic aquatic ecosystem, and a hydrothermal vent, respectively, but are almost not present in the other metagenomes.

Oceans	Mean (%)	SD (%)	Coral atolls	Mean (%)	SD (%)	Deep oceans	Mean (%)	SD (%)
Candidatus-Pelagibacter	13.66	8.11	*Prochlorococcus*	9.01	8.49	*Sulfurovum*	9.75	19.93
Roseobacter	3.20	2.33	*Synechococcus*	6.38	4.62	*Thiomicrospira*	3.04	4.08
Prochlorococcus	2.59	3.78	*Alteromonas*	6.11	4.90	*Pseudoalteromonas*	2.73	4.69
Flavobacterium	2.37	1.27	*Candidatus-Pelagibacter*	5.27	4.49	*Psychrobacter*	2.19	4.31
Silicibacter	2.18	1.28	*Arabidopsis*	4.07	5.69	*Flavobacterium*	2.05	1.84
Roseovarius	2.17	1.37	*Pirellula*	2.09	2.04	*Pseudomonas*	2.05	1.05
Synechococcus	2.01	1.80	*Odontella*	2.06	2.27	*Magnetospirillum*	1.77	1.17
Pseudomonas	1.82	0.49	*Silicibacter*	2.01	1.84	*Maricaulis*	1.60	3.18
Flavobacteriales	1.65	0.90	*Thermosynechococcus*	1.86	1.95	*Roseobacter*	1.48	1.37
Tenacibaculum	1.58	1.65	*Cyanidioschyzon*	1.77	2.02	*Tenacibaculum*	1.47	1.84

Antarctic lakes	Mean (%)	SD (%)	Arctic snows	Mean (%)	SD (%)	Soils	Mean (%)	SD (%)
Prosthecochloris	10.55	23.37	*Pseudoalteromonas*	16.23	11.92	*Bradyrhizobium*	6.90	1.89
Candidatus-Pelagibacter	6.94	8.65	*Flavobacterium*	7.83	4.79	*Solibacter*	6.29	2.40
Polaribacter	5.97	10.89	*Nostoc*	6.55	0.94	*Blastopirellula*	4.42	2.74
Tenacibaculum	2.93	5.08	*Psychrobacter*	6.10	5.01	*Acidobacteria*	3.66	1.95
Clostridium	2.30	2.23	*Cytophaga*	3.64	2.43	*Rhodopseudomonas*	3.34	0.90
Flavobacterium	2.24	1.98	*Bacteroides*	2.17	1.33	*Burkholderia*	2.43	0.19
Paramecium	2.07	2.79	*Shewanella*	2.11	0.95	*Pirellula*	2.37	1.28
Acanthamoeba	1.88	3.43	*Psychromonas*	2.09	1.23	*Nitrobacter*	2.21	0.53
marine actinobacterium	1.24	1.89	*Flavobacteriales*	1.97	0.05	*Mycobacterium*	1.95	0.86
Thermoanaerobacter	1.19	1.43	*Anabaena*	1.60	0.14	*Geobacter*	1.92	1.06

Sediments	Mean (%)	SD (%)	Sludges	Mean (%)	SD (%)	MFC	Mean (%)	SD (%)
Nostoc	4.98	4.62	*Dechloromonas*	20.68	4.40	*Geobacter*	34.15	15.49
Geobacter	4.90	1.29	*Azoarcus*	6.09	1.25	*Pseudomonas*	7.64	4.02
Chloroflexus	4.02	1.30	*Xanthomonas*	4.90	0.99	*Acidovorax*	6.55	3.62
Pirellula	2.77	0.62	*Ralstonia*	3.41	0.25	*Delftia*	6.38	1.88
Bacteroides	2.76	0.88	*Pseudomonas*	3.16	0.55	*Xenopus*	2.75	1.89
Blastopirellula	2.49	0.98	*Burkholderia*	2.81	0.25	*Agrobacterium*	1.67	1.17
Methanosarcina	2.16	0.70	*Rhodoferax*	2.70	0.50	*Desulfuromonas*	1.62	0.72
Clostridium	2.14	0.64	*Polaromonas*	2.07	0.42	*Rhizobium*	1.55	1.01
Desulfovibrio	2.11	0.59	*Geobacter*	1.88	0.28	*Burkholderia*	1.33	0.38
Bacillus	2.07	0.55	*Thiobacillus*	1.87	0.09	*Aeromonas*	1.00	0.31

Acid mine	Mean (%)	SD (%)	Polluted air	Mean (%)	SD (%)	Human feces	Mean (%)	SD (%)
Ferroplasma	21.01	17.10	*Caulobacter*	30.30	4.78	*Bacteroides*	33.60	17.07
Thermoplasma	8.85	0.91	*Stenotrophomonas*	13.01	4.52	*Clostridium*	15.00	4.36
Geobacter	6.90	2.22	*Streptococcus*	3.12	0.32	*Bifidobacterium*	5.59	6.17
Erwinia	4.43	2.83	*Erwinia*	3.03	1.73	*Bacillus*	4.26	1.33
Picrophilus	3.31	0.08	*Xanthomonas*	2.89	1.49	*Desulfitobacterium*	3.86	1.07
Staphylococcus	1.86	0.95	*Sphingomonas*	2.10	0.43	*Enterococcus*	2.86	0.85
Burkholderia	1.70	0.66	*Pseudomonas*	1.89	0.21	*Thermoanaerobacter*	2.42	0.79
Methanosarcina	1.59	0.04	*Delftia*	1.78	1.27	*Streptococcus*	2.06	0.68
Pelobacter	1.48	0.44	*Sphingopyxis*	1.59	0.05	*Porphyromonas*	1.29	0.66
Sulfolobus	1.46	0.12	*Burkholderia*	1.43	0.25	*Fusobacterium*	1.21	0.34

Chicken cecum	Mean (%)	SD (%)	Mouse cecum	Mean (%)	SD (%)	Cow rumen	Mean (%)	SD (%)
Bacteroides	56.40	11.72	*Bacteroides*	29.06	11.00	*Clostridium*	20.28	1.09
Clostridium	9.08	2.71	*Clostridium*	11.93	4.91	*Bacteroides*	17.04	3.16
Lactobacillus	3.29	0.12	*Bacillus*	5.92	2.25	*Psychrobacter*	6.93	0.83
Desulfitobacterium	2.56	0.78	*Porphyromonas*	4.61	1.89	*Parabacteroides*	5.12	0.98
Bacillus	2.46	0.73	*Lactobacillus*	3.85	1.09	*Bacillus*	2.44	0.16
Enterococcus	2.02	0.74	*Desulfitobacterium*	3.62	1.26	*Alkaliphilus*	2.37	0.25
Streptococcus	1.81	0.69	*Enterococcus*	3.08	1.21	*Streptococcus*	2.27	0.08
Porphyromonas	1.57	0.06	*Escherichia*	2.15	0.23	*Porphyromonas*	1.82	0.35
Thermoanaerobacter	1.17	0.31	*Streptococcus*	2.05	0.57	*Thermoanaerobacter*	1.81	0.25
Staphylococcus	0.93	0.34	*Thermoanaerobacter*	1.66	0.49	*Desulfitobacterium*	1.78	0.09

Table 1: List of the 10 most represented genera (relative proportion in percentage) for 15 distinct ecosystems. Means and standard deviations calculated from all the metagenomic datasets from the same environment.

The 11 non-animal associated microbiomes appear to be quite different from each other with different genera and evenness, while the microbiomes from human feces, chicken cecum, mouse ceacum and cow rumen appear to be similar with the genus Bacteroides, Clostridium, Bacillus, Desulfitobacterium, Thermoanaerobacter, Streptococcus and Porphyromonas dominating these four animal microbiomes (Table 1). However, the species corresponding to these genera are different in the different animal microbiotas (see the animal section below).

While comparing principal genera provides general information about ecosystem differences, these comparisons do not necessarily help understand specific functional characteristics associated with minor but critical members of the microbial community. Thus, in spite of their relatively low proportion, the distribution of some genera and species of economical, clinical, or environmental interest were compared among the 77 metagenomes.

Pathogenic microorganisms

To highlight the importance of pandemic factors and human health risks, genera and species associated with some illnesses were tracked among the compiled metagenomes (Figure 5). Since the SEED annotation system was used, sequences related to one already sequenced genome might actually be from a similar gene in an unknown and/or related species. Based on the E-value criteria, annotated sequences are relatively close to genes from the genomes of reference and so provide crucial information about environmental taxonomy. For example, the causative agent of tetanus disease, Clostridium tetani [97], is detected more frequently in animal and human microbial populations. On the other hand, Clostridium botulinum appears to be more present in the environment and in particular in Antarctic aquatic environments although it is well represented in cow rumens. Pseudomonas aeruginosa is more detected in sludges and sediments and sequences related to Yersinia pestis and Pseudotuberculosis are commonly detected in Antarctic metagenomes. Finally, sequences related to Listeria monocytogenes EGD-e are more detected in animals except for cow rumens (Figure 5).

However, the genes of some of the toxic proteins involved in these diseases (Tetx, BoNT, ToxA, YPM, CTB, and Ymt) were not detected by local blasts (BLASTall software). This might demonstrate the difficulty in correlating pathogenic islands and known pathogens in the environment. In addition to exploring pandemic factors, genes coding for prions and H1N1 proteins were BLASTed but were also not detected in the metagenomes. Interestingly, HIV sequences were found in two environments: ocean and cow rumen (see animal section). This is the first case of HIV sequences being detected in the environmental metagenomes. HIV sequences were also tracked in the environmental NCBI database (nucleotide blast using environmental samples and whole genome shotgun reads). Genetic fragments corresponding to this retrovirus (up to 600 nt in length) were detected in freshwater, termite gut, human mouth, and even in Neanderthal fossil metagenomes, highlighting an unexpected possible presence of this virus or its ancestor in the environment.

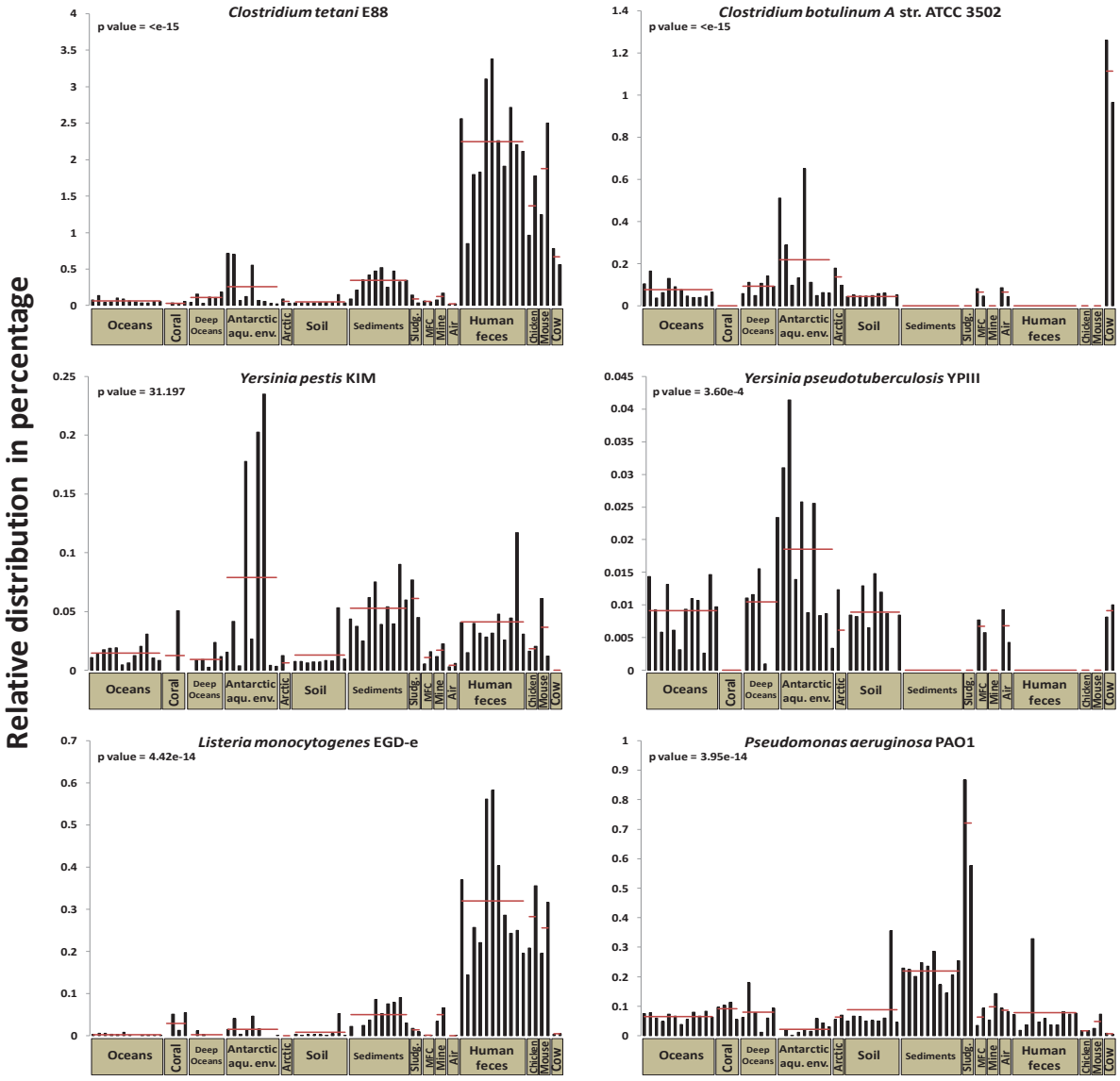


Figure 5. Relative distribution (in percentage of annotated reads) of selected pathogens (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Species of economical or environmental interest

The distribution of few species of environmental or economical interest was also compared (Figure 6). The archaeon Methanosarcina acetivorans strain C2A, which generates two green house gases: carbon dioxide and methane [98,99], is detected more in sediments, acid mine drainage biofilms, and in some Antarctic aquatic environments samples than in oceans or soil. Dehalococcoides BAV1, known to be involved in chlorinated compound biodegradation (chlororespiration) with, for example, the capacity to dechlorinate vinyl chloride [100], is

detected more in deep oceans and Antarctic aquatic environments. Geobacter metallireducens, which uses extracellular iron for the electron transport chain [101] and which is involved in aromatic degradation [102], is detected in all the metagenomes, but sequences related to its genome are prevalent in microbial fuel cells. Marinobacter hydrocarbonoclasticus, a hydrocarbon-degrading marine bacterium [103], is commonly detected in aquatic ecosystems, but was also identified in soil.

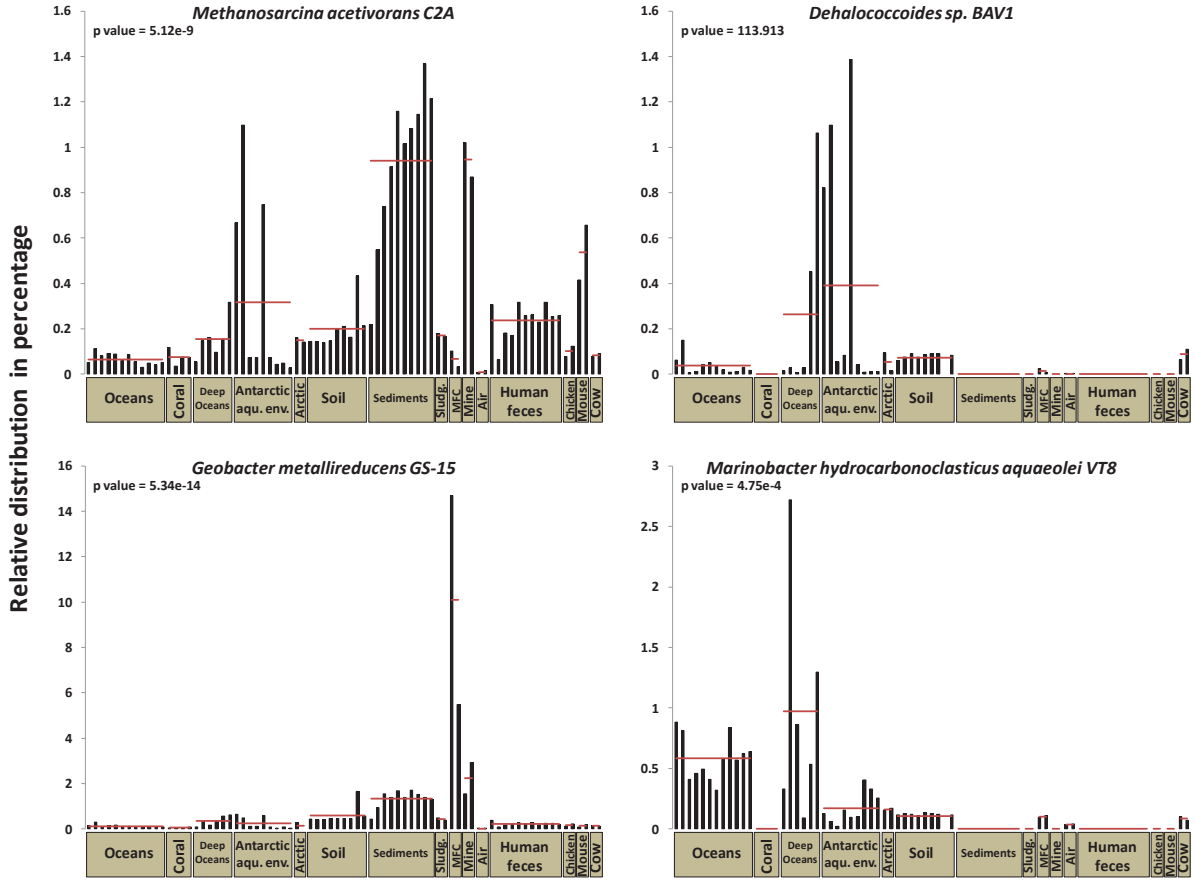


Figure 6. Relative distribution (in percentage of annotated reads) of selected pollutant degraders (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Principal viruses and bacteriophages distribution and consequences

In addition to Bacteria and Archaea, viral DNA fragments can be detected in some metagenomes and provide substantial information about the specificities of the studied metagenomes. An important viral diversity was detected in the 77 metagenomes, but the majority of them are apparently less often detected than bacterial and archaeal organisms (supplement data). However, a few classes of viruses and bacteriophages are unexpectedly distributed in specific environments and can influence functional distributions (Figure 7). The

icosahedral ssDNA bacteriophage family Microviridae, known to possess small genome sizes [104], is prevalent in one ocean metagenome (>8%) and could have an effect on the microbial communities present and their evolution.

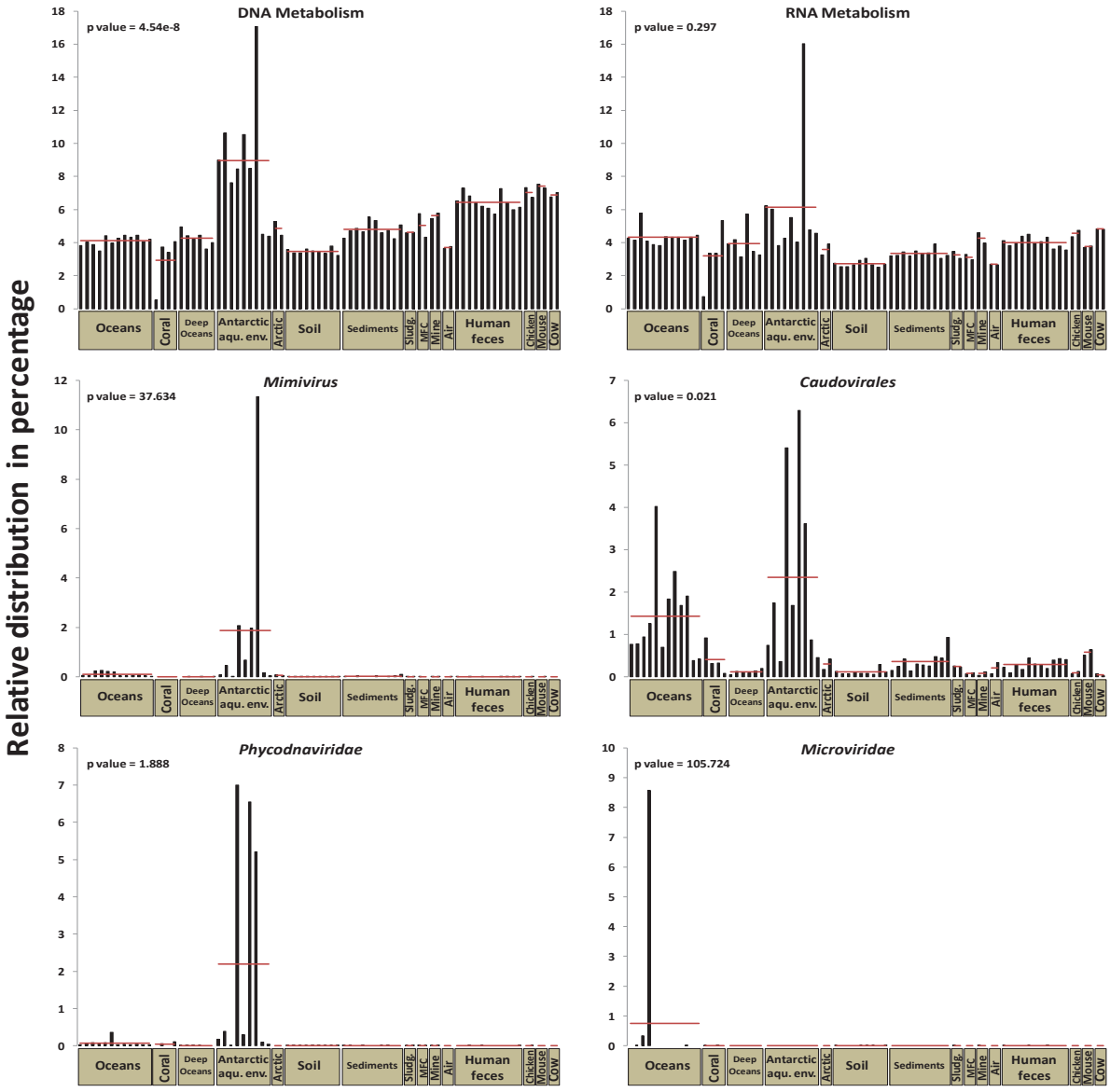


Figure 7. Relative distribution (in percentage of annotated reads) of viruses and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Moreover, the Caudovirales order [105], composed of the tailed bacteriophages, appears to be prevalent in different ocean and Antarctic aquatic metagenomes and can also impact the evolution of the microbial hosts. Interestingly, more than 10% of the sequences annotated in one Antarctic aquatic metagenome correspond to Mimivirus, the largest known virus [106]. Moreover, because of its genetic originality (e.g. the presence of four aminoacyl-tRNA synthetases (RNA metabolism)), this virus stimulated scientific debate about how the

eukaryotic nucleus emerged [107] and its prevalence in one Antarctic aquatic metagenome might provide information about not only how, but also where, this structure emerged. Similarly, other nucleocytoplasmic large DNA viruses are highly present in Antarctic aquatic metagenomes. This is also the case of the families Phycodnaviridae, Poxviridae and Asfarviridae, so highlighting the importance of large eukaryotic DNA viruses in this environment. However, all of these viruses appear to possess a simple and ancient viral ancestor with a small subset of genes [108] and this ancestor possibly emerged in this type of ecosystem. The presence of nucleocytoplasmic large DNA viruses appears to influence the distribution of DNA and RNA metabolism, which is relatively stable in the other environments.

Principal Eukaryote phyla distribution and consequences:

In spite of focusing on bacterial and archaeal communities, Eukaryotic sequences can be found in high proportion in few metagenomes and appear to influence some other community characteristics such as viral functional distributions. Among the 77 metagenomes studied in this work, Eukaryotes were principally found in coral atolls. This was the case of Viridiplantae, Rhodophyta and Stramenopiles which were highly present in these four coral metagenomes (e.g. up to 18% of Viridiplantae in one of them) (Figure 8).

In addition, Alveolata is also highly present in oceans and in particular in one metagenome corresponding to an Antarctic aquatic environment. This same metagenome was found to possess a high proportion of Mimivirus, which is consistent with the relationship between it and amoebae [106]. These Eukaryotic sequences stimulate the distribution of functions related to photosynthesis and respiration in atoll corals. Respiration functional subsystems are highly stable in all the ecosystems except in these atoll corals and some Antarctic aquatic environments where nucleocytoplasmic large DNA viruses are in high proportion.

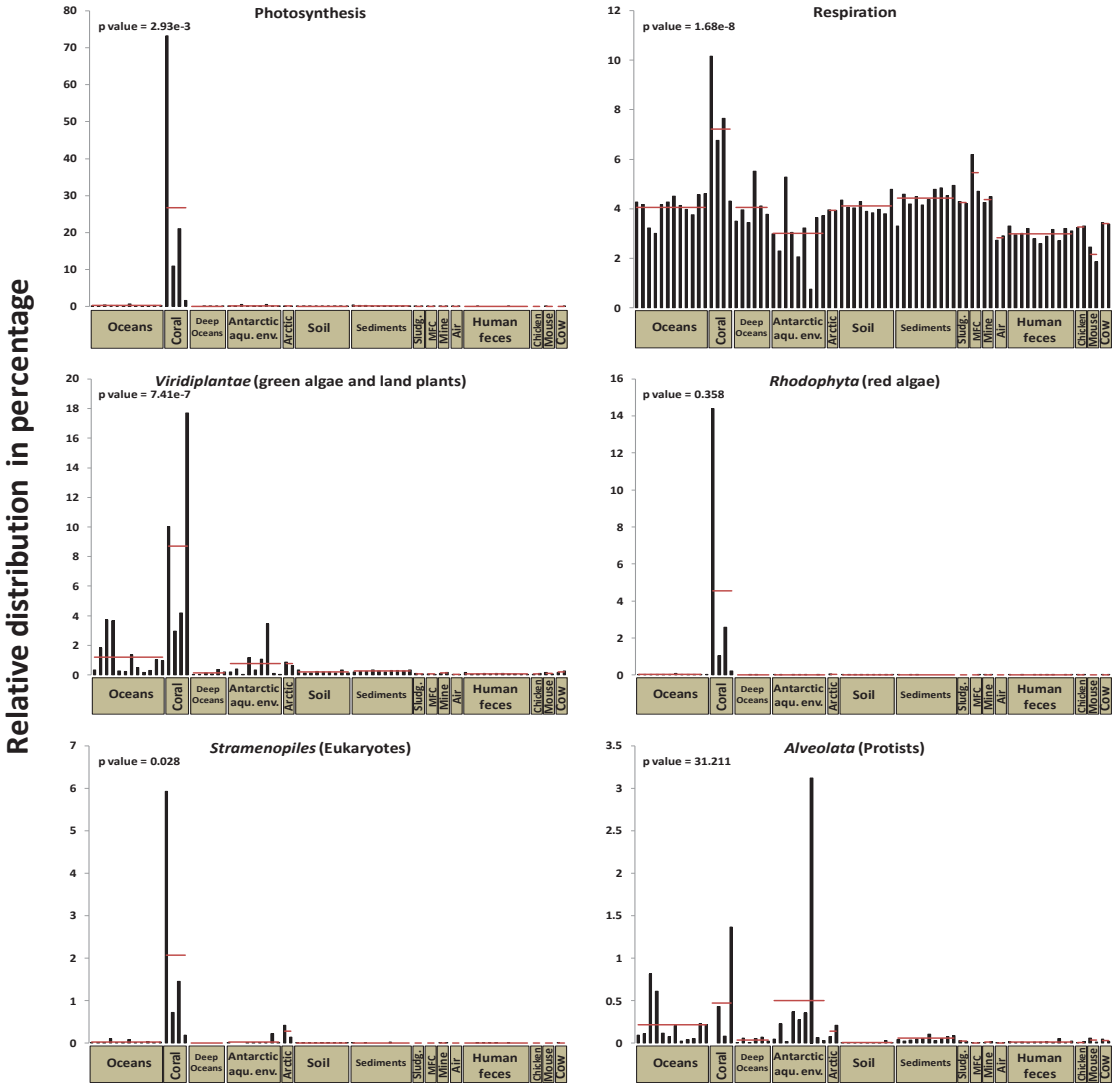


Figure 8. Relative distribution (in percentage of annotated reads) of Eukaryotes and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Functional comparison:

General functions

The global ocean survey provides information about specific sampling points corresponding to the surface of oceans around the world. However, and because the same strategy was applied for a majority of the samples (depth, filters, DNA extraction, sequencing technology) and because microbial communities living in this environment appear to be relatively stable with the actual accessible technologies (here shotgun sequencing), differences between

metagenomes are not important even if the sampling location varied considerably geographically (Figure 9). Differences are limited among coastal habitats, open oceans, and even between the two defined marine environments (coastal versus open oceans).

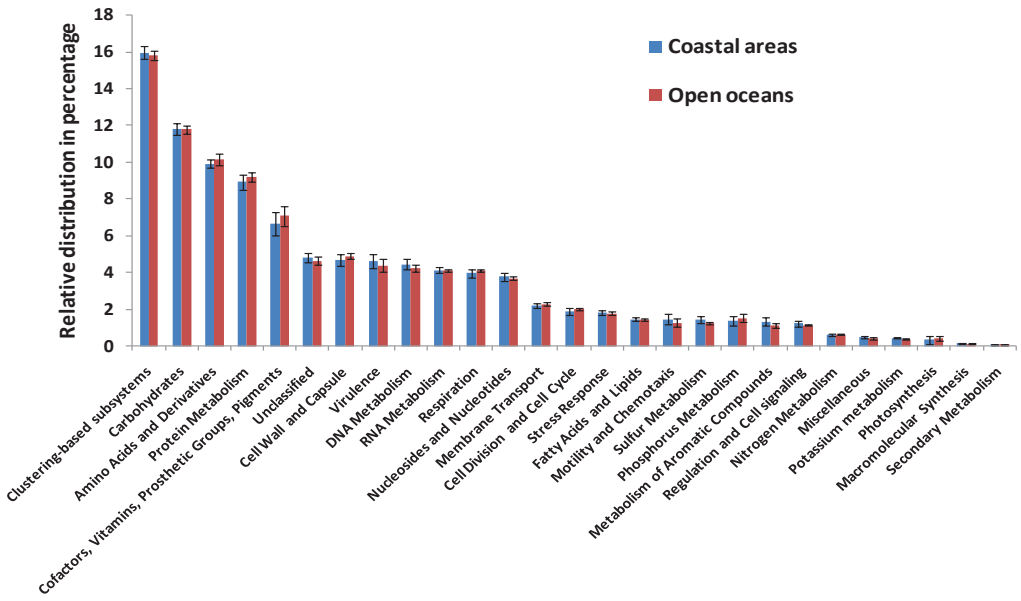


Figure 9. Relative distribution (in percentage of annotated reads) of general functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for 36 metagenomes from the global ocean survey. 18 datasets were selected to represent two sub-environments: coastal area and open oceans. Error bars represent the standard deviation among the 18 metagenomes for the two groups.

However, when comparing these 36 metagenomes from coastal and open oceans in a principal component analysis (Figure 10), differences can be observed, and as an example, sequences related to photosynthesis, cell division and cell cycle, protein metabolism, membrane transport and respiration appear to be more prevalent in open oceans. On the other hand, sequences related to mobility and chemotaxis, sulfur and potassium metabolism, RNA metabolism, regulation and cell signaling, and perhaps more interestingly, metabolism of aromatic compounds and virulence are more detected in coastal environments. However, these differences are limited (see figure 9) and these results would need to be confirmed in individual cases by other approaches (e.g., quantitative PCR).

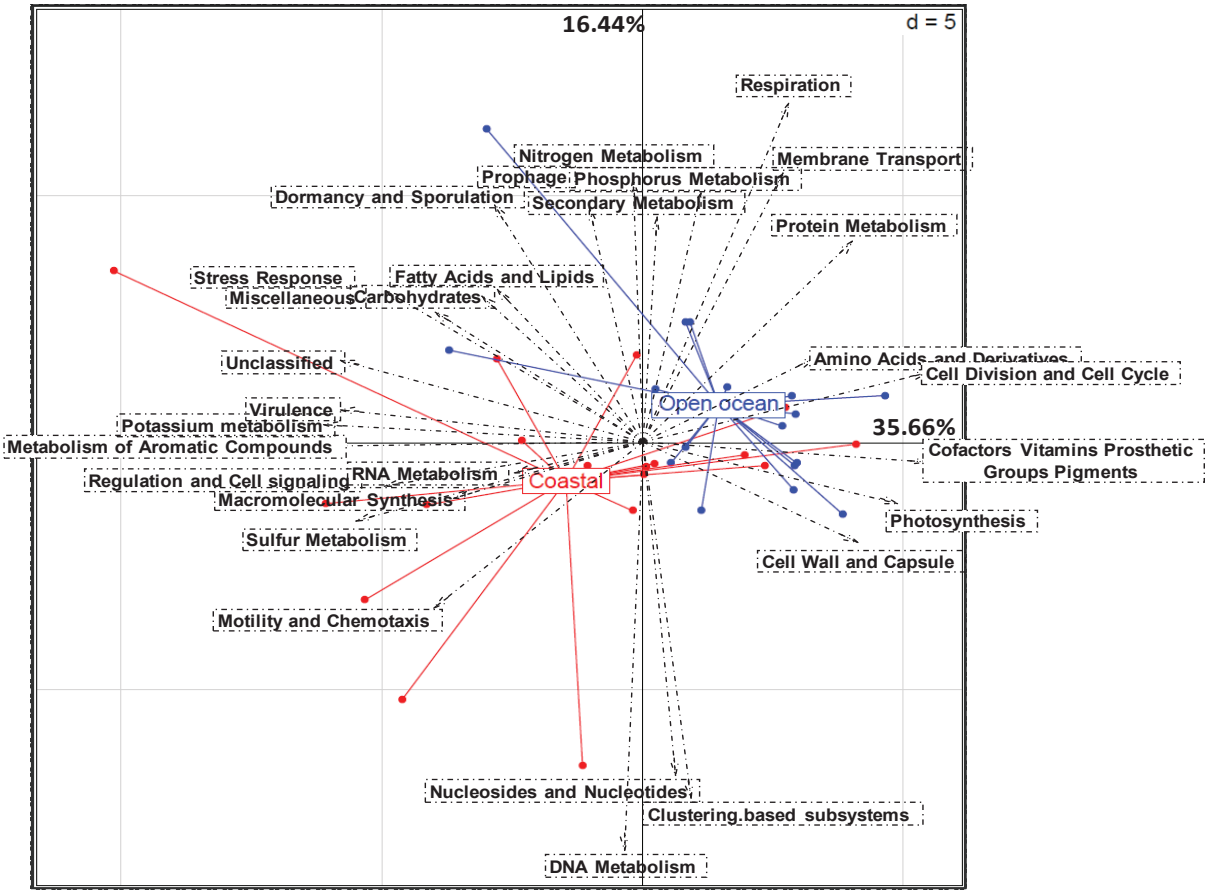


Figure 10. Principal component analysis based on the relative distribution (in percentage of annotated reads) of general functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for 36 metagenomes from the global ocean survey. 18 datasets were selected to represent two sub-environments: coastal area and open oceans. The percentages of the illustrated two major axes correspond to the fraction of the total variance that they represent

This metagenomic comparative approach can be extended to many different ecosystems. In the following example (Figure 11), metagenomes from 13 different ecosystems were compared. In order to produce a legible PCA, sequences from two environments were excluded due to the large quantity of eukaryotic sequences (coral atolls, see figure 7) and viral sequences (Antarctic aquatic environments, see figure 8). Twenty-nine general functions were detected among the 13 compared ecosystems. Phyla are unevenly distributed in the metagenomes (Figure 11). Of course, the low number of general functions used in figure 11 simplifies the system, but is useful for visualizing global microbial community variations (similarly to that shown by Dinsdale and colleagues, [90]).

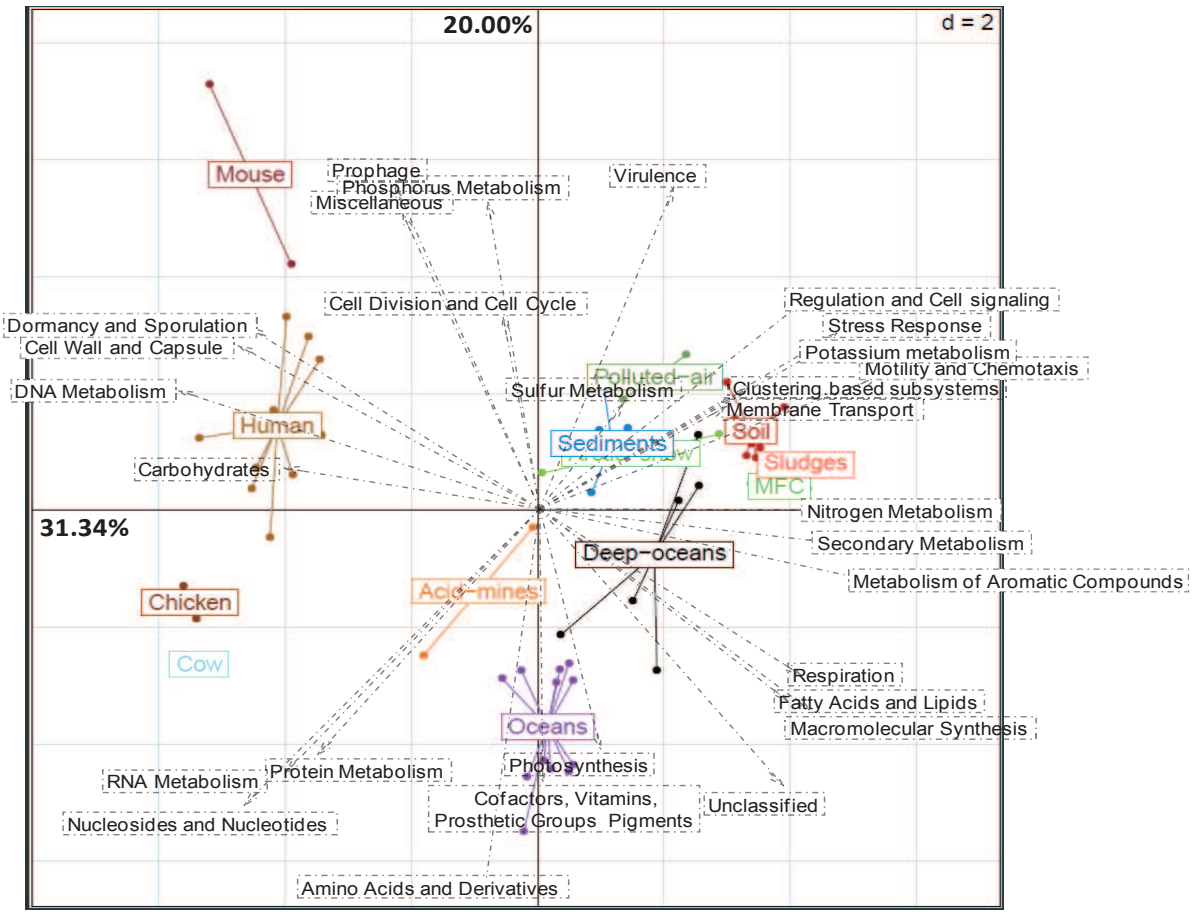


Figure 11. Principal component analysis based on the relative distribution (in percentage of annotated reads) of general functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. The percentages of the illustrated two major axes correspond to the fraction of the total variance that they represent

For example, sequences related to carbohydrate metabolism, DNA metabolism, cell wall and capsule, and dormancy and sporulation are more represented in animals and in particular in human microbial populations. In contRAST, sequences related to metabolism of aromatic compounds, secondary metabolism and nitrogen metabolism are less represented in these environments. Sequences related to RNA metabolism, protein metabolism, and nucleosides and nucleotides are more represented in acid mine drainage biofilms, photosynthesis, cofactors, vitamins, prosthetic groups and pigments, and amino acids and derivatives in oceans. Sequences related to respiration, fatty acids and lipids, and macromolecular synthesis are specific to deep oceans. Those related to stress response, potassium metabolism, regulation and cell signaling, motility and chemotaxis, and membrane transport are more prevalent in soils, sediments, sludges, MFCs, arctic snow and polluted air (which possess also an important proportion of sequences related to virulence and sulfur metabolism).

Principal functions

Although general functions emphasize global community specificities, they are not sufficient to understand in detail how microorganisms have adapted and function in different environments. Therefore, functional distributions were also compared at a more specific level.

Oceans	Mean (%)	SD (%)	Coral atolls	Mean (%)	SD (%)	Deep oceans	Mean (%)	SD (%)
tRNA aminoacylation	2,24	0,34	Photosystem II	21,95	22,32	tRNA aminoacylation	1,64	0,47
Ribosome SSU bacterial	1,65	2,28	Photosystem I	4,76	5,43	Ton and Tol transport systems	1,62	0,54
Serine-glyoxylate cycle	1,38	0,28	F0F1-type ATP synthase	3,60	2,81	DNA-replication	1,26	0,14
Ton and Tol transport systems	1,38	0,45	tRNA aminoacylation	2,33	1,00	Serine-glyoxylate cycle	1,23	0,20
Phosphate metabolism	1,33	0,19	Serine-glyoxylate cycle	1,07	0,49	Phosphate metabolism	1,21	0,26
DNA-replication	1,28	0,09	Phosphate metabolism	0,93	0,40	Iron-sulfur experimental	1,17	0,52
Peptidoglycan Biosynthesis	1,11	0,13	DNA-replication	0,84	0,39	Glutathione-regulated potassium-efflux system*	1,15	0,53
DNA repair, bacterial	1,06	0,10	Universal GTPases	0,83	0,46	DNA repair, bacterial	1,14	0,14
Branched-Chain Amino Acid Biosynthesis	1,05	0,17	Methionine Biosynthesis	0,83	0,40	tRNA modification E.coli	1,10	0,75
Ribosome LSU bacterial	1,01	0,21	Peptidoglycan Biosynthesis	0,78	0,34	Flagellum	1,07	0,43

Antarctic aqu. env.	Mean (%)	SD (%)	Arctic snows	Mean (%)	SD (%)	Soils	Mean (%)	SD (%)
tRNA aminoacylation	2,71	1,40	Ton and Tol transport systems	1,87	0,20	cAMP signaling in bacteria	3,14	0,56
DNA-replication	2,60	0,80	DNA-replication	1,80	0,04	Ton and Tol transport systems	1,73	0,28
DNA repair, bacterial	1,75	1,17	Phosphate metabolism	1,72	0,07	Phosphate metabolism	1,41	0,04
RNA polymerase II	1,64	4,05	tRNA aminoacylation	1,44	0,57	Bacterial Chemotaxis	1,27	0,14
CBSS-335283.3.peg.454	1,60	1,45	Glutathione-regulated potassium-efflux system*	1,37	0,19	Cobalt-zinc-cadmium resistance	1,27	0,26
CBSS-258594.1.peg.3339	1,49	0,90	Cobalt-zinc-cadmium resistance	1,22	0,38	CBSS-258594.1.peg.3339	1,26	0,13
Chitin and N-acetylglucosamine utilization	1,43	3,20	cAMP signaling in bacteria	1,19	0,03	Galactosylceramide and Sulfatide metabolism	1,10	1,71
Peptidoglycan Biosynthesis	1,23	0,48	Flagellum	1,18	0,23	tRNA aminoacylation	1,09	0,05
Ribonucleotide reduction	1,23	1,21	DNA repair, bacterial	1,16	0,08	Serine-glyoxylate cycle	1,04	0,10
Phosphate metabolism	1,12	0,50	CBSS-258594.1.peg.3339	1,09	0,38	DNA-replication	1,04	0,12

Sediments	Mean (%)	SD (%)	Sludges	Mean (%)	SD (%)	MFC	Mean (%)	SD (%)
CBSS-258594.1.peg.3339	1,76	0,23	Glutathione-regulated potassium-efflux system*	1,61	0,13	Bacterial Chemotaxis	1,98	0,14
Cyanobacterial Circadian Clock	1,63	0,40	Phosphate metabolism	1,50	0,00	Copper homeostasis	1,75	0,30
cAMP signaling in bacteria	1,58	0,30	Cobalt-zinc-cadmium resistance	1,38	0,54	Flagellum	1,74	0,12
Phosphate metabolism	1,41	0,19	Serine-glyoxylate cycle	1,23	0,04	Phosphate metabolism	1,47	0,02
tRNA aminoacylation	1,36	0,15	tRNA aminoacylation	1,20	0,00	tRNA aminoacylation	1,40	0,11
Bacterial Chemotaxis	1,29	0,19	cAMP signaling in bacteria	1,18	0,08	DNA-replication	1,37	0,08
DNA-replication	1,09	0,20	DNA-replication	1,09	0,05	Glutathione-regulated potassium-efflux system*	1,35	0,05
DNA repair, bacterial	1,08	0,07	Restriction-Modification System	1,08	0,06	Cobalt-zinc-cadmium resistance	1,29	0,00
Ton and Tol transport systems	0,99	0,11	DNA repair, bacterial	1,07	0,03	Restriction-Modification System	1,19	0,40
Maltose and Maltodextrin Utilization	0,87	0,16	Flagellum	1,07	0,10	Iron-sulfur experimental	1,18	0,07

Acid mine	Mean (%)	SD (%)	polluted Air	Mean (%)	SD (%)	Human feces	Mean (%)	SD (%)
Lactose and Galactose Uptake and Utilization	3,47	1,71	Ribosome SSU bacterial	7,90	3,46	tRNA aminoacylation	2,19	0,29
Tn552	3,44	2,15	Toxin-antitoxin systems**	5,52	0,39	Multidrug Resistance Efflux Pumps	1,75	0,17
tRNA aminoacylation	2,23	0,07	Galactosylceramide and Sulfatide metabolism	3,85	0,87	Cellulosome	1,52	0,94
Cobalt-zinc-cadmium resistance	1,37	0,35	Ton and Tol transport systems	3,83	0,11	DNA-replication	1,45	0,08
Glutathione-regulated potassium-efflux system	1,28	0,43	Cobalt-zinc-cadmium resistance	1,23	0,02	Peptidoglycan Biosynthesis	1,44	0,18
Trehalose Biosynthesis	1,17	0,14	Phosphate metabolism	1,23	0,10	Phosphate metabolism	1,43	0,11
DNA repair, bacterial	1,13	0,14	Bacterial Chemotaxis	1,21	0,02	DNA repair, bacterial	1,35	0,07
Flagellum	1,11	0,45	tRNA aminoacylation	1,14	0,03	Maltose and Maltodextrin Utilization	1,27	0,12
CBSS-258594.1.peg.3339	1,07	0,19	DNA-replication	1,09	0,04	Sucrose Metabolism	1,18	0,10
TCA Cycle	1,06	0,26	DNA repair, bacterial	1,07	0,02	Methionine Biosynthesis	1,18	0,18

Chicken cecum	Mean (%)	SD (%)	Mouse cecum	Mean (%)	SD (%)	Cow rumen	Mean (%)	SD (%)
tRNA aminoacylation	3,76	0,15	Alkylphosphonate utilization	3,19	0,46	tRNA aminoacylation	4,63	0,13
Universal GTPases	1,83	0,26	Multidrug Resistance Efflux Pumps	1,81	0,05	Universal GTPases	2,10	0,11
Peptidoglycan Biosynthesis	1,51	0,03	tRNA aminoacylation	1,77	0,19	DNA-replication	2,01	0,04
Galactosylceramide and Sulfatide metabolism	1,50	0,22	Sucrose Metabolism	1,75	0,94	Ribosome LSU bacterial	1,98	0,03
DNA-replication	1,43	0,06	Restriction-Modification System	1,64	0,19	RNA polymerase bacterial	1,84	0,02
RNA polymerase bacterial	1,41	0,09	Phosphate metabolism	1,60	0,30	Methionine Biosynthesis	1,69	0,05
Ribosome LSU bacterial	1,37	0,01	DNA-replication	1,45	0,18	Peptidoglycan Biosynthesis	1,50	0,09
DNA repair, bacterial	1,36	0,01	Bacterial Cytoskeleton	1,40	0,00	Iron-sulfur experimental	1,46	0,20
Methionine Biosynthesis	1,28	0,09	DNA repair, bacterial	1,36	0,12	DNA repair, UvrABC system	1,44	0,03
Multidrug Resistance Efflux Pumps	1,21	0,08	CBSS-258594.1.peg.3339	1,29	0,01	Maltose and Maltodextrin Utilization	1,36	0,05

Table 2: List of the 10 most represented functional subsystems (relative proportion in percentage) for 15 distinct ecosystems. Means and standard deviations calculated from all the metagenomic datasets from the same environment.

A total of 838 functional subsystems were detected at least once among the 77 studied metagenomes. As shown for phylogenetic distributions, some of them appear to be significantly and unequally represented in the different environments and highlight microbial community specificities.

The distribution of the ten more represented functional subsystems was compiled for the 15 ecosystems (Table 2). The standard deviations correspond to variance among metagenomes from the same environment and provide information about intra-ecosystem variations.

In contRAST to phylogeny, major functions vary less between the different ecosystems. In particular, tRNA aminoacylation genes are predominant in all the metagenomes. Phosphate metabolism, DNA replication, peptidoglycan biosynthesis and bacterial DNA repair are also highly represented in a majority of environments. However, ecosystems also have specific functions that are more highly represented in comparison to the other ecosystems (e.g. cellulosomes in human feces or cyanobacterial Circadian clock in hypersaline sediments).

The most represented functions are important to highlight biogeochemical specificities of the different environments. Nevertheless, an important part of the functions appears to be lightly represented in all the metagenomes but with significant proportional differences. Some groups of functional genes were selected on the basis of the information they provided about their environments. Some specific functions can highlight environmental stresses and pollution, or the possible presence of compounds of interest.

Metamobilome:

The metamobilome can be defined as the mobile part of a metagenome. It includes integrons, transposons, prophages, plasmids, genomic Islands, and inteins [2]. These genetic structures are called mobile genetic elements (MGEs) and are largely involved in bacterial and archaeal adaptation and evolution. However, the intra- and inter-genomic mobility strategy varies between these MGEs, so providing microorganisms with a range of tools to increase cellular fitness or for adapting to different stresses. Moreover, the quantity of these elements in an ecosystem might reflect the adaptation capacities of the specific microbial communities.

The comparison of the relative proportion of these MGEs in different environments could uncover their environmental specificities and possible provide clues to their roles in these microbial communities. The relative proportion of six genes involved in microbial adaptation was compared in the 77 metagenomes (Figure 12). These genes are unevenly represented between and for some of them among the different environments. Plasmids appear to be more active in some deep oceans, in one coral atoll and in sediments. Antarctic aquatic ecosystems appear to possess considerable adaptation capacities (relatively high proportion of inteins, integrons and prophages). On the other hand, the relative proportion of the six

MGEs appears to be limited in oceans. More studies need to be performed to understand the role of each of these elements and why they are so unevenly represented in the different microbial communities.

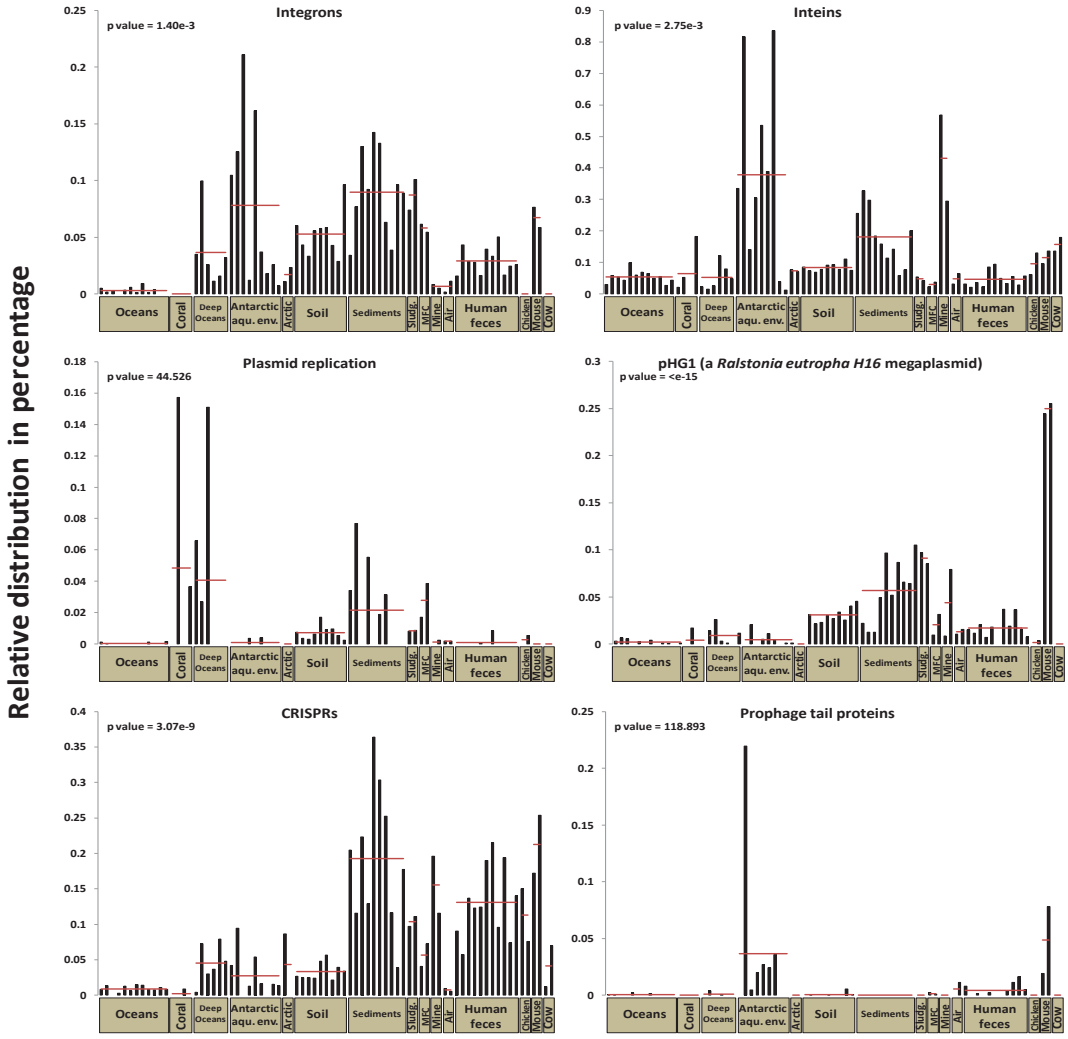


Figure 12. Relative distribution (in percentage of annotated reads) of functional subsystems related to the metamobilome (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Genes involved in aromatic compound degradation:

Aromatic hydrocarbons are a class of hydrophobic compounds that are often persistent in the environment due to a relatively low solubility in water and are, in some cases, cytotoxic, genotoxic, mutagenic and carcinogenic [109-111]. Human activities are largely involved in the release of these molecules in nature (e.g. industry, herbicides, insecticides, vehicular emissions) although natural seepage from petroleum reservoirs can also occur. Microbial

culture approaches have demonstrated that some microorganisms have the genetic capacities to use these compounds as a source of carbon and energy via a variety of aromatic degradation pathways [112]. In spite of the discovery of various pathways, the number of microorganisms studied is still largely limited by the cellular culture steps and it is possible that non cultivable species are largely involved in the degradation of these compounds in nature.

The proportion of the detected pathways is commonly lowly represented (less than 0.1%) in the different metagenomes (Figure 13). These capacities are limited in animals and varied considerably in the others environments. Genes involved in toluene, naphthalene, anthracene and p-cymene degradation capacities are more represented in oceans, deep oceans, soils and polluted air.

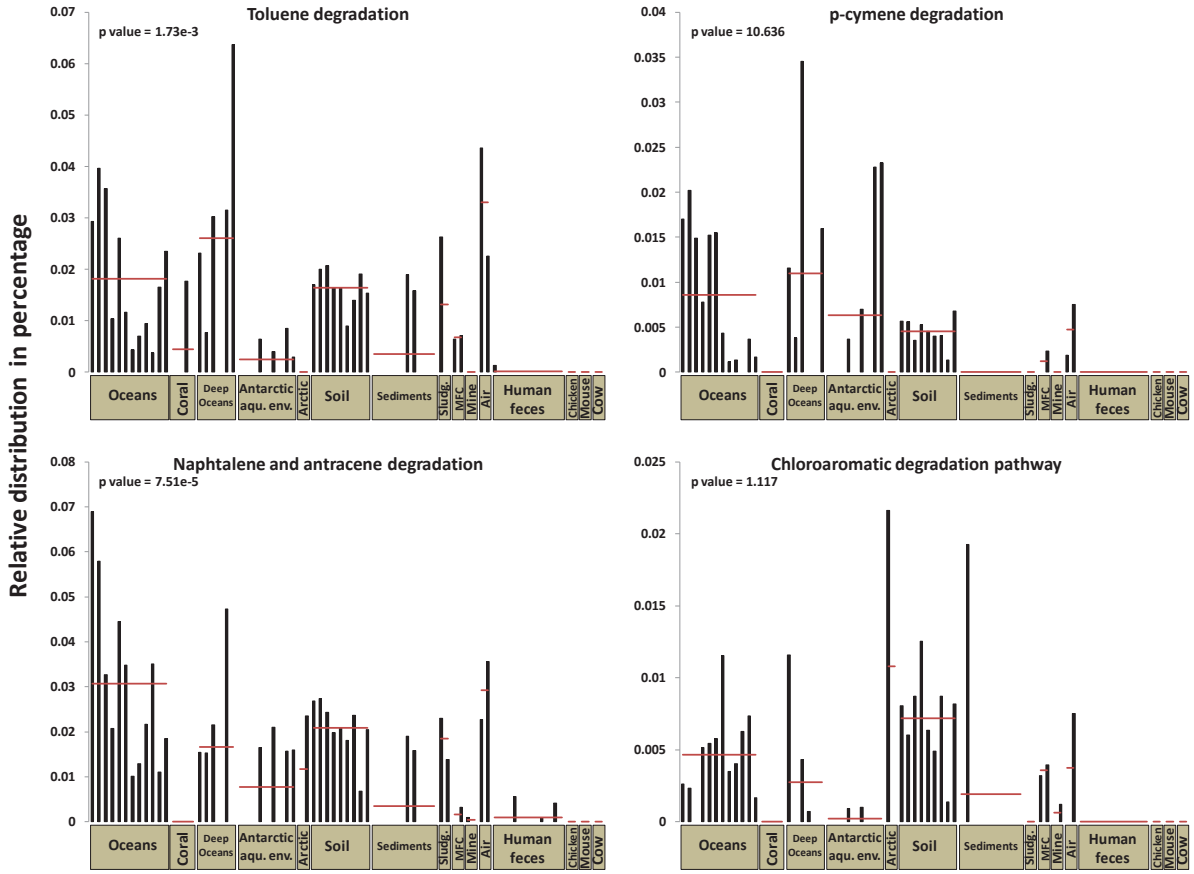


Figure 13. Relative distribution (in percentage of annotated reads) of functional subsystems related to aromatic compounds degradation (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Interestingly, in spite of the recent emergence of chloroaromatic compounds from human industry (e.g. PCB,) [113], chloroaromatic degradation pathway is already present in non negligible proportion in various natural environments. In particular and based on the

distribution of this subsystem, oceans appear to be in majority contaminated by these toxic molecules. Arctic snow and soils possess also these degradation capacities. It is interesting because a part of the soil metagenomes were extracted from a natural grass land preserved from pollution since more than 150 years [83]. So these compounds are probably propagated by air. However Antarctic aquatic environments and animals are probably not yet considerably contaminated by chloroaromatic compounds.

Species known to be involved in aromatic or chloroaromatic compounds [112] and which are detectable in the metagenomes studied are not correlated to the genes involved in these degradation pathways (supplement data). Of course it is necessary to provide metatranscriptomic data to confirm these results, but it is possible that unknown microorganisms are strongly implicated in aromatic compound degradations in environments and that new species and genes of interest will be discover in future metagenomic studies.

Resistance to metals:

A majority of metals play an essential role in the life processes of microorganisms [114]. It is the case of zinc, nickel, manganese, sodium, chromium, cobalt, calcium, copper, iron, potassium, and magnesium which are essential nutrients. On the other hand, gold, aluminum, silver, cadmium, lead and mercury are not essential for cell processing and have a greater affinity to oxygen sites and thiol-containing groups, implying a higher cellular toxicity.

However, all the metals are toxic for cells at high concentration. To respond to this toxicity, bacteria and archaea possess various resistance systems [115,116] which are known to be mainly supported by plasmids and transposons. For example, the arsenate reductase (ArsC) is used by different species to reduce arsenic [117]; czcA, czcB and czcC are genes coding the CzcABC protein complex which mediates resistance to cobalt, zinc and cadmium [118,119]; the mer operon provides the capacity to detoxify organic mercury [120]. The proportion of these resistances, which can reflect a metal pollution in a given environment, varied considerably between the 77 metagenomes (Figure 14).

Arsenic resistance is more present in polluted air, acid mine drainage biofilms, microbial fuel cells and sludges, but is also highly present in chicken cecum, some deep-oceans and in sediments. Zinc resistance alone is clearly more present in sediments, instead the proportion of sequences related to cobalt, zinc and cadmium resistance is higher in sludges, acid mine drainage biofilms, polluted air, soil and some deep-sea oceans, so underlining cobalt or cadmium presence in these environments. Mercury resistance operon is fortunately undetected in animals, but is present in diverse environments (e.g. sludges and deep-oceans) and in particular in polluted air.

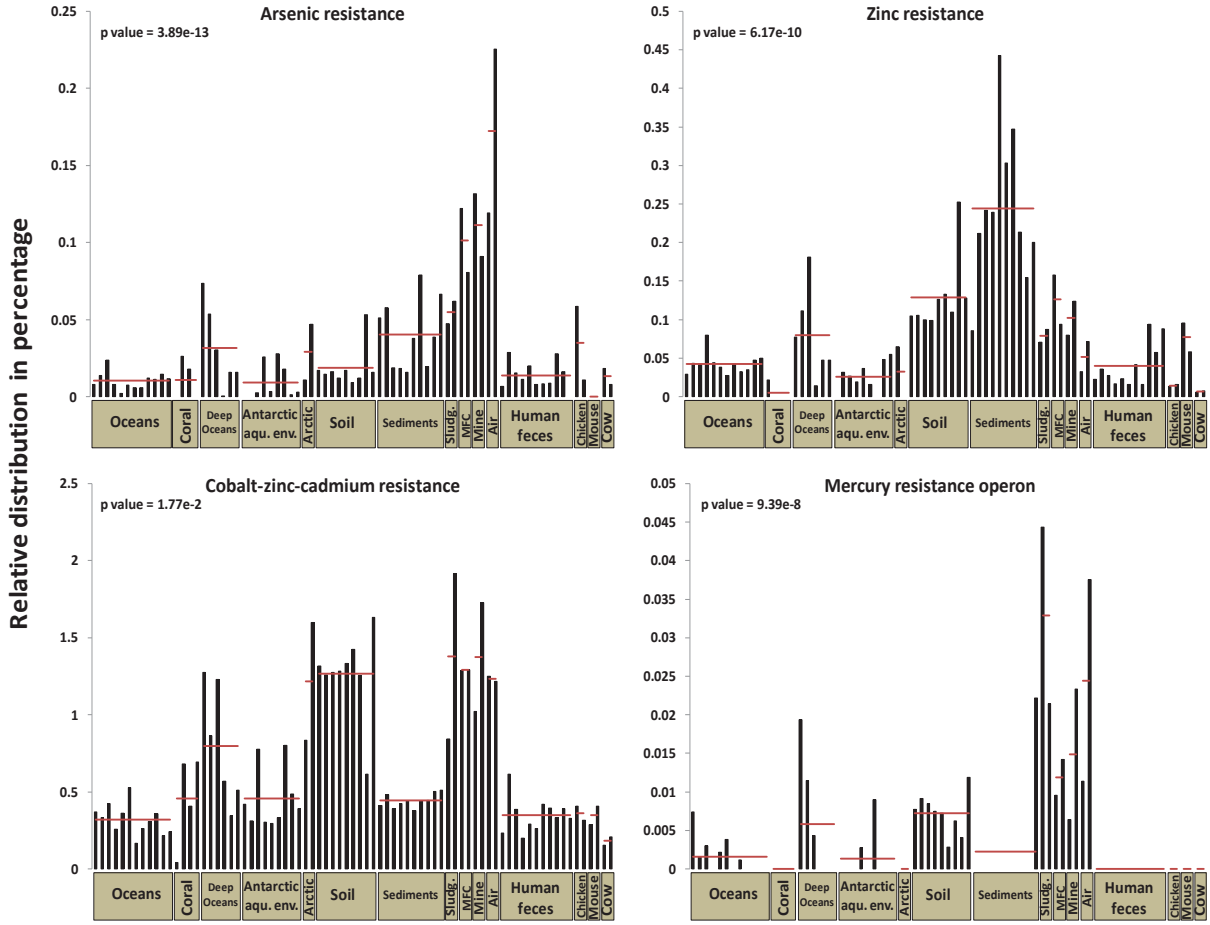


Figure 14. Relative distribution (in percentage of annotated reads) of functional subsystems related to metal resistances (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Resistance to antibiotics:

The widespread utilization of natural or synthetic antibiotics to limit human microbial infections has quickly induced the emergence and propagation of resistance mechanisms within clinical microbial communities [121]. These mechanisms can appear by genomic mutations or by horizontal transfers with plasmids, transposons or integrons for vectors [121,122]. For example, betalactamases are enzymes that hydrolyse the betalactame cycle of betalactamines, so inactivating its antibiotic properties [123]. The blaR1 gene is an antirepressor regulating betalactamase expression [124]. This gene involved in antibiotic resistances is unusually highly represented in human feces and mouse cecums (especially in

the obese mouse with more than 0.2% of the annotated functions) (Figure 12), underlining human impact in these metagenomes. In contRAST, this gene family is almost undetected in oceans. Moreover, the blaR1 gene is known to be present in staphylococci, but no correlation between the function and these bacteria was found based on their relative distribution among the 77 metagenomes (supplement data).

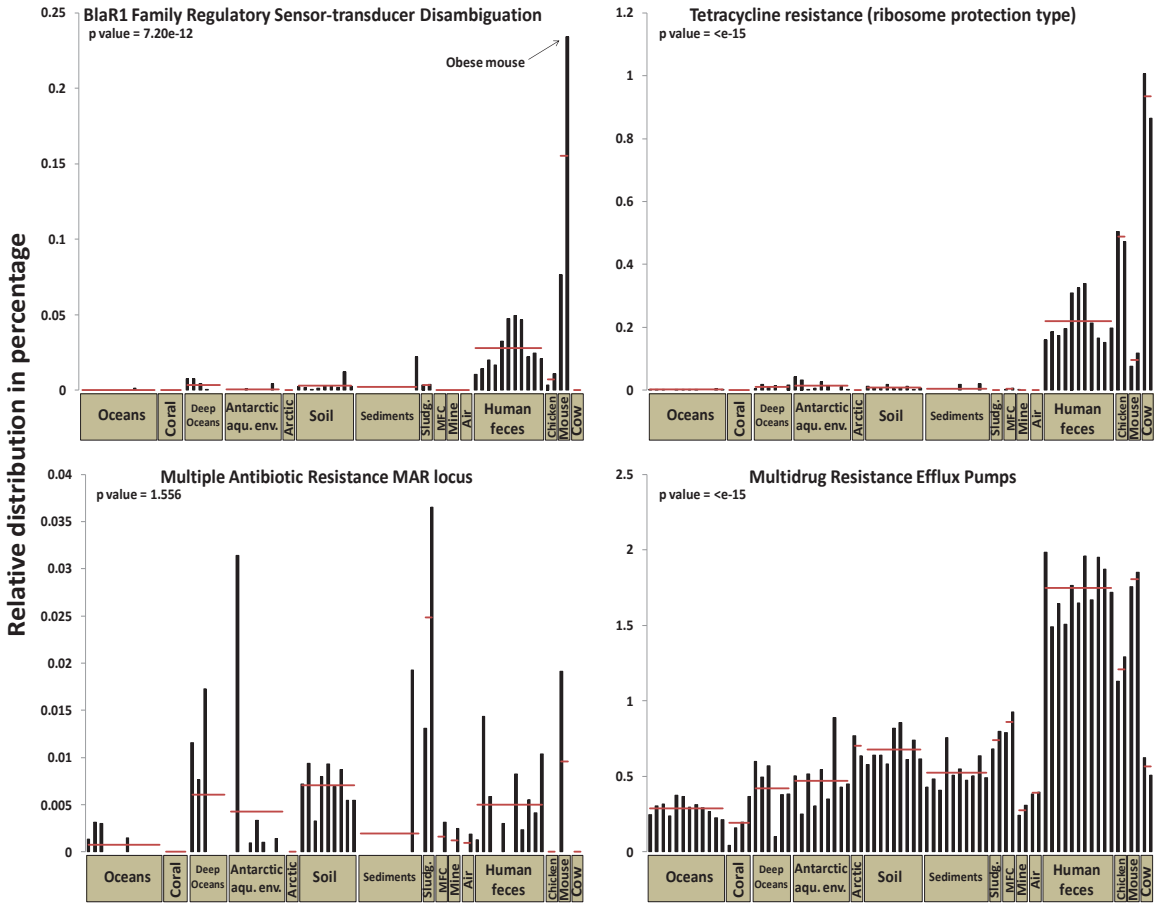


Figure 15. Relative distribution (in percentage of annotated reads) of functional subsystems related to antibiotic resistances (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Another example of mechanism is the resistance to tetracycline. Produced molecule prevents the action of the antibiotic between aminoacyl tRNA and the ribosome [125] and is an important mechanism of resistance for bacteria [126]. This mechanism is detected in relatively high proportion in animals and especially in cow rumens in comparison to all the other environments (Figure 15). This observation provides important information about possible xenobiotic impacts on human and animal microbial communities. Of course, additional experiments have to be done to define with precision the importance of this dissemination and the exact role of human health strategy. In addition, genetic environment

of these genes have to be define to identify their different hosts (pathogenic or not) but also if mobile genetic elements are involves in this dissemination.

In some species, these mechanisms can be grouped into multi-drug resistance systems. In particular, the multiple antibiotic resistance (mar) loci, known to be present in E. coli and Salmonella, can be involved in different antibiotic families (e.g. penicillins, fluoroquinolones, tetracycline, chloramphenicol) [127-129]. Interestingly, this locus is detected in various environments (Figure 15), like deep oceans, Antarctic aquatic environments, and sludges. Furthermore, no correlations were founds between this locus and E. coli or Salmonella (supplement data). Results confirm that this locus can be involved in resistances other than for antibiotics in the nature. This structure is yet known to be involved in resistance to oxidative stress agents [130,131], to organic solvants [132,133] or to disinfectants [134].

Finally, the relative proportion of multidrug resistance (MDR) [135] efflux pumps was compared into the 77 metagenomes. MDR efflux pumps contribution to antibiotic resistance in pathogens was largely studied [136] [137] [138]. However, these structures are present in all living organisms and are not restricted to antibiotic compounds [139]. MDR efflux pumps are capable of extruding heavy metals [140] [141], solvents [142], and antiseptics [143]. They are known to be largely present in soil and in association with plants [144]. However, based on this global metagenomic comparison, they appear to be more present in human feces and chicken and mouse cecum (Figure 12). This element is also present in all the environments studied so confirming its multifunctional role, but is limited in oceans, deep oceans, polluted air and acid mine drainages biofilms.

Photosynthesis:

Oxygenic photosynthesis process is used by plants, algae and cyanobacteria to convert solar energy into chemical energy [145]. This process is essential for oxygen stabilization on earth and so indispensable for an important part of life [146]. Moreover, oxygenic photosynthesis converts around 1014 grams of carbon from CO2 into biomass every year [147] and so could limit future human induced climate changes by limiting carbon dioxide augmentation.

Photosystems I and II are two large membrane protein complexes which catalyze the primary step in this energy conversion [145]. Due to high presence of algae related sequences in atoll coral metagenomes, these two subsystems are prevalent in this environment (Figure 16). As expected, they are practically undetectable in animals and present in various proportions in nature. When excluding atoll coral metagenomes, they are in high proportion in an oceanic metagenome corresponding to a marine bacterioplankton environment.

On the other hand, photorespiration (oxidative photosynthesis) involves oxygen fixation and carbon dioxide release [148]. This process decreases the efficiency of photosynthesis and appears to be unevenly represented in the 77 metagenomes (Figure 16). In oceans, oxidative photosynthesis is approximately 2.5 times less represented than photosystems I and II. But in other environments, like soil, sediments and animals, the ratio is inversed, so underlining different balances between oxidative and oxygenic photosynthesis as a function of the ecosystems.

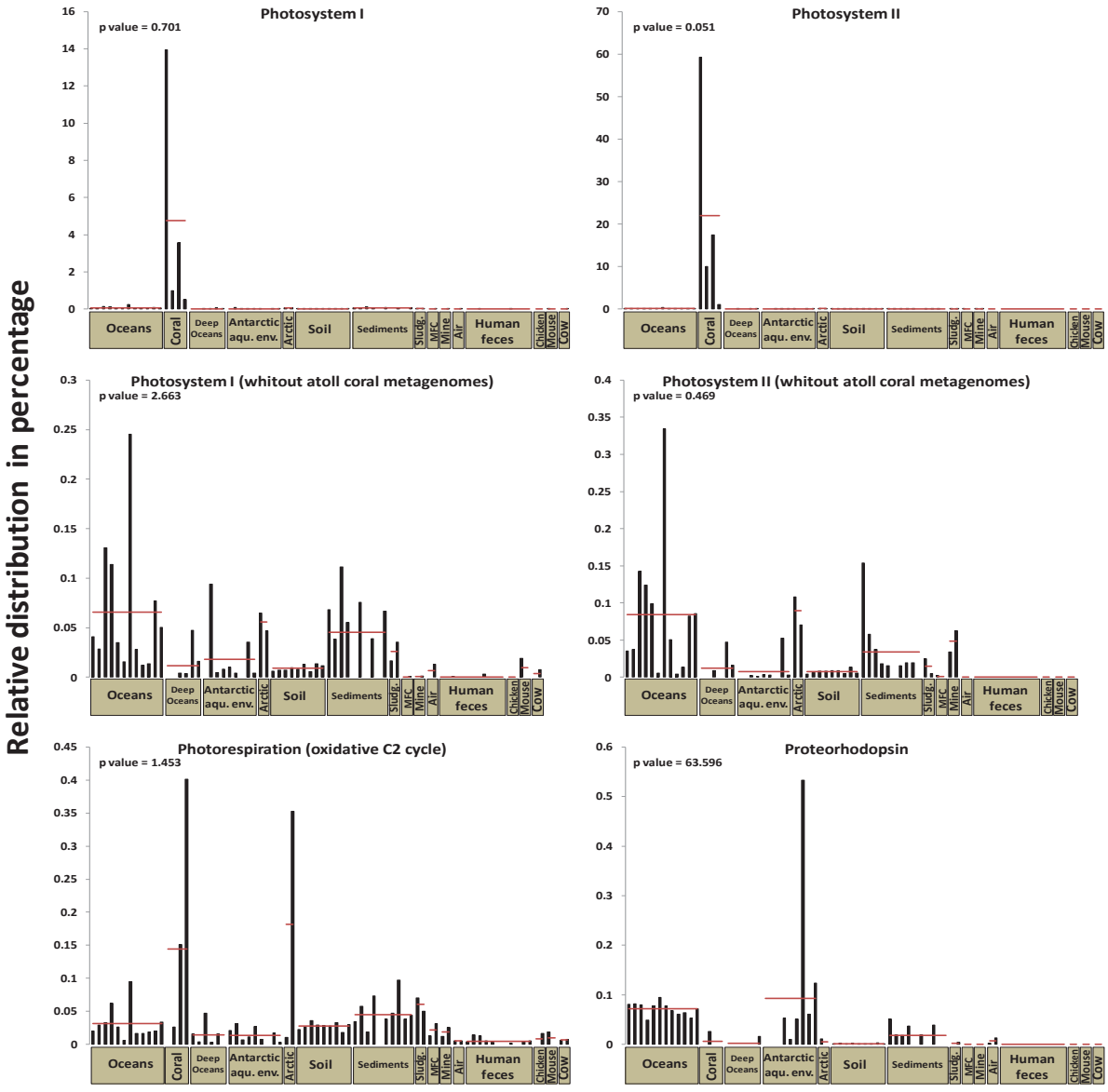


Figure 16. Relative distribution (in percentage of annotated reads) of functional subsystems related to photosynthesis (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Interestingly, ten years ago a new mode of mediated light-driven energy generation, different than the largely known chlorophyll-based photosynthesis, was discovered and is now highly studied [149]. This process is generated by retinylidene membrane proteins named proteorhodopsins and is known to be used by different marine proteobacteria in the trophic zone of the ocean [150]. However, new investigations indicate that proteorhodopsins could have numerous physiological functions, different than light-driven energy generation [149].

As expected, proteorhodopsin is undetectable in animals and is present in (surface) oceans, more than in deep oceans were light is low. But interestingly, its relative proportion is considerable in one Antarctic aquatic metagenome and could have a specific or more important role in this environment. Moreover, this process known to be present only in aquatic environments is not detectable in soil but appears to be present at different sediments depths.

Nitrogen cycle:

Nitrogen is an essential element for all organisms in particular because of its crucial role in protein structure. To fixe nitrogen from air, some microorganisms express specific genes coding nitrogenase enzyme complexes which convert atmospheric dinitrogen to NH_4^+ [151,152]. These cells, called diazotrophs, are present in all ecosystems and are an indispensable source of nitrogen for life [153,154]. Some environments appear to fixe important quantities of nitrogen (Figure 17). It is the case of sediments, sludges, microbial fuel cells and acid mine drainage biofilms. Furthermore, one Antarctic aquatic metagenome possesses an important nitrogen fixation potential. However, oceans, soil and animals appear to be lowly involved in this process.

As a result of diazotrophs, the nitrogen cycle is largely driven by some microorganisms which utilize nitrogen compounds and redox process to generate energy [155]. It is the case of denitrification and dissimilatory nitrite reductase which reduce respectively nitrate [156] and nitrite [157] to N_2. These processes imply the loss of accessible nitrogen for cells and so induce important problems for agriculture activities. On the other hand, they are known to remove excessive quantity of nitrogen from sewage and wastewaters. The importance of this process in sludges (Figure 17) in comparison to other environments corroborates this aspect. Moreover, deep oceans appear to be strongly involved in denitrification and could have because of the importance of this environment in term of volume an unexpected role in nitrogen cycle.

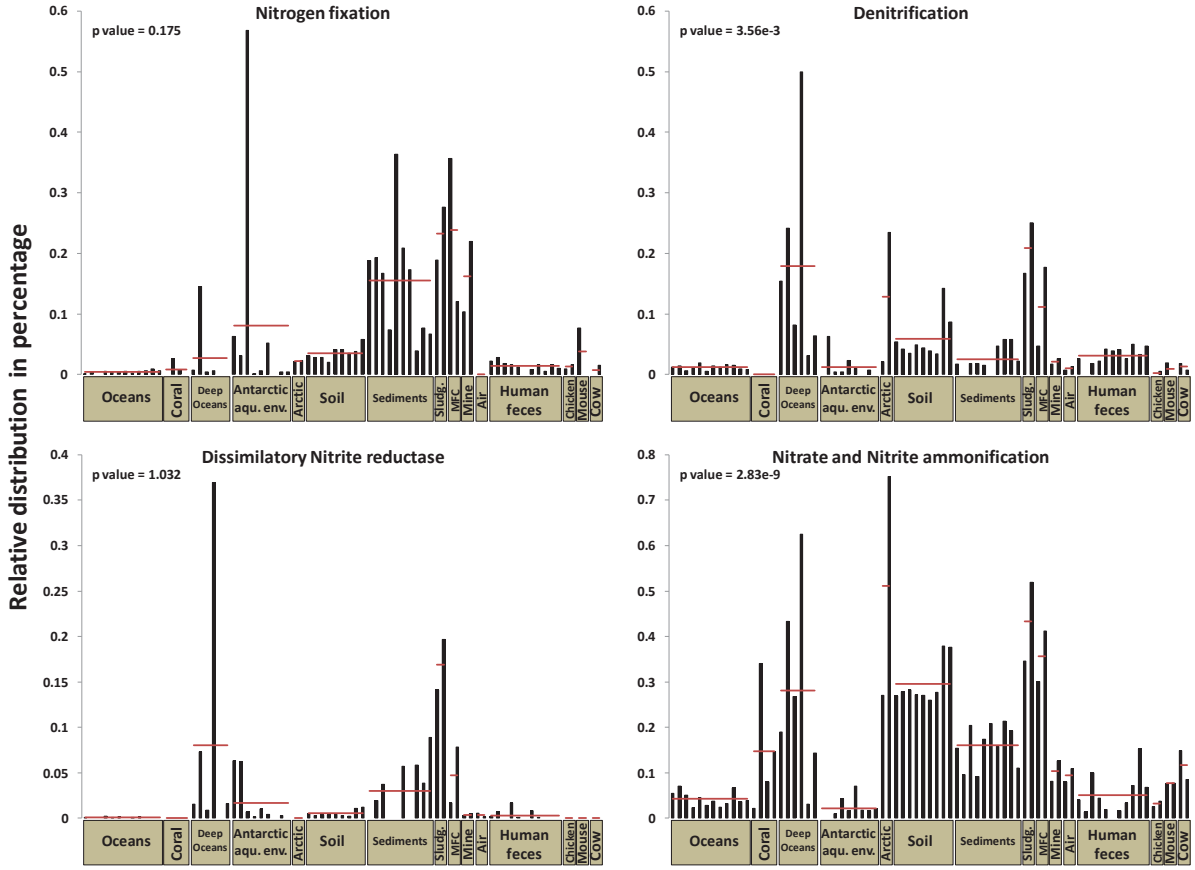


Figure 17. Relative distribution (in percentage of annotated reads) of functional subsystems related to the nitrogen cycle (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

But nitrate and nitrite can also be reduced in ammonium. This process is called ammonification and is used by bacterial and archaeal cells to transform organic nitrogen into ammonia [158]. Like for denitrification, this process appears to be mainly present in deep oceans and sludges, but is also commonly detected in soil, sediments, microbial fuel cells and specially in Arctic snows

Ecosystems specificities:

While it is largely known that microorganisms are adapted to their environments, approaches providing information about their specificities at the community level are relatively limited. We assume that i) the study of one environment using metagenomic approaches is necessary to generate a dataset corresponding to the related microorganisms, that ii) this study alone is insufficient to understand the peculiarities of this dataset, and finally that iii) global metagenomic comparisons are unique tools to visualize microbial communities' specificities at the ecosystem level by confronting a compilation of datasets.

Although a function can be relatively stable in the majority of metagenomes sequenced from a studied environment (e.g., global ocean survey, see figures 9 and 10), it is difficult to know if its distribution is normal or unique without comparing its distribution in other environments. Related species and especially functions possessing unusual distributions provide unique information about microbial communities' adaptations, interactions and impact for climate and biogeochemical cycles present in the different studied environments [92]. We present some functions and taxa unusually represented in the 15 compared environments and a part of what these observations provide in term of information and hypotheses elaborations. "Give me your metagenome, I will tell you what you are" could be an appropriate expression to explain the interest of global metagenomic comparisons to understand environmental peculiarities and its consequences in term of microbial life styles.

Oceans:

Microbes mediate fluxes of matter and energy in the ocean (see [159] as a review) and represent a crucial bio engine for the maintenance of climate at the planet level. As an example, marine phytoplankton carries out almost half of the Earth photosynthesis [147]. As a consequence, this environment is highly studied by microbiologists to understand the different roles of bacterial and archaeal communities living there. Because oceans were probably until now the most studied environments using metagenomic approaches (e.g. global ocean survey, [82]), an important number of oceanic metagenomes are accessible. 12 metagenomes corresponding to various oceanic localizations generated during different projects were selected to characterize this environment.

Based on this global metagenomic comparison, oceans appear to possess the most important distribution of sequences related to dimethylsulfoniopropionate (DMSP) breakdown (Figure 18). DMSP occurs in considerable amounts in marine algae, for which this molecule and its breakdown products probably serve as an antioxidant system [160]. But probably more important, this reaction can release dimethyl sulfide molecules (DMS) in the atmosphere, which can potentially improve cloud formation and so limit solar radiations at the planet level [161]. In addition, DMS can have signaling roles for bacteria [162].

The important distribution of sequences related to this reaction in oceans can play a considerable role in global climate stabilization, but can also be a widespread bacterial signaling mechanism in this environment.

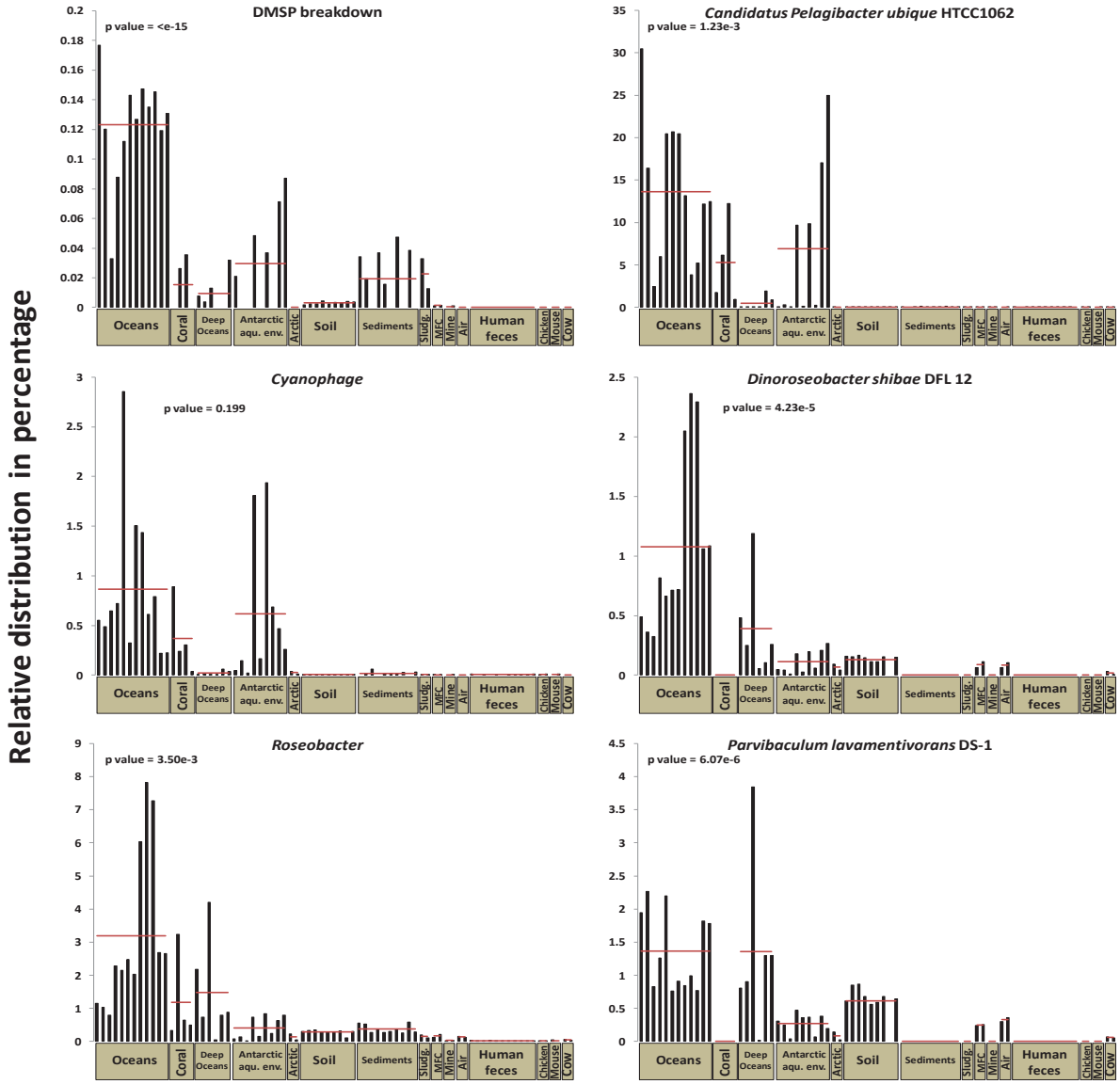


Figure 18. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Sequences related to Candidatus Pelagibacter ubique are highly detected in oceans but also in coral atolls and antarctic aquatic environments. This species is the first cultured member of the SAR11 clade, known to represent 25% of microbial cells in oceans [163]. In addition, its genome contains few mobile genetic elements (no introns, inteins, transposons and plasmids). This genomic characteristic explains for a part the low proportion of genes involved in mobile genetic elements in oceans (see figure 12). Sequences related to Dinoroseobacter shibae are also more represented in this environment. These cells were isolated from marine dinoflagellates (Flagellate protists) and are aerobic anoxygenic phototrophic cells [164]. These two species are known to use DMSP as a substrate

[165,166], which is for a part in concordance with the distribution of sequences related to DMSP breakdown among the 77 metagenomes.

Cyanophage related sequences are specific to oceans, coral atolls and antarctic aquatic environments. These bacteriophages are known to be specific to marine Cyanobacteria cells (Synechococcus and Prochlorococcus; see [167] as a review). The distribution of these phages and Cyanobacteria are not correlated, in particular in the hypersaline sediments were sequences related to Cyanobacteria are prevalent and were cyanophages are almost undetected (see figure 3 and 18). So these observations confirm that cyanophages are specific to only parts of Cyanobacteria (e.g. Prochlorococcus, but not Anabaena and Nostoc which are highly represented in the hypersaline sediments; see figure 26). Finally, Roseobacter, known to be largely represented in oceans (see [168] for more details), and Parvibaculum lavamentivorans, known to alkylbenzenesulfonate [169] are also globally more represented in this environment.

Coral atolls:

Metagenomes related to coral atolls were generated to elucidate the roles of microorganisms and viruses on coral reef ecosystems and to study they response to coral reef degradations [47]. While these metagenomes are highly contaminated by eukaryotic sequences (examples are provided in the figure 8), this sequencing effort provides some information about the life style of coral atoll microbial populations.

In particular, sequences related to photosystems I and II (protein complexes involved in photosynthesis and found in plants, algae and Cyanobacteria, see figure 16), cytochrome B6-F complex (electron transfer between photosystems I and II), and membrane-bound ATP synthases (F0F1-type ATP synthase) are more detected in the four coral atolls related metagenomes [47] in comparison to all the other sequenced environments. This unusual distribution of sequences involved in photosynthesis mechanism is principally due to the considerable presence of sequences related to algae in the two coral atoll metagenomes (see figure 6).

But some taxa are also highly represented in this environment. In particular, sequences related to Prochlorococcus, Synechococcus and Thermosynechococcus are highly detected in this environment (Figure 19) and these cyanobacteria can also be involved in photosynthesis mechanisms. In addition, Alteromonas macleodii-like bacteria, highly detected in some coral atoll related metagenomes, are known to play a particular role in the transfer of organic carbon from coral mucus to the pelagic microbial food webs of coral reefs [170].

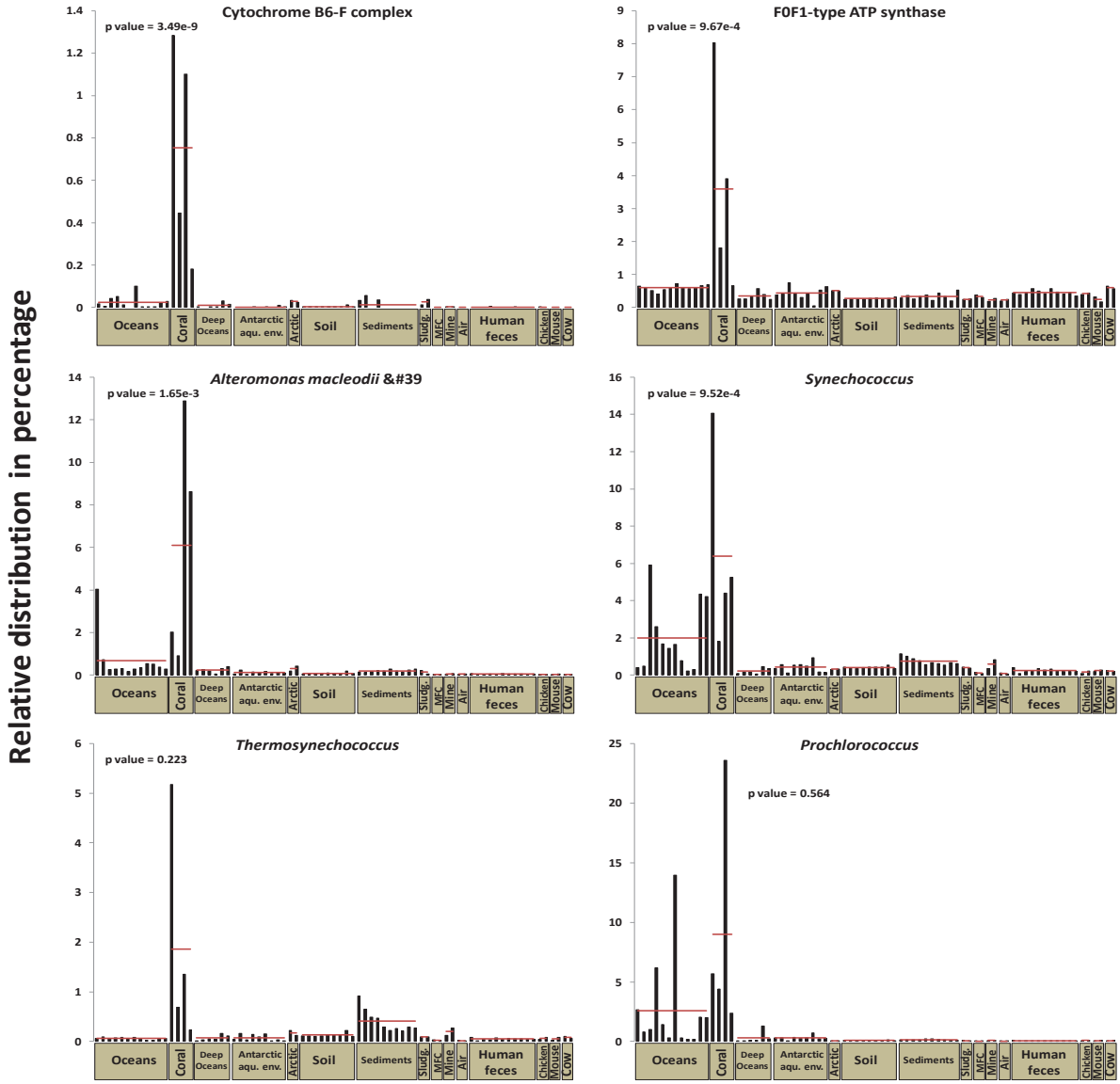


Figure 19. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Deep oceans:

The average depth of the ocean is 4,267 meters and Deep Ocean represents a considerable part of the water of our planet. While this environment is largely unknown in term of biodiversity, some metagenomic studies were done using samples extracted from the deep ocean [38,171,172] and the generated sequences provide information about the related microbial communities. However, these different metagenomes are not similar in term of functional and taxonomical distributions (e.g., distribution of the Epsilonproteobacteria

class, figure 4) so reflecting important variations in the different locations studied and a probably rich environment in term of microbial communities' life styles. In other words, the community diversity differences among the "deep ocean" samples prevent an extensive characterization of this ecosystem. Future analyses might uncover significant structural and functional differences between the sampled sites.

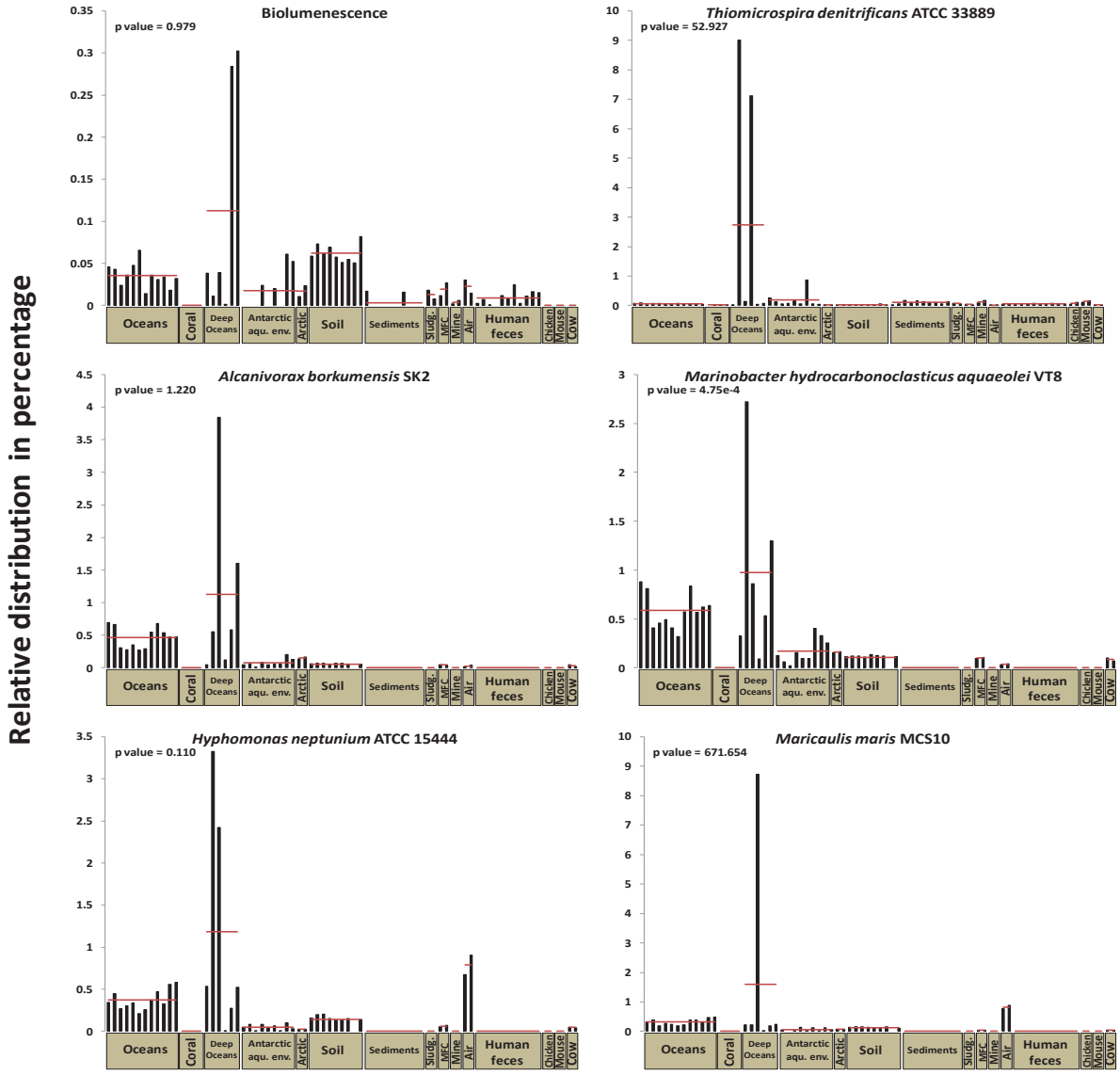


Figure 20. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

In deep oceans related metagenomes (more than 400 meters depth), four functions are noticeably distinct. Sequences related to bioluminescence (Figure 20), denitrification, dissimilatory nitrite reductase (see figure 17), and protection from reactive oxygen species (supplement data) are in average more distributed in these metagenomes in comparison to

the other. These functions emphasize specific microbial adaptations in this environment and an unexpected role in nitrogen cycle but are largely unevenly represented among this environment.

Alcanivorax borkumensis is unusually represented in some of these metagenomes (Figure 20). This species is a marine bacterium which uses petroleum oil hydrocarbons as the only sources of carbon and energy [173,174]. This bacterium is known to be lowly represented in oceans but in important proportion in oil-contaminated aquatic environments, and especially when nitrogen and phosphorus nutrients are highly present [175]. The important proportion of sequences related to genes involved in nitrogen cycle suggests that this element is highly present in these deep oceans. The presence of Alcanivorax borkumensis suggests that oil is also present.

To corroborate this hypothesis, sequences related to a Marinobacter hydrocarbonoclasticus species (Figure 6), known to use various hydrocarbons as the sole source of carbon and energy [103,176] and Parvibaculum lavamentivorans DS-1, known to degrade the linear surfactant alkylbenzenesulfonate [177] are also highly represented in some of these metagenomes. In addition, sequences related to toluene and p-cymene degradation are in average more represented in deep oceans than in all the other environments (see figure 13). In addition, a majority of these metagenomes appear to possess relatively more aliphatic oil degrading genes than the surface of oceans and soils [92].

Other species or groups of species are also unusually more represented in some of these metagenomes. In particular, due to its importance for human disease, it is important to note that Campylobacter species and C. jejuni in particular are highly represented in some cases, and especially in an hydrothermal vent [172] were this species is more represented than in human feces. C. jejuni is perhaps the most common cause of human gastroenteritis, but is possibly also present in unexpected environments. This observation feeds the hypothesis of deep oceans pathogenic microorganism's emergences and could explain why the C. jejuni genome possesses an unusual high proportion of hypervariable regions as a strategy for short adaptations and survival in unstable environments [178]. However, Sulfurovum sp. NBC37-1, a species closed to C. jejuni, appears to be predominant in this metagenome (more than 50% of detected species, [172]). So it is possible that sequences related to C. jejuni belong in reality to another genome due to the limits of the SEED annotation.

A dimorphic prosthecate bacterium that uses its stalk as a reproductive structure, Hyphomonas neptunium [179] is also highly represented in two deep ocean metagenomes (Figure 16). It possesses the particularity to produce a motile cell from a non motile cell as a strategy to colonize new habitats. In addition, Maricaulis maris MCS10, an oligotroph species also possessing a stalk is highly detected (almost 9%) in one deep ocean metagenome [38]. These observations provide information about microorganisms' life style in some deep oceans. Sulfurimonas denitrificans DSM 1251, previously named Thiomicrospira denitrificans ATCC 33889, is highly detected in two deep ocean metagenomes, and probably plays a

particular role for denitrification in these ecosystems (see figure 20). While these metagenomes provide some information about specific deep ocean locations, a more complete survey has to be done to study all microbial community life styles present at the bottom of the oceans. Unique biodiversities and functional distributions will probably be discovered from these surveys.

Antarctic aquatic environments:

Due to the interest in the Antarctic as an extreme environment and as a global warming indicator, Antarctic is the target of various scientific communities. Especially, microbial communities resisting to these extreme conditions of life are of interest for microbiologists to study specific adaptation capacities. So far some metagenomes from Antarctic aquatic environment were generated and made available for the international community (Ricardo Cavicchioli, Antarctic metagenome program). However, due to a lack of information about the different metagenomes, we decided to not decrypt these nine datasets. This effort to emphasize Antarctic aquatic environment peculiarities will be done in further global comparisons and after the publication of these metagenomes by the Cavicchioli group.

Arctic snows:

As for the Antarctic, Arctic environments are of interest for a wide range of scientists and to study microbial communities adapted to this environment will probably stimulate the discovery of new life styles and adaptations capacities. Two Arctic snow metagenomes were generated so far to explore microbial communities present in this environment (Catherine Larose, Arctic snow program) and sequences related to alginate metabolism, cold shock (CspA family of proteins), nitrate and nitrite ammonifcation, and finally synthesis of osmoregulated periplasmic glucans are unusually represented in the two Arctic snow metagenomes (Figure 21). Interestingly, alginate is known to have a particular role in Pseudomonas aeruginosa biofilm resistance [180]. This specific bacterium is not prevalent in this environment (Figure 3) but it is possible that alginate has an important biofilm resistance role in Arctic snow microbial communities. Unfortunately, actual metagenomic approaches cannot connect functions and species with certitude and so further experiments have to be done to know exactly which species possess an unusual distribution of genes related to this function. It is important to note tha alginate is also known to play a role on osmotic stress and exogenous oxidants.

Overexpression of psychrophilic bacteria CspA proteins in E.coli increases highly cold resistance, so emphasizing the role of these proteins in polar environments survival [181]. The high proportion of sequences related to cspA genes provides news insights about microorganisms' adaptation in this cold environment. Due to the presence of a "cold box"

untranslated region which cause important mRNA instability at 37°C, cspA is inducted only by temperature without any necessity of transcriptional factors in contRAST to heat shock induction [182]. Thus, the unusual presence of this function at the metagenomic level implies not only a response to cold at the RNA level but an adaptation at long term to this environment by replicating this genes or by the selection of species possessing them. Moreover, osmoregulated periplasmic glucans are an important component of gram negative bacterial envelope and have an important role in extreme conditions [183]. They also play crucial roles in pathogenesis and symbiosis [184]. This subsystem is more represented in the two arctic related metagenomes than in all the other compared, so confirming its specific role in this extreme environment.

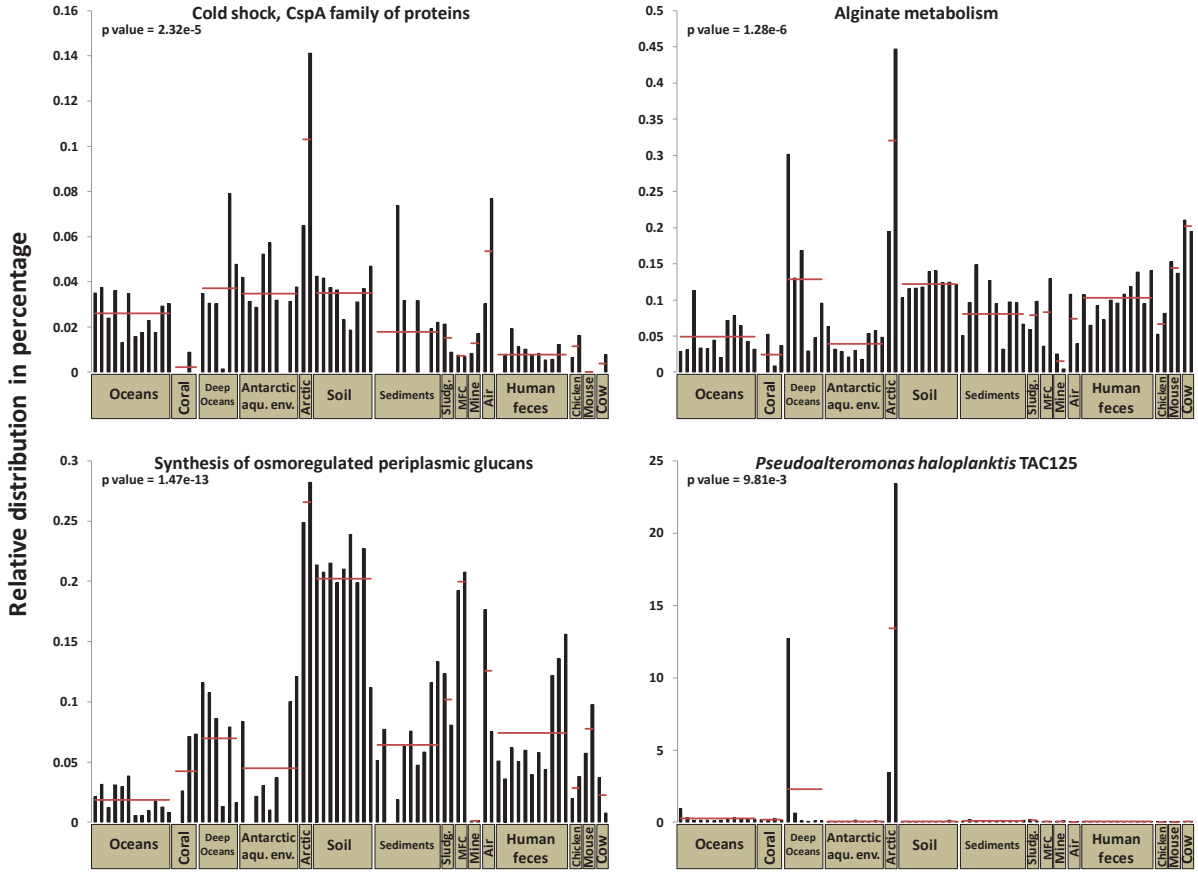


Figure 21. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Pseudoalteromonas haloplanktis is a psychrotolerant bacterium considered as a model organism of cold adapted bacteria [185] and a considerable potential for industrial purposes [186]. The strain TAC125 was isolated from a coastal Antarctic ocean [187] [188] and is able to grow at 0 °C [189]. Due to this unusual capacity, this strain is studied for its bioremediation potential in cold habitats. As an example, the recombinant strain TAC125 can

convert aromatic compounds to catechol molecules in low temperatures when expressing a toluene-o-xylene monooxygenase [190]. Sequences related to this strain are highly detected in three datasets: one from deep ocean [38] and the two corresponding to Arctic snows. In particular, one of them contains more than 23% of annotated sequences related to this strain. The strain Pseudoalteromonas atlantica T6c is also more detected in this environment in comparison to the other (supplement data).

Cytophaga hutchinsonii is highly detected in these two metagenome (Figure 22). This species in known to be an abundant aerobic cellulotic bacterium highly present in soil, but is unexpectedly more detected in Arctic snows. This gram-negative bacterium possesses two unusual characteristics: a rapid gliding motility over surfaces and a capacity to use crystalline cellulose as a source of energy. However, both motility mechanisms and cellulose utilization are not identified in this species [191], and so their distributions cannot be compared. So in this particular case, because the genes involved in these processes are unknown a phylogenetic comparison is indispensable to understand the strategies and life style of a part of Arctic snow microbial communities. In addition, by analyzing unannotated sequences, it could be possible to define clusters present only in these two metagenomes and that correspond to these functions.

Sequences related to Legionella species, involved in human disease are highly detected in one of the two Arctic snow metagenomes (more than 1.5%, higher than the 76 other). This global metagenomic comparison provides complementally information to better understand their environmental sources and risks for human health (see [192] as a review). In addition, Legionella can enter in a viable non cultivable state, be in association with protozoa, and grows in biofilms, so complicating its detection in the environment [193]. So metagenomic approaches (direct DNA extraction) possess particular potentials to detect and quantify these species in the different ecosystems of our planet.

Nostoc sequences are highly detected in the surface of the hypersaline sediment but also in Arctic snows (see figure 26). These Cyanobacteria are known to possess a strategy to respond to desiccation stress by accumulating trehalose [194]. Nostoc commune responds to water deficit by elaborating extracellular glycan [195]. The strain Nostoc commune is even capable to grow after long periods under anhydrobiosis conditions and occur in some deserts. These species appear to be adapted to Arctic desiccation stress due to the form of water not available for microscopic life in this environment, so explaining their important distribution there. This distribution observation provides a new hypothesis: are Nostoc species psychrophiles in cold habitats?

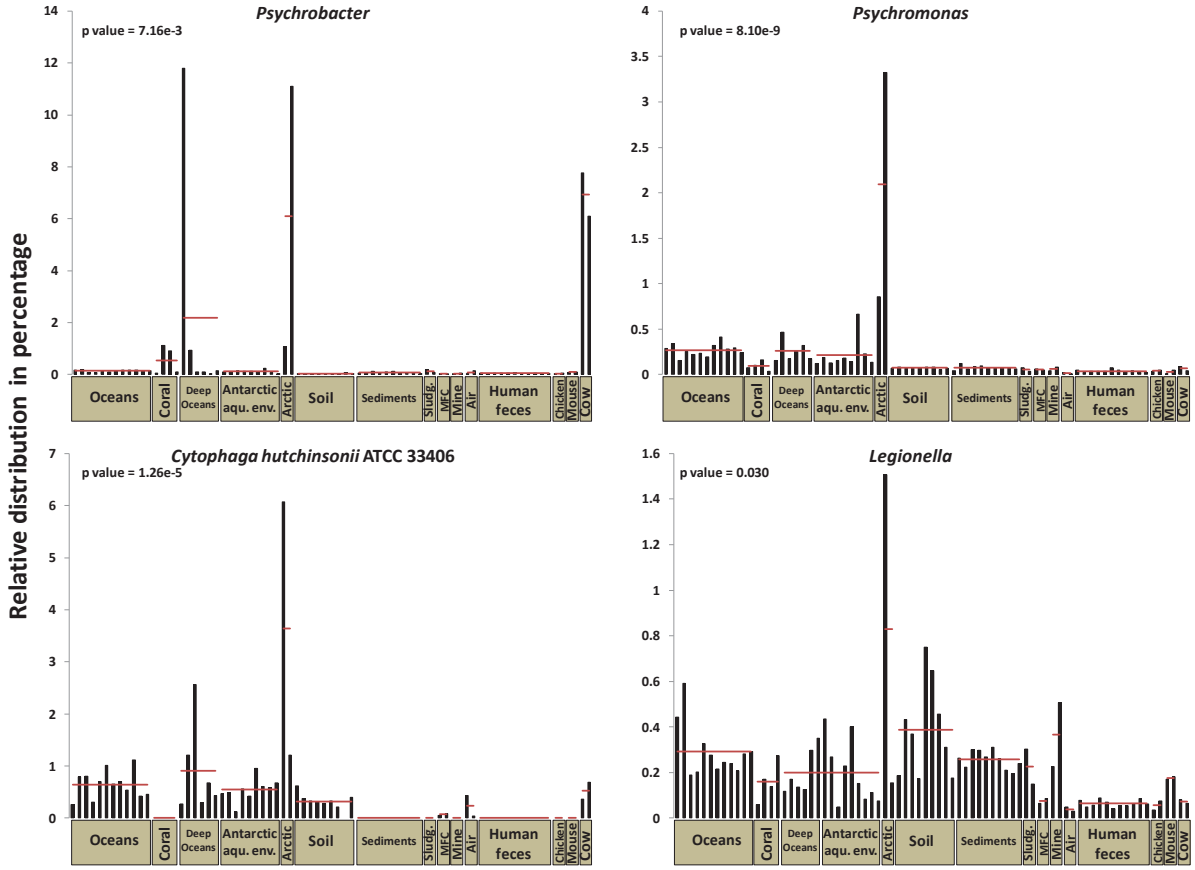


Figure 22. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Psychrobacter species live in cold habitats and can grow at negative temperatures. They were isolated from various environments: Antarctic sea ice [196], deep sea [197], marine environment [198] and Siberian permafrost [199]. Psychrobacter cryohaloentis experiments showed that as a strategy to survive cells increase their adenylate concentration when temperature is decreasing and continue to generated ATP at -80°C [200]. Analyses of the Psychrobacter arcticus 273-4 genome reveal membrane composition changes and synthesis of cold shock proteins as cold survival strategies [201]. Psychrobacter species are highly detected in one deep ocean (the same possessing many sequences related to Pseudoalteromonas), the two Arctic snows, and unexpectedly in the two cow rumens (Figure 22). However, sequences related to CspA family of proteins are correlated to this genus in Arctic snows but not in cows where these proteins are lowly represented. So even if genomes similar to Psychrobacter species are present in these rumens, they have not the same functions and are probably not adapted to cold habitats.

Sequences related to Psychromonas species are as expected more detected in the two Arctic snow metagenomes than in all the other. Psychromonas ingrahamii for example was isolated

from an Arctic sea ice and possesses the particularity to grow exponentially at -12°C [202]. Its genome was sequenced [203] and reveals some unusual characteristics (e.g. important proportion of regulators of cyclic GDP, "orphan" hypothetical proteins, three-subunit TRAP systems).

Psychrophile cells are clearly over-represented in Arctic snow communities when compared to other environments. These microbial communities are adapted to cold by using specific strategies (e.g. cold shock proteins and desiccation resistance). Some of them potentially form biofilms and are pathogenic, other probably play a particular role in nitrogen cycle (see figure 17). Liquid water can exist as low as -20°C [204], an overlap with Mars temperatures for example. So this adaptation to snow habitat provide unique information about life potential in extreme environments and so in Mars subsurface or specific exoplanets. The question is: could extreme environment life emerge without a first evolution in a more life friendly (e.g. more temperate) habitat.

Soils:

Soil is considered as the most biodiverse environment on Earth with an estimation of diversity varying between 104 [205] and 107 [206] species per gram and even more when using different DNA extraction protocols to stimulate the accessibility of the soil metagenome [62]. This environment is of interest to study the impact of intensive agriculture, global warming and human induced pollutions on soil microbial communities, but also to discover genes of interest (e.g., new antibiotics and biodegradation capacities). Three soil metagenomes are actually available to the international community and were selected to represent this environment. They are located in Central America (Luquillo Rain Forest Soil, Puerto Rico, unpublished), North America (Waseca farm soil, [38]), and England (Rothamsted, under submission).

In these soils, sequences related to benzoate transport and degradation cluster, central metacleavage pathway of aromatic compounds, salicylate and gentisate catabolism, phenylpropanoid compound degradation, bacterial cAMP signaling, nitric oxide synthase [207], PQQ dependent quinoprotein dehydrogenase (figure 23), and almost as much as in Arctic snow, synthesis of osmoregulated periplasmic glucans (see figure 21) are relatively more detected in comparison to the majority of the other environments.

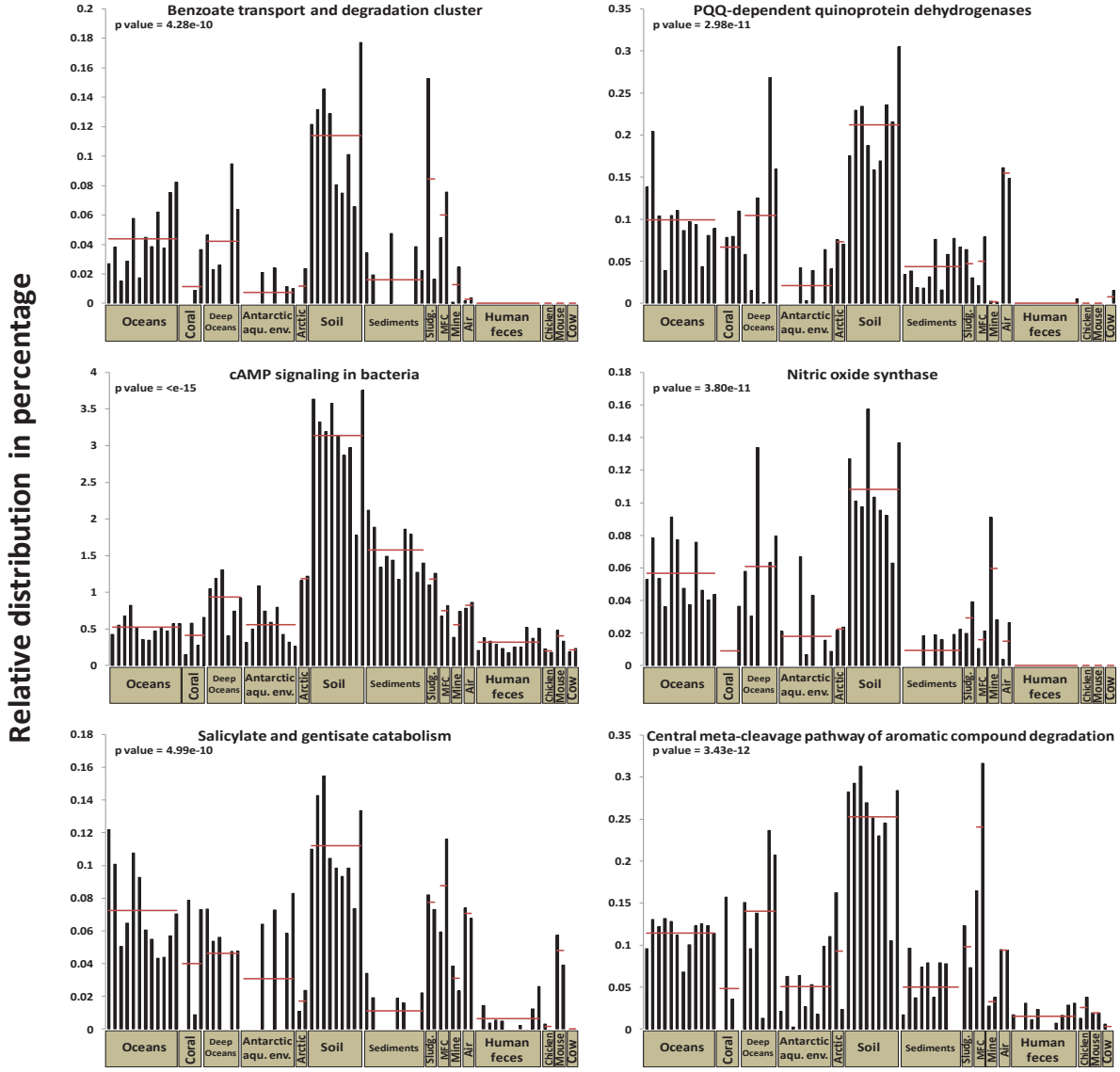


Figure 23. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Based on these observations, this environment is more involved in the catabolism of aromatic compounds than the other. Gentisate molecule for example serves as a key intermediate for the aerobic metabolism of many aromatic compounds (e.g. conversion of naphthalene to central metabolites via gentisate, [208]). A part of these compounds are secreted by plants. For example, methyl benzoate molecules are emitted by flowers to attract pollinators, with a maximal emission during the day [209]). But plants can also secrete aromatic compounds in defense against pathogenic species (e.g. phenylpropanoid compounds, [210]). So soil microorganisms probably degrade these compounds to generate energy or in reaction to plants defense.

Sequences related to bacterial cAMP signaling are also more represented in soil. cAMP is an important second messenger in both eukaryotic and prokaryotic organisms. Interestingly, as a cAMP subversion mechanism, some bacterial pathogens inject adenylate cyclase protein toxins into plants to increase host adenylate cyclases rate [211]. For example, genes involved in cAMP signaling mechanism are overrepresented in Mycobacterium species and have specific roles in pathogenesis [212]. Thus soil bacterial and archaeal communities possess a considerable cAMP subversion system potential for plant signaling mechanisms deceiving.

Nitric oxide synthases are multidomain metalloproteins present like for cAMP signaling in both eukaryotic and bacterial organisms. In Bacteria, nitric oxide can play specific roles in oxidative stress and radiation damage protection, but also in toxin biosynthesis [213]. The subsystem representing this function is detected in several ecosystems but is more represented in soil related metagenomes than in all the other compared environments.

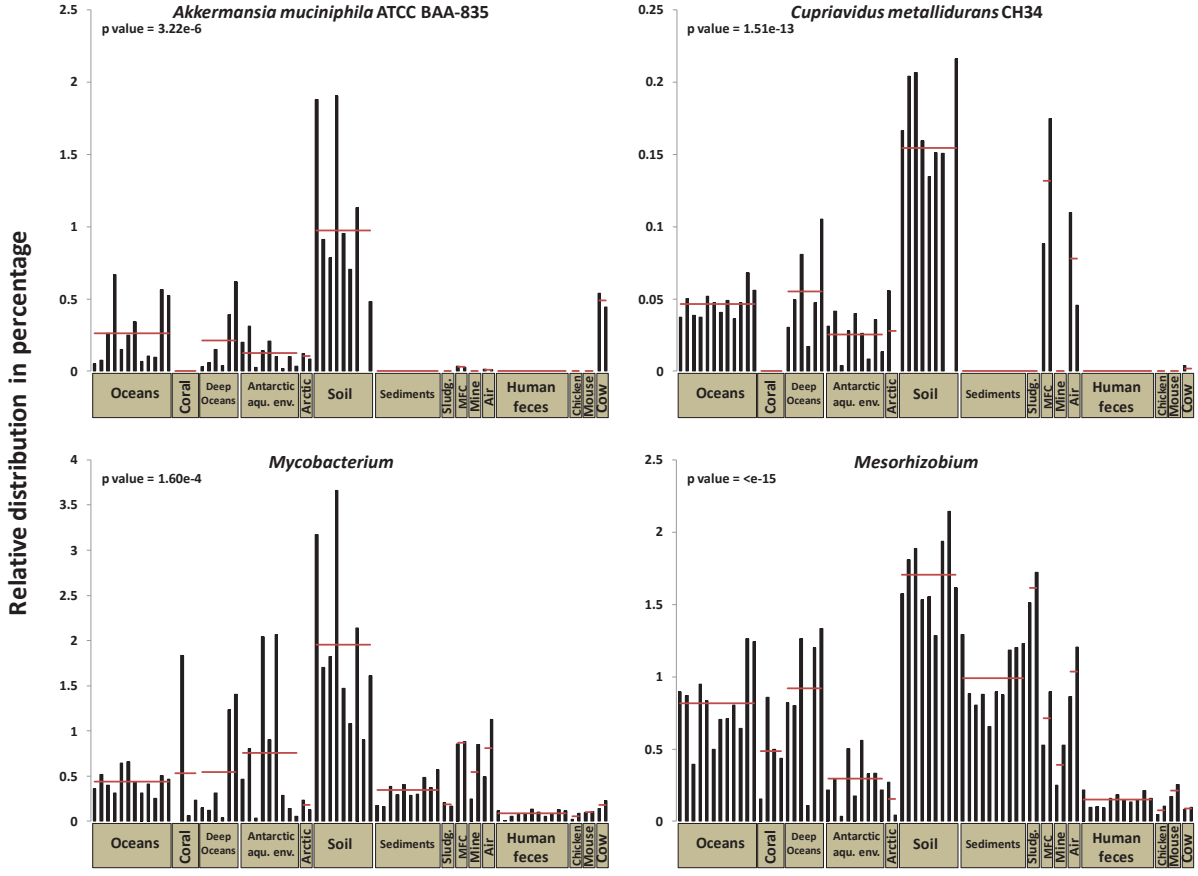


Figure 24. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Quinoproteins are mainly involved in the direct oxidation of alcohols, sugars and amines [214]. For example, quinoprotein dehydrogenases use the pyrroloquinoline quinone (PQQ) cofactor to catalyze the oxidation of alcohols. However the mechanisms involved are not completely understood [215]. The important detection of sequences related to PQQ dependent quinoprotein dehydrogenase in soil emphasizes another specificity of the communities present in this environment and make this system a good matrix to study more in detail mechanisms involved on this pathway. In addition to unusually distributed functions, some genera appear to be uncommonly detected in soil related metagenomes and some of them are represented in the figure 24.

Akkermansia muciniphila was isolate from a human intestine [216] using mucin as the sole carbon and nitrogen source. This bacteria is a gram negative strictly anaerobic. Using a 16S rRNA-targeted probe, this organism was estimated to represent from 1 to 3% of the total human feces microbiota [217]. However, this bacterium appears to be more represented in soils in comparison to human feces where it is undetected. Sequences related to this species are also present in cow rumens, oceans, deep oceans, Antarctic environments, Arctic snows, and in microbial fuel cell anode biofilms even if its distribution is very low. Because this bacterium was isolated from human microbial populations, it was only tracked in this specific environment and so its role in other environments is not studied. Experiments using soil and a culture with mucin should be performed to confirm these observations. If this microorganism can be isolated from this environment, to sequence its genome could emphasize genetic differences and be a good model to study adaptation mechanisms when one microorganism colonizes human intestines.

Cupriavidus metallidurans was isolated from a metal processing factory [218] and recently sequenced [219]. This strain is highly resistant to Zn2+, Cd2+, and Co2+, and genes conferring these resistances are located in two plasmids. Sequences related to this strain are more represented in soils. Interestingly, sequences related to cobalt-zinc-cadmium resistance are also largely detected in soils (see figure 14) and could for a part belong to this species.

Intercellular communications are used by microorganisms to modulate cellular processes using chemical signals. But these signaling processes can also be used between bacteria and eukaryotes (e.g. soil bacteria and legumes, [220]). Rhizobium, Mesorhizobium, Sinorhizobium and Bradyrhizobium genus have the capacity to colonize leguminous plant roots and to induce the formation of nodules (see [221] as a review). These benefic organs cannot be formed by these plants without this nitrogen-fixing symbiotic association. Genes involved in this symbiosis are largely linked to quorum sensing and are present in plasmids in Rhizobium species and in chromosomic islands in Bradyrhizobium, Azorhizobium, and Mesorhizobium (see [222] as a review). Sequences related to these four genera are common in soils (Figures 24, 25 and supplement data). However, only Bradyrhizobium is overrepresented in this environment in comparison to the 14 other. The metagenome

possessing the higher distribution of sequences related to Rhizobium and Sinorhizobium was extracted from a microbial fuel cell.

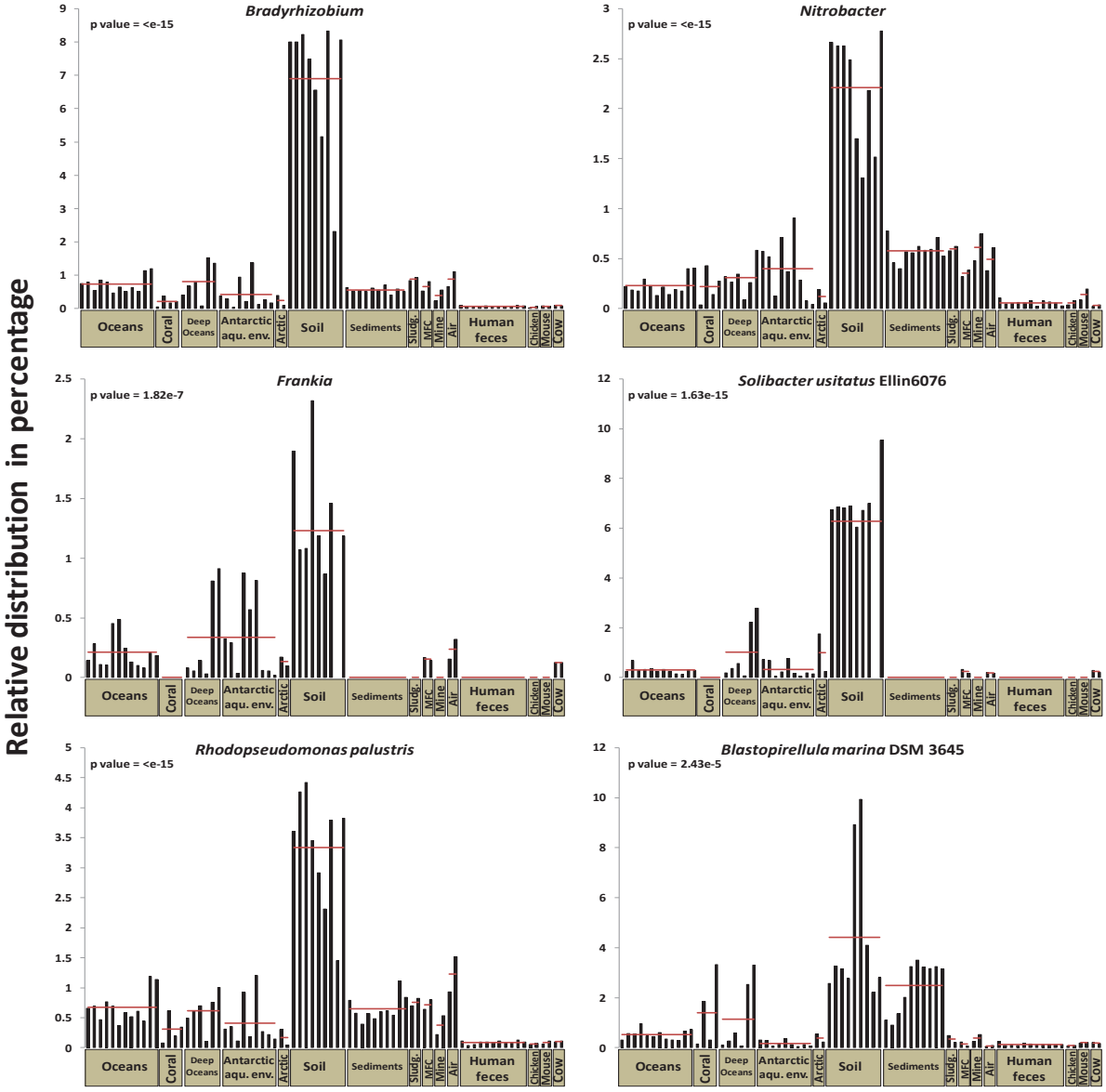


Figure 25. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Frankia related species possesses the ability to colonize and induce N2-fixing root nodules in other plants, named Actinorhizal (see [223] and [224] as reviews). These plants, helped by Frankia cells for the fixation of nitrogen, are known to grow in poor soils and to be pioneers in plant community development in extreme environments (e.g. Arctic tundra) but also in new habitats (e.g. deglaciated soils, [225]). Sequences related to the Frankia genus are in

majority more detected in soils than in the other environments, but are also present in Antarctic aquatic environments and Arctic snows for example, where plants are absent.

Mycobacterium genus comprises pathogenic species and so is well studied by scientists. It is the case of the well known M. tuberculosis and M. leprae, but also several other species which can be at times deadly pathogens (see [226] as a review). Due to this risk for human health, genomes corresponding to several species were already sequenced, and the distribution of these microorganisms can be tracked among the different metagenomes. Globally, sequences related to the Mycobacterium genus are more detected in soils than in all the other environments. However, as a function of the species observed, the predominant ecosystem can varied. As an example, both M. avium paratuberculosis (str. k10) and M. leprae TN are more detected in one coral atoll metagenome [47]. But other species, like M. marinum M and M. smegmatis (str. MC2 155) are more present in soils (supplement data).

Nitrobacter species are detected in all compared metagenomes but appear to be lowly represented in animals and highly present in soils (Figure 25). They are known to oxidize nitrite into nitrate in soil and so play an important role in nitrogen cycle. However, based on subsystem distributions, sequences involved in this process are not particularly more represented in this environment (see figure 17). As a consequence, other microorganisms are more involved in the nitrogen cycle elsewhere than in soil.

Rhodopseudomonas palustris is a purple bacterium and a model organism to study aromatic compound degradation mechanisms [227]. It possesses a wide range of metabolic pathways, and to sequence its genome revealed the presence of three nitrogenases and five benzene ring cleavage pathways [228]. This species is more detected in soils than in the other environments. It is probably in part due to its capacity to acquire carbon from green plant-derived compounds (e.g. dicarboxylic acids, [229]). The distribution of sequences related to this species can explain for a part the important presence of sequences involved in aromatic compounds catabolism in soil (e.g. benzoate transport and degradation cluster).

Sequences related to Blastopirellula marina, usually isolated form aquatic environments ([230] are unexpectedly more represented in soils and sediments than in oceans for example. However sequences matching with this species possess high e-values in contRAST to those matching with Bradyrhizobium for example (Delmont et al., under submission). It is highly probable that a majority of microorganisms predominant in soils are not yet sequenced and this gap of data induces biases during the SEED annotation.

Solibacter usitatus [231] is highly detected in two of the three soils, undetected in the third. In contRAST, sequences related to Novosphingobium aromaticivorans [232] and Deinococcus geothermalis (see figure 28) are more detected in the third, so emphasizing soil specificities even if the distribution of the majority of functions and species are similar in the compared soil datasets.

Obviously, soil microorganisms are specialized on aromatic compound metabolism (and in particular catabolism) by using plant secreted molecules as substrate. They can also for some of them (e.g. Mycobacterium species) use specific mechanisms to attack (cAMP signaling, nitric oxides, osmoregulated periplasmic glucans) or to be in symbiosis with plants (e.g., Frankia), and potentially degrade plant defense molecules.

Hypersaline sediment:

To investigate a stratified hypersaline microbial mat, ten different depth horizons where sequenced from a hypersaline sediment between the surface to 49 millimeters depth [50]. This works aimed to study both fine scale variations in microbial communities and their adaptations to an extreme environment (especially salinity and radiations). The metagenomes are represented from the surface (left) to the deeper horizon (right) in all graphs. A wide range of genera and functions appear to be more represented in these sediment horizons than in the other environments. Some of them decrease with depth, other increase or are stable among the different metagenomes.

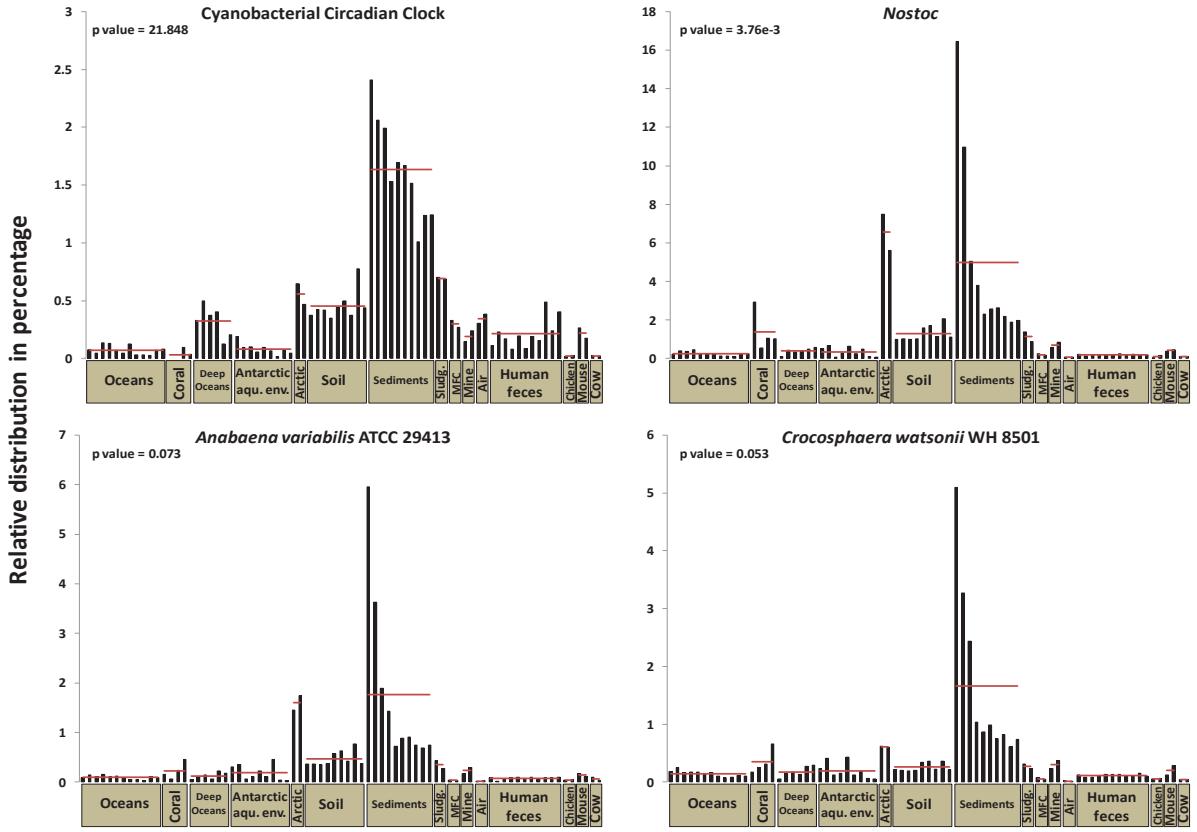


Figure 26. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Sequences related to genes involved in the Cyanobacterial Circadian Clock [233] are prevalent in the surface of this sediment (Figure 26). This process confers an adaptative advantage by synchronizing the cell with the environment [234]. In addition, Nostoc punctiforme, Anabaena variabilis and Crocosphaera watsonii are also highly detected in this sediment (Figure 26). These three species are affiliated to the Cyanobacteria phylum (see figure 1) and are known to fixe nitrogen [235-237]. For example, Anabaena variabilis fixes nitrogen and CO2 and produces hydrogen during photosynthetic process.

The distribution of genes involved in Cyanobacterial Circadian Clock is decreasing with sediment depth and is correlated to the distribution of the Nostoc genus, Anabaena variabilis and Crocosphaera watsonii which decrease also with this variable. However, these genes are lowly represented in oceans and especially in coral atolls where other Cyanobacterium (Synechococcus and Prochlorococcus) are highly represented (see figure 19). So genes involved in this process are not ubiquitous in this phylum. This result is confirmed by recent cultural observations suggesting that the genus Prochlorococcus is deleted by one of the three principal genes involved in the circadian clock in Cyanobacteria [238].

The rapid decreasing distribution of Cyanobacteria in this sediment can easily be explained by the necessity they have to live in environments exposed to solar radiations to perform photosynthesis. Because they possess various UV radiations defense mechanisms and use these radiations to generate energy, they have a clear advantage and so are predominant in the first millimeters of the sediment.

In contRAST to sequences related to the Cyanobacteria phylum decreasing with depth, some species are highly detected in this sediment in comparison the other environments, but with a distribution increasing with depth (in relative percentage). It is in particular the case of Methanosarcina genus, Methanopyrus kandleri, Chloroflexus aurantiacus, Haloarcula marismortui, Natronomonas pharaonis and Halobacterium sp. NRC-1 (Figure 27).

Methanosarcina species are known to generate energy by converting CO to methane [239] [240] but can produce other molecules. As an example, Methanosarcina acetivorans produces acetate and formate in addition to methane during CO-dependent growth [241]. These species play a role in carbon cycle and so have a particular potential to counter human industry CO generation.

The methanogen Methanopyrus kandleri was isolated from a black smoker at 2OOO meters deep where temperature is from 84 to 110 °C [242]. It is for the moment the only representative of the Methanopyrus genus. Interestingly, even if this sediment in not a thermophilic environment, enzymes isolated from M. kandleri require high salt concentrations for stability and activity [243] [244]. This particularity can explain the important distribution of sequences related to this species in the deeper sequenced horizons of this hypersaline ecosystem.

Chloroflexus aurantiacus is a phototrophic gliding filamentous bacterium isolated from hot springs [245]. This thermophilic photosynthetic bacterium can grow using CO2 as sole carbon source [246].

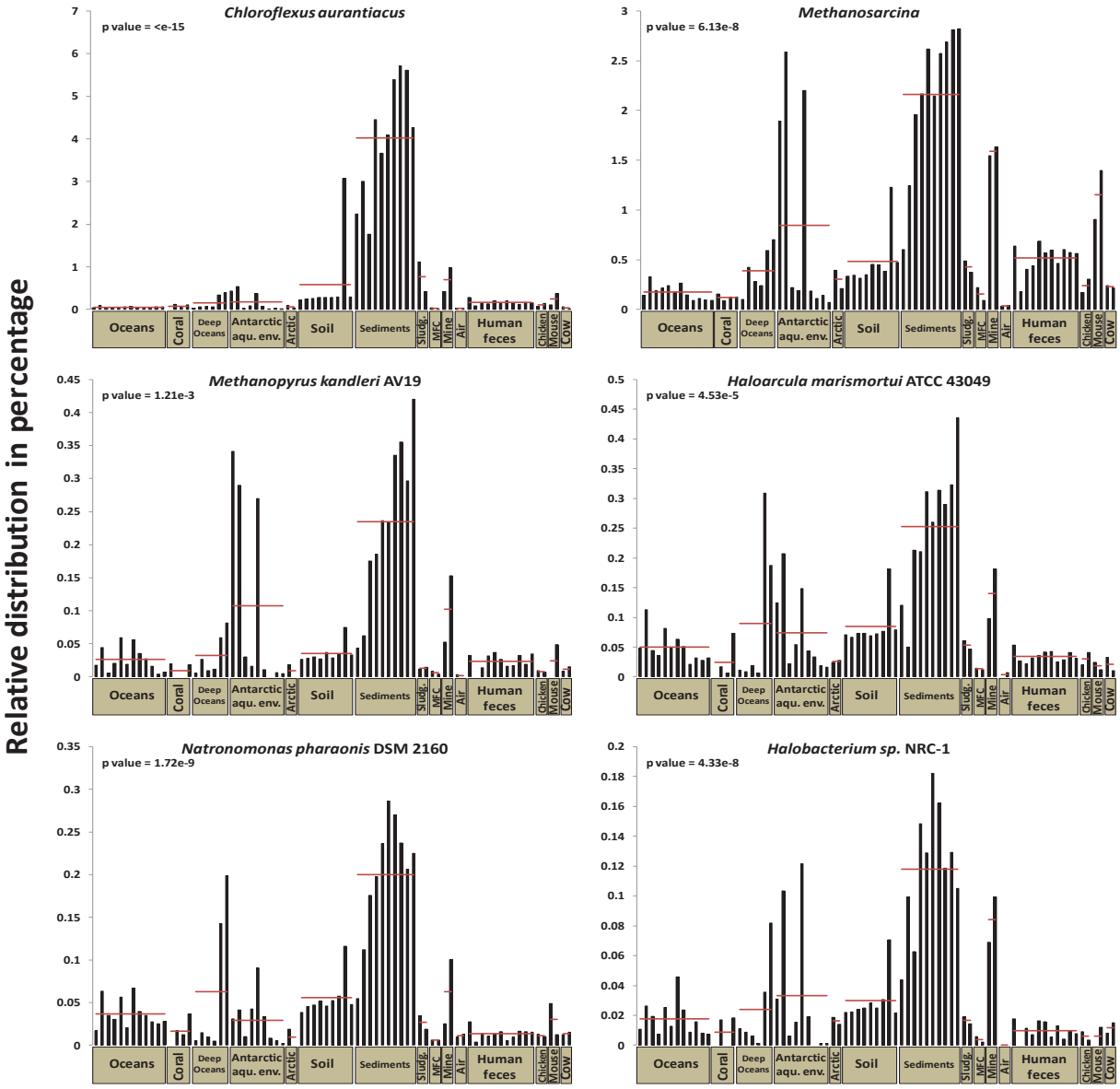


Figure 27. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Natronomonas pharaonis is an archaeon adapted to both alkine pH and high salt concentrations. It was first isolated from saline lakes, then sequenced [247] and is now a model to study haloalkaliphile life (e.g. [248]).

Halobacterium sp. NRC-1 is a halophilic archaeon which grow optimally near the NaCl saturation concentration (4.5 M, [249]). This strain is also highly resistant to UV-radiations [250]. Its genetic structure is composed of a large genome and two minichromosomes [251]. It possesses also 91 insertion sequences which provide a considerable flexibility to this organism. Because of its extreme life style, this halophile is a model among the archaea domain and is intensively studies to improve scientific knowledge about elemental cellular processes (e.g. [252]).

Haloarcula marismortui is a halophilic red Archaeon isolated from the Dead Sea [253]. Its genome is organized into nine circular replicons, and its comparison to Halobacterium sp. NRC-1 genetic structure suggests a common ancestor for these two species [250].

Sequences related to Pelobacter carbinolicus, Desulfotalea psychrophila, Chlorochromium aggregatum and Deinococcus geothermalis are also unusually represented in this environment (Figure 28).

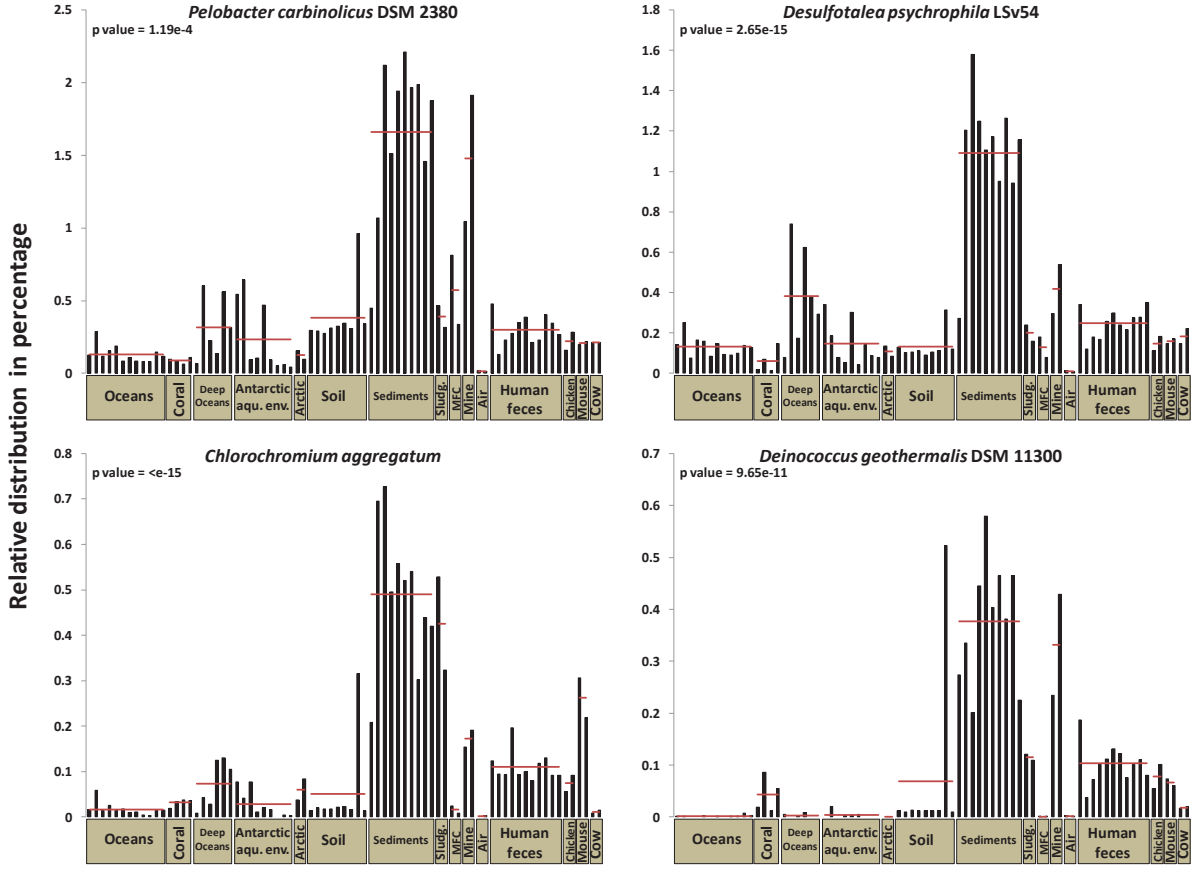


Figure 28. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

The strict anaerobe and delta-proteobacterium Pelobacter carbinolicus is known to use only 2,3-butanediol, methylacetoin, acetoin, and ethylene glycol to grow [254] and to reduces Fe(III) indereictly via sulfide production [255]. Interestingly, Pelobacter species are known to grow fermentively and to generate hydrogen for consumption by methanogens. So this species can have a positive impact for methanogens (e.g. Methanosarcina and Methanopyrus, see figure 24).

The gram negative and sulfate-reducing delta-proteobacterium Desulfotalea psychrophila was isolated from cold marine sediment [256] and can grow in water below 0°C. Its sequence revealed an unusual presence of genes involved in the TCA cycle, and other coding several two-component regulatory systems and especially putative cold shock proteins [257].

The phototrophic consortium Chlorochromatium aggregatum consists of green sulfur bacteria (epibiont cells) surrounding a motil chemotrophic bacterium [258] [259]. It represents the most developed interspecific association of bacteria known and so is a unique model to study signal transduction mechanisms and bacterial coevolution [260].

Deinococcus geothermalis is an extremely radiation resistant strain first isolated from hot springs [261]. In addition to its unusual high resistance to radiation, it grows optimally between 45 and 50°C. So this strain has a unique potential for bioremediation in high temperature radioactive mixed waste ecosystems but needs the acquisition of plasmids possessing genes involved in this process [262]. Finally, the sequencing of its genome [263] provided new insights about DNA repair systems with a possible important role of chromosome order alignment structure in resistance to radiations.

In addition to species highly detected in this sediment, few functions are unusually highly detected and provide unique insights about functional specificities of these microbial communities. It is in particular the case of sequences related to tungsten ABC transporter, carbon monoxide induced hydrogenase, methanogenesis from methylated compounds, pyrrolysine (Figure 29), reductive dechlorination, hexose phosphate uptake system, sigmaB stress response regulation and finally zinc resistance (Figure 30).

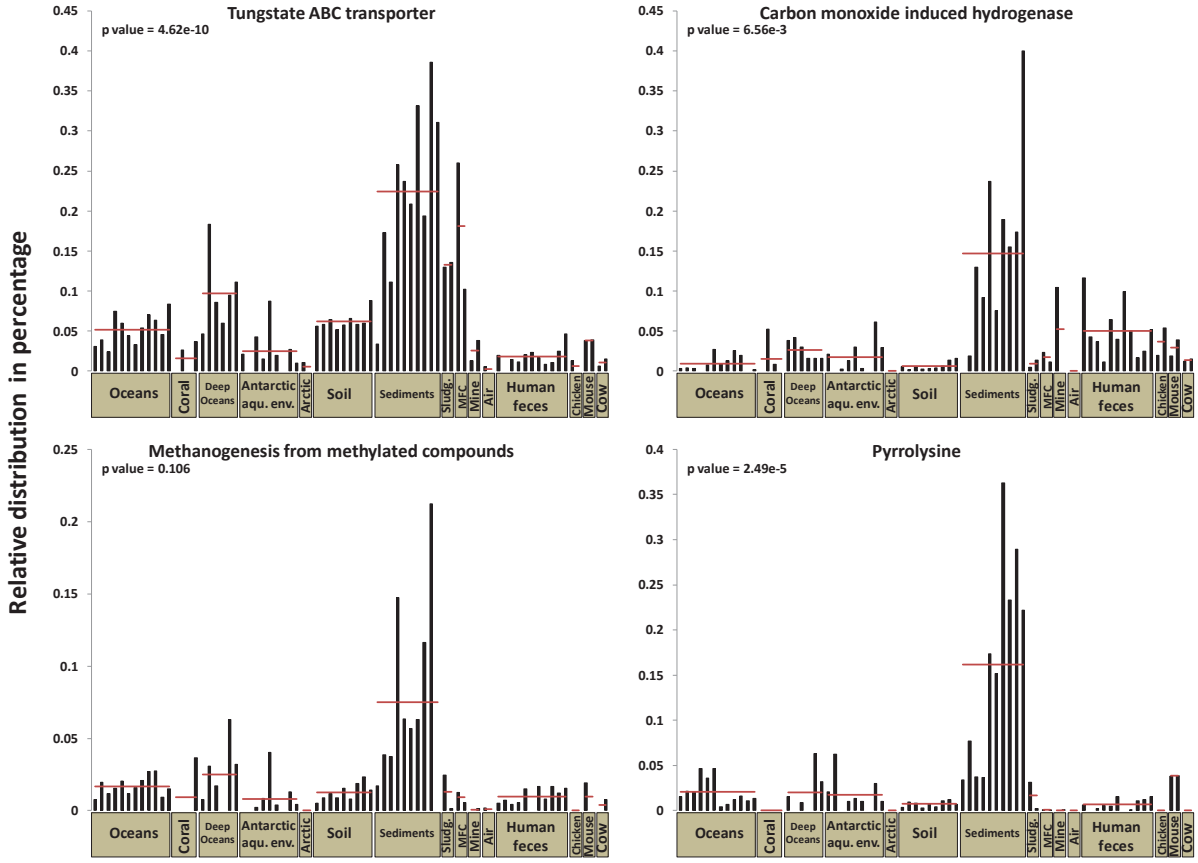


Figure 29. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

ABC transporters are the most important family of membrane transport systems and control the translocation of substrates across the membrane by hydrolyzing ATP to by-pass negative concentration gradients [264]. Tungstate ABC transporter is a specific transporter group present in some cultivable microorganisms (e.g. Methanococcus maripaludis, Campylobacter jejuni, Methanosarcina acetivorans [265-267]. In addition, some species cannot grow without the presence of tungsten (e.g. Pyrococcus furiosus, [268,269]). Interestingly, the distribution of Methanococcus maripaludis, Methanosarcina acetivorans and the genus Pyrococcus is also highly represented in this environment in comparison to the majority of the other compared ecosystems (supplement data). However, it is not the case of Campylobacter jejuni which is relatively lowly represented in this hypersaline sediment.

Sequences related to carbon monoxide (CO) induced hydrogenase are also highly represented in this environment and increase with depth. In spite of a high toxicity for the majority of living matter on our planet, carbon monoxide (CO) can be use as a source of carbon for some organisms [270]. These microorganisms are known to possess this specific function which allows them to grow in a CO dependant manner in the dark by catalyzing the

reaction: CO + H2O --> CO2 + H2 (e.g. Rhodospirillum rubrum, [271,272]). No correlations were found between the distribution of Rhodospirillum rubrum and this function, but it is clear that other species are involved in this process in the deeper horizons of the studied sediment.

Methanogenesis is a process occurring in methanogen organisms to generate methane. However, a majority of methanogenic bacteria use very restricted substrates for methane production (generally hydrogen and carbon dioxide, acetate, methanol, or formate; [273]. But few species are known to use N-methyl compounds (e.g. Methanosarcina barkeri; [274]) and to reduce methylated sulfur compounds (e.g. dimethyl sulfide (DMS) molecules, [275]) to generate methane. DMS molecules can be release by bacteria into the ocean (important distribution of sequences related to DMSP breakdown in oceans, see figure 18) and have the potential to play an important role in cloud formation [161]. But the important distribution of sequences related to methanogenesis from methylated compounds in the studied sediments can limit the release of DMS in the atmosphere and emphasize another cycle for these compounds with the formation of methane, what confirms precedent conclusions of Kiene and collaborators.

Pyrrolysine was discovered in 2002 [276,277] as the 22nd amino acid in a Methanosarcina barkeri methyltransferase. This protein initiates methanogenesis and carbon assimilation from methylamines, and this amino acid appears to be essential for methane formation from these substrates [278]. Interestingly, Methanosarcina species (including M. barkerii) are also highly detected in sediments in comparison to the majority of the other environments (Figure 24). So in the base of Methanosarcina barkerii species, pyrrolysine, and methanogenesis from methylated compounds functional distributions in these sediments, it is clear that this species (or species possessing a genome equivalent) and these two functions are involved in the production of methane using methylamine substrates, especially in the deeper sequenced horizons (lower proportion near the surface). Moreover, this correlation emphasizes common results between cultural and metagenomic approaches. In addition, the oxidation of CO is known to be coupled with methanogenesis [270]. So sequences related to carbon monoxide induced hydrogenase can also play a particular role in this process. However, in spite of positive distribution correlations, information is insufficient to conclude about a hypothetical role of Tungstate ABC transporters.

In addition, the sigmaB stress response regulation modulates the stress response of several Gram-positive bacteria (e.g. Bacillus subtilis). Protection against membrane, protein and DNA damages could be one of the principal roles of the sigmaB stress response regulation [279] . The high proportion of sequences related to this stress response is probably due to the important presence of species known to be resistant to radiations.

Finally, the important distribution of sequences related to nitrosative stress, reductive dechlorination process and zinc resistance emphasizes additional adaptations of these microbial communities in an extreme environment.

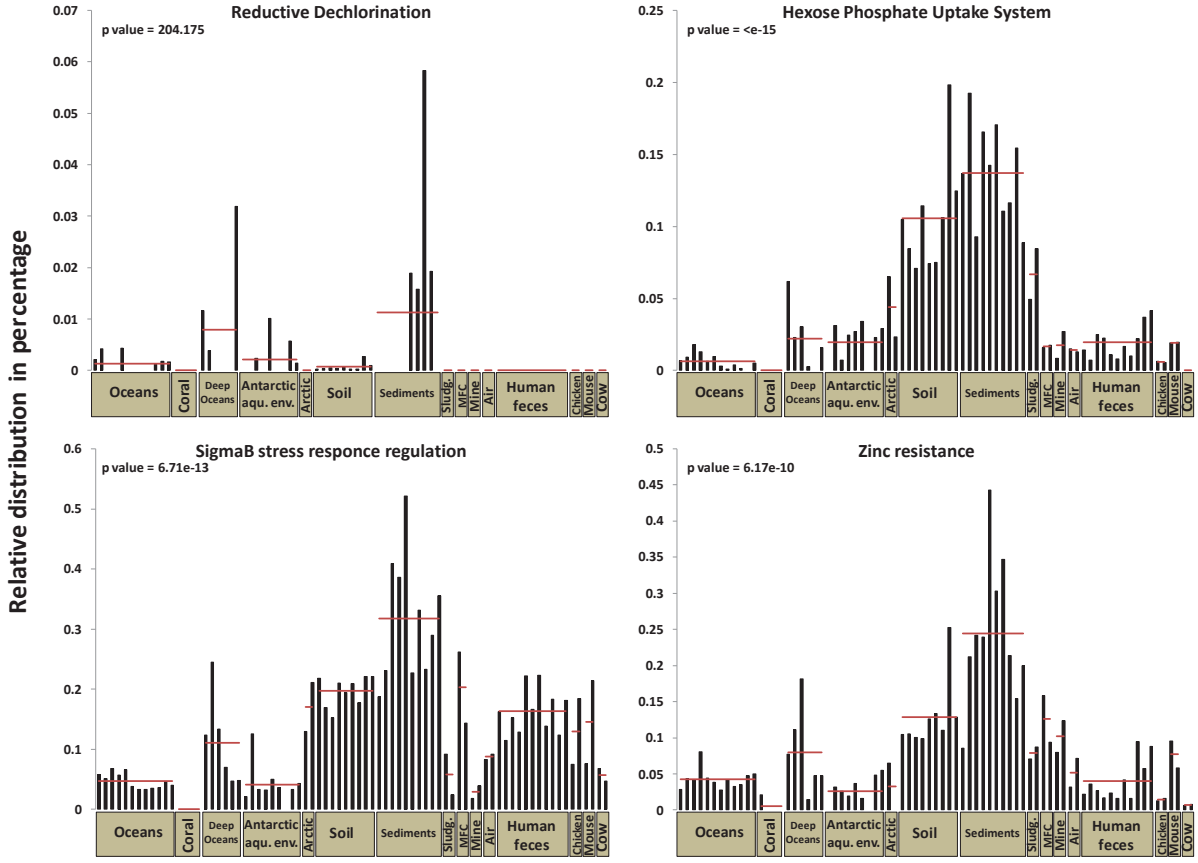


Figure 30. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Based on these metagenomic observations (Functional and taxonomical distributions) among the different horizons of this hypersaline sediment and the different environments compared, it is clear that this ecosystem possesses unique characteristics reflected by the presence of unusual extremophilic microorganisms and functions. Obviously these organisms are highly adapted to high concentration of NaCl and radiations. In addition, the distribution of sequences related to methanogenic species (e.g. Methanosarcina) emphasizes a particular role in carbon cycle in this sediment.

Activated sludges:

Enhanced biological phosphorus removals (EBPR) are wastewater treatments designed for the removal of phosphate in activated sludges. Phosphorous is a major nutrient contributing

to lakes and natural waters eutrophication and to control its concentration is essential to limit its effects on the environment. Microbial communities present are involved in this process and to study these microorganisms (which species and functions) is crucial to optimize treatments. In this aim, two metagenomes corresponding to communities from two distinct EBPR (in Australia and North America) were generated and studied [280]. Based on this study, the uncultured Candidatus Accumulibacter phosphatis was found to be the predominant phosphate removal agent (respectively 80% (left) and 60% (right) in the graphs). The correspondent genome was partially assembled and provides insights about mechanisms involved in phosphate removal. In these two metagenomes, sequences related to dissimilatory nitrite reductase (see figure 17), inorganic sulfur assimilation, membrane-bound Ni, Fe-hydrogenase, mercury resistance operon (see figure 14), orphan regulatory proteins and phenylpropionate degradation are unusually highly represented (see figures 31 and 33).

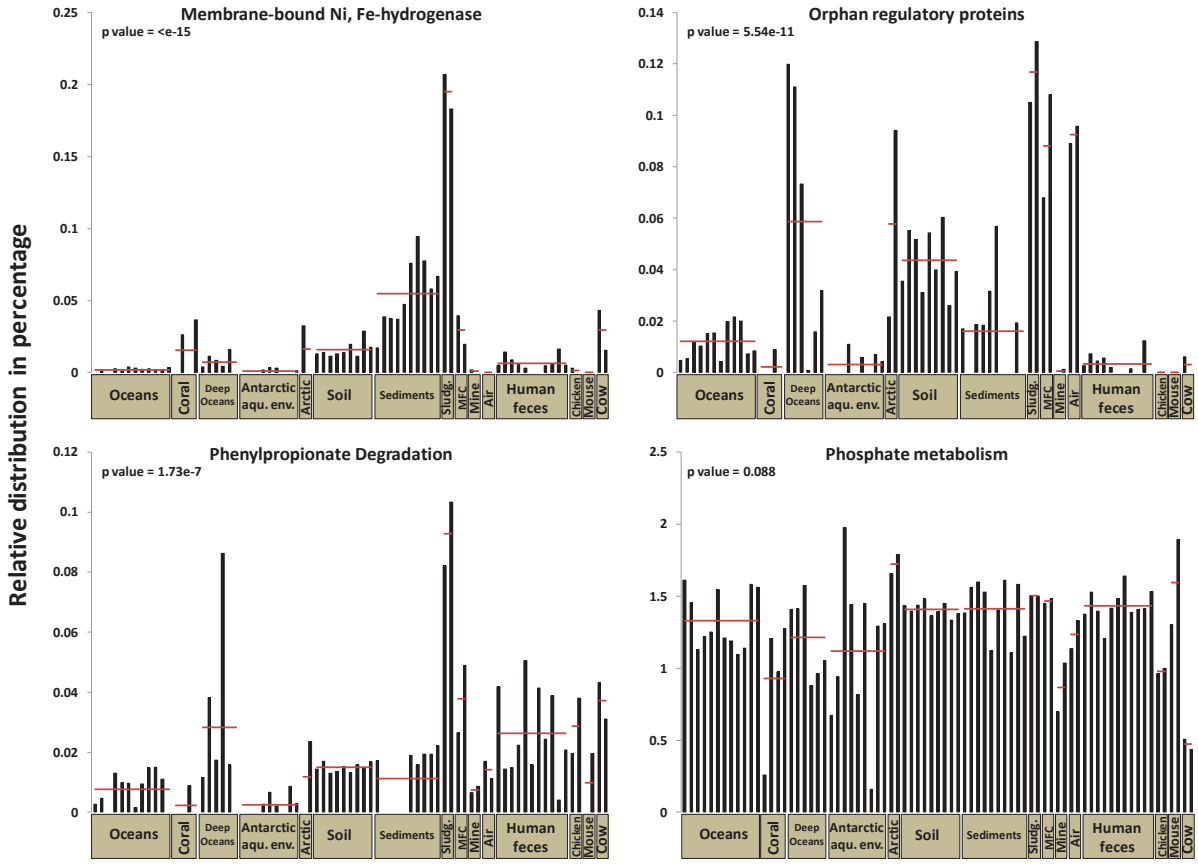


Figure 31. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

The two microbial communities are highly similar in spite of different localizations. This artificial ecosystem (industrial processes) possesses unique microbial communities (see

figure 11), with a particular involvement in specific nitrogen and sulfate transformations. However, in spite of studies emphasizing the predominance of the uncultured Candidatus Accumulibacter phosphatis to remove phosphate in sludges (e.g. [280,281]), known functions related to phosphate uptake and metabolism are not unusually more represented in these metagenomes (Figure 31 and supplement data).

NiFe-hydrogenases are enzymes which catalyze the oxidation of H2 to protons and electrons. This process is generally sensitive to dioxygen [282]. However, the membrane-bound NiFE-hydrogenase is an O2 tolerant dehydrogenase and is well studied in Ralstonia eutropha for example [283]. Sequences related to this function are more detected in this environment. Interestingly, Ralstonia eutropha JMP134 and Nitrosomonas eutropha C71 are also highly detected in these metagenomes in comparison to other. However, these two species are common in soils and acid mine drainage biofilms without a corresponding proportion of this function. So these sequences can be related to uncultured bacterium, so confirming the previous results [280].

The important presence of sequences related to orphan regulatory proteins, involved in quorum sensing processes [284], provides information about microbial interactions in these sludges. Phenylpropanoid compounds are a breakdown product of plants and can be degraded by several species (e.g., [285-287]). The distribution of the subsystem related to this process could reflect an important concentration of phenylproponiate compounds in these EBPR. Finally, the important presence of sequences related to a mercury resistance operon can potentially be the consequence of an unusual presence of mercuric compounds in these wastewaters due to human pollution.

In spite of a predominance of uncultured species in this environment, to observe already genomes sequenced matching with these sequences can provide helpful information about similarities between Candidatus Accumulibacter phosphatis and other bacterial groups. In addition, because this species is more represented in one of the two metagenomes (the second in the figures), all species more detected in the other metagenome don't correspond to A. phosphatis. Sequences related to species unusually highly represented in this environment and positively correlated to A. phosphatis are Dechloromonas aromatica RCB, Azoarcus sp. EbN1, the Burkholderia genus, the Ralstonia genus, Rhodoferax ferrireducens and the Xanthomonas genus (Figure 32). Not all of these sequences are related to A. phosphatis, but the predominance of sequences related to Dechloromonas aromatica RCB for example is probably due to a similarity between a part of genomes corresponding to it and A. phosphatis.

Dechloromonas aromatica RCB is a gram negatve bacterium first isolated from a river sludge contaminated by benzene, toluene, ethylbenzene and xylene compounds [288] and known to be found in soil environments. It possesses the particularity to degrade anaerobically chlorobenzoate. Azoarcus strain EbN1 is an anaerobic aromatic-degrading and denitrifying bacterium [289] metabolizing numerous aromatic compounds. As a particularity in

comparison to other Azoarcus species, this strain does not possess genes involved in nitrogen fixation. The Burkholderia genus represents numerous beta-proteobacteria known to occur in various ecosystems (see [290] as a review) and possess particular characteristic, from human, animal and plant pathogeneicity to biodepollution. For example, Burkholderia mallei is the causative agent of Glanders, Burkholderia pseudomallei of Melioidosis (see [291] as a review). These two species are studied in part due to their potential for bioterrorism utilities, and appears to be unusually highly represented in these sludges.

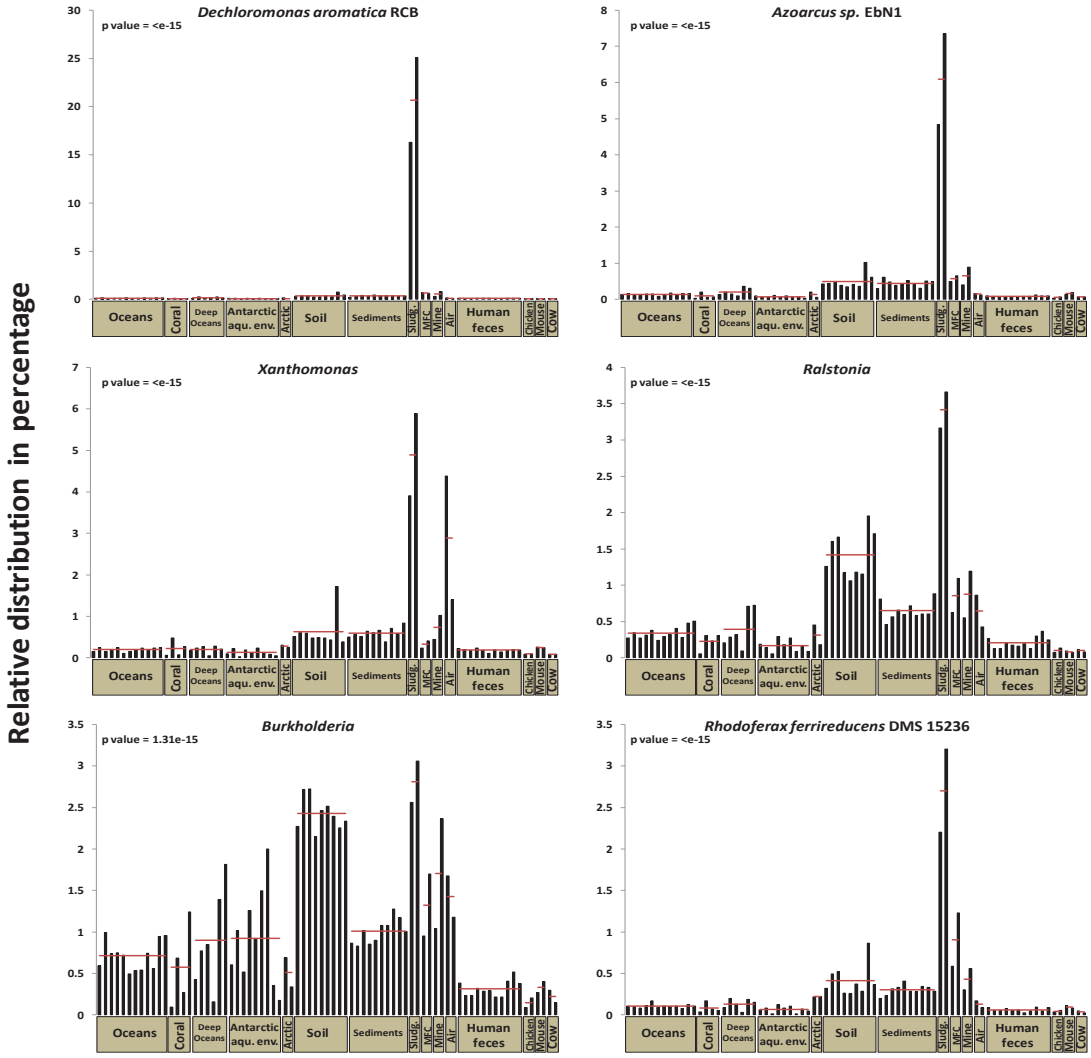


Figure 32. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Ralstonia eutropha, metallidurans, and solanacearum are highly detected in these metagenomes. These species are known to be found in water, soil and wastes (e.g. [292] [293]) and for some of them possess biotechnological potentials (e.g. R. eutropha, [294]). Rhodoferax ferrireducens was isolated from subsurface sediments [295] and appears to have

the genetic material to use or resist to various stresses comprising heavy metal or aromatic compound presence, and oxidative stress [296]. Xanthomonas are gram negative rod-shaped bacteria known as common plant pathogens and to grow almost exclusively in plants. However, some of them have the potential to create biofilms [297] and so can survive in the environment inside these structures.

Some species are unusually highly detected in these two metagenomes but inversely correlated to the presence of A. phosphatis. So it is probable that these sequences correspond to other genomes, and can provide new insights about microbial communities present in this environment. These species are Chromobacterium violaceum, Nitrococcus mobilis and Nitrosococcus oceani (Figure 33). In addition, sequences related to inorganic sulfur assimilation are more represented in these two metagenomes and not correlated to A. phosphaztis. So this particular process in probably presents in other species than A. phosphatis and can play an important role for the production of cysteines and methionines for example.

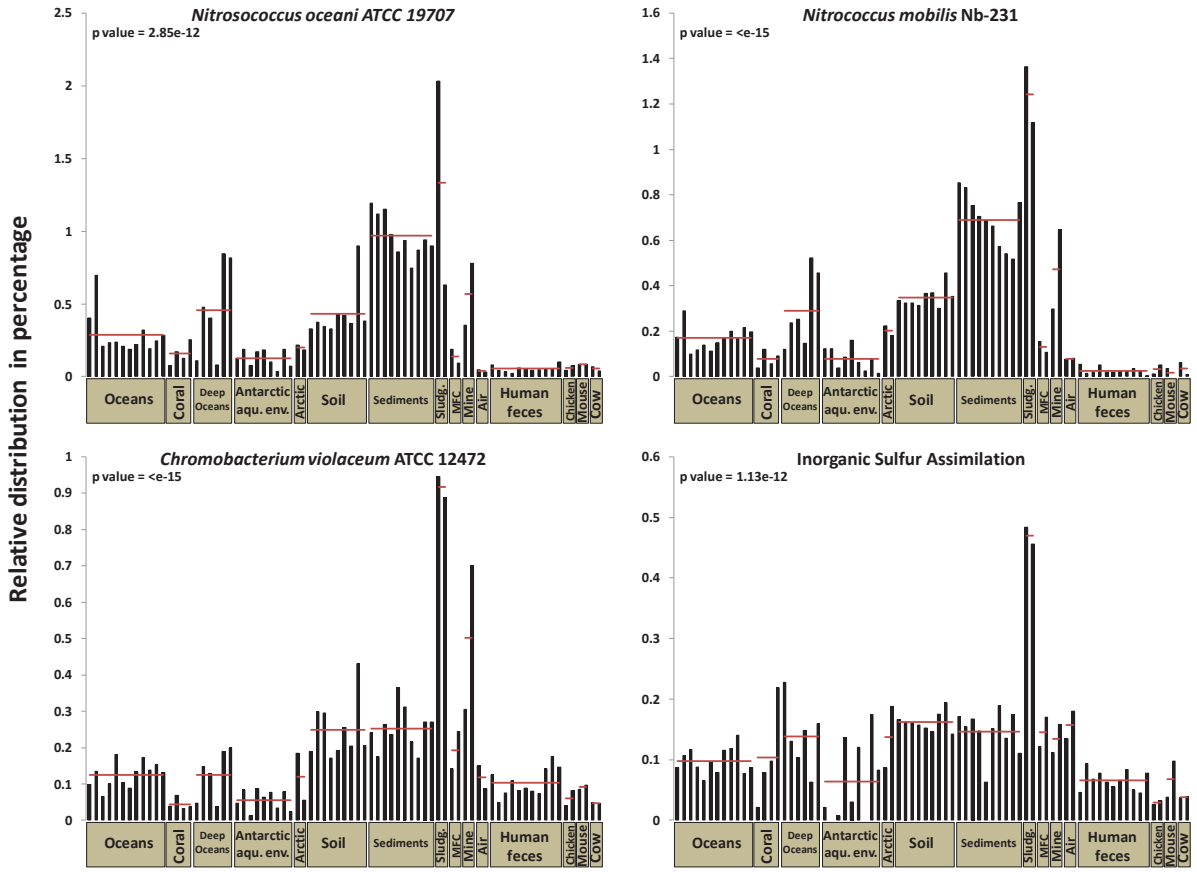


Figure 33. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and a functional subsystem (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Chromobacterium violaceum is a gram-negative bacterium possessing a versatile energy-generating metabolism providing a particular potential to survive under diverse environmental conditions [298]. This bacterium is known to be primarily found in water and soil [299]. Interestingly, this species possesses important industrial, pharmacological and ecological interests [300] but is also a rare pathogen that can cause potentially fatal infections in both humans (e.g. [301]) and animals (e.g. [302]). Nitrococcus mobilis was isolated in 1971 by Watson and Waterbury [303]. It is a nitrifier (nitrite oxidation to nitrate) gammaproteobacterium known to play a role in the nitrogen cycle in marine environments. Nitrosococcus oceani was isolated from seawater [304] and in known to produce energy from the oxidation of ammonia to nitrite. Its genome was sequenced by Klotz and collegues in 2006 [305].

Microbial communities present in these two EBPR are similar each other in spite of their distinct localization (USA and Australia) and possess distinct characteristics in comparison to other communities. In particular, they appear to possess an important genetic potential to assimilate inorganic sulfur, the oxidize H2 or to degrade phenylpropionate compounds. However, their genetic potential in term of phosphate metabolism is not particularly elevate and could be improved with the creation of artificial genomes [306] optimized for EBPR processes.

Wastewater microbial fuel cell anode biofilms:

Based on electrical effects accompanying the decomposition of organic compounds, the concept of using microbial cells in an attempt to produce electricity was proposed at the beginning of the twentieth century [307]. This bioprocess is now largely studied under the form of microbial fuel cells (MFC). MFC are bio-electrochemical systems that convert chemical energy into electricity through the catalytic activities of microorganisms and provide an important potential as alternative energy sources, novel wastewater treatment processes, and biosensors for oxygen and pollutants [308]. Two wastewater microbial fuel cell anode biofilms (anaerobic chamber) were sequenced on the aim of studying species and functions involved in the generation of electricity to improve the efficiency of this promising process (Jean-Michel Monier, MFC exploitation program).

This sequencing effort provided unique information about microbial communities present in these chambers (Tables 1 and 2, and figure 11). Sequences related to bacterial chemotaxis, bacterial motility (gliding), formate hydrogenase, acetyl CoA pathway of CO_2 (Figure 34), bacterial polyadenylation, selenocysteine metabolism, copper homeostasis, and tetrathionate respiration (Figure 35) are more represented in this environment in comparison to the other.

Bacterial chemotaxis allows swimming cells to follow chemical gradients in the environment [309]. However, the distribution of sequences related to this function is more important in

these MFC than in other environment. In particular, this function is lowly represented in oceans and deep oceans. Because the anode allows energy to bacteria we hypothesize that, among all the species present in the wastewater, those colonizing the biofilms found first the anode by using chemotaxis. This function is a considerable advantage to colonize first areas of interest. Gliding was defined as a movement of a non flagellated cell in the direction of its long axis on a surface [310]. It allows microorganisms to travel in low water content environments like soil but also biofilms [311]. The distribution of sequences related to gliding bacterial motility is important in the MFC, but not in acid mine drainage biofilms. This function could play a particular role in the production of electricity by microorganisms at the anode.

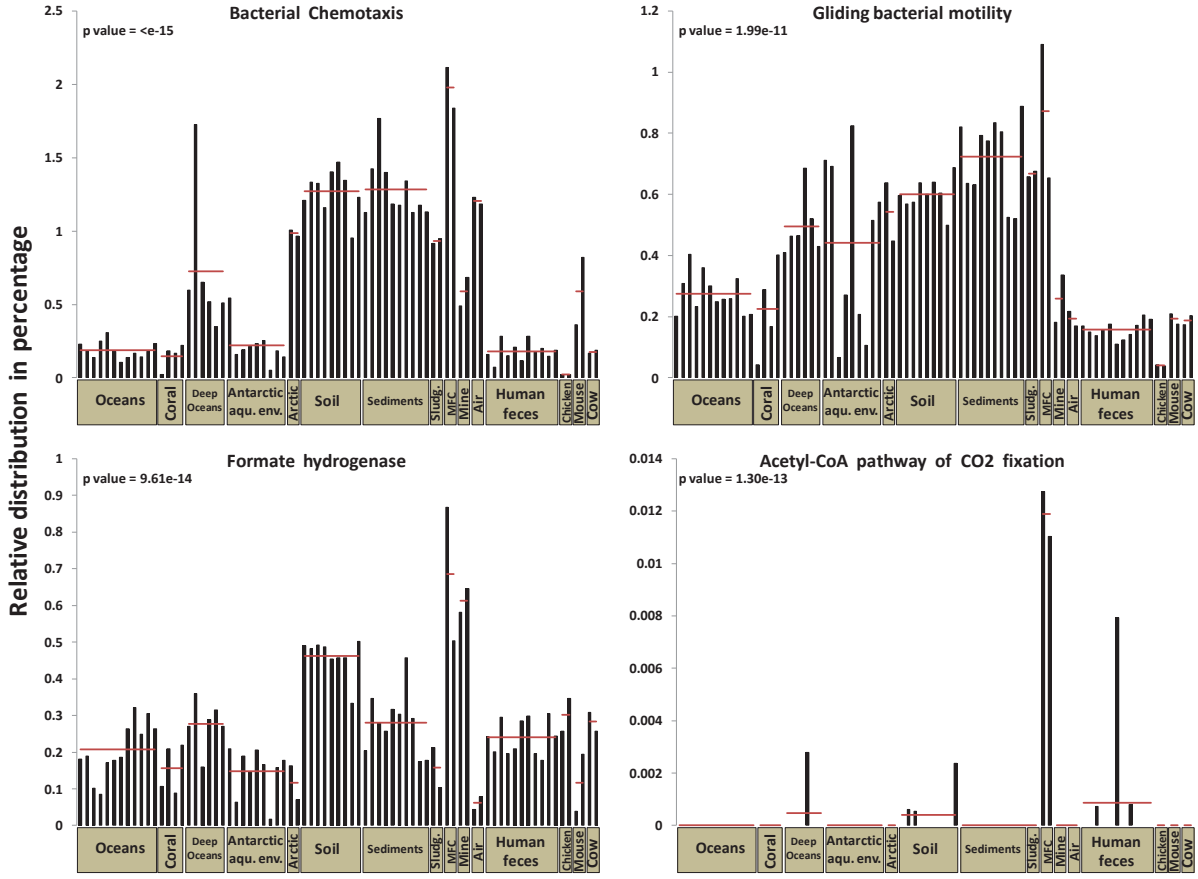


Figure 34. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Formate hydrogenases are membrane-bound complexes highly studied in Escherichia coli and possessing the particularity to oxidize formate to hydrogen and carbon dioxide [312]. Because the production of electricity is limited by the production of hydrogen in the anode compartment, this function is crucial for MFC energy generation. Interestingly, sequences related to this subsystem are more represented in these two metagenomes than in all the

other environments compared (except acid mine drainage biofilms which possess a similar distribution).

The acetyl CoA pathway of CO_2 is the major mechanism of CO_2 fixation under anaerobic conditions by transforming this molecule into acetyl CoA [313]. In spite of a low representation, this function is more represented in these two metagenomes than in the 75 other, where it is generally undetected. This observation can easily be explained by the fact that a majority of the 15 environments are aerobic. Bacterial RNA polyadenylation appears to plays a role in RNA stability by providing a signal for RNA degradation [314]. However, this process is still lowly understood, and the fact that sequences related to this mechanism are unusually represented in the two anode biofilms can provide new hypothesizes for both anode biofilms and bacterial RNA polyadenylation studies.

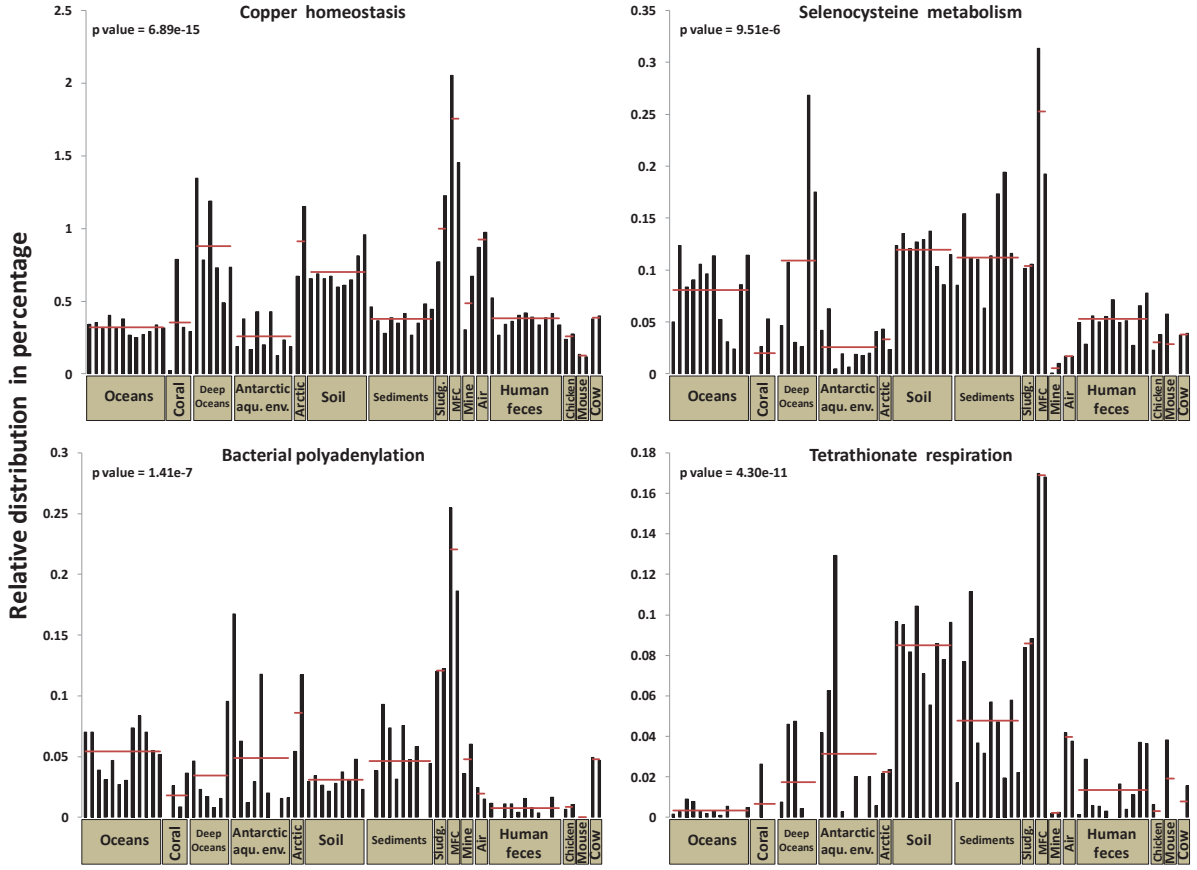


Figure 35. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Selenocysteine, the 21st amino acid [315], is identical to cysteine except that selenium substitutes the sulfur atom. The position of selenium in the periodic table makes selenoproteins ideal catalysts for many biological redox transformations [316]. Due to a lower pKA, the selenium atom confers a higher reactivity to selenocysteine than cysteine.

Thus most selenoproteins use their higher nucleophilic activity to catalyze redox reactions. Sequences related to the selenocysteine subsystem are more represented in this specific environment (anode biofilm) than in the other (except in two deep ocean metagenomes which possess an equivalent distribution of this metabolism). Thus selenoproteins probably stimulate MFC biological redox reactions. In addition, selenoproteins are known to represent one of the main enzymatic antioxidant systems [317], and so could play an important role in microbial communities present in the anode biofilms in response to oxidative stresses.

Under anaerobic conditions, bacteria can use tetrathionate as a terminal respiratory electron acceptor [318] during carbohydrate metabolism. The reaction is: S4062- + H2 = 2S2O32- + 2H+. This reaction needs the presence of a tetrationate reductase, known to be present in some heterotrophic bacteria (Salmonella, Proteus, Citrobacter) and a marine pseudomonad [319-321]. The distribution of sequences related to this respiration is higher in the two anode biofilm metagenomes than in the 75 other. Because this reaction generates hydrogen electrons, it is involved in the generation of electricity at the anode biofilms. In addition, no correlations were found between the distribution of this process and the species known to possess the genetic necessary machinery, so emphasizing a lack of information on this domain. Some uncultured bacteria are probably more involved in this anaerobic respiration in the nature.

Copper ions play an important role as a redox co-factor in cell metabolism, and in particular in photosynthesis and respiration. However, copper ions are responsible to the intracellular generation of reactive oxygen species [322]. In addition, this ion appears to be more toxic under anaerobic conditions [323] [324]. To respond to this stress, some bacteria used copper homeostatic mechanisms, which comprise the sensing, the chelation and finally the transport outer the membrane of the ions [325] [326].

In addition, few organisms appear to be prevalent in this environment. It is in particular the case of the genera Geobacter, Pseudomonas, Acidovorax, and Aeromonas, and the species Delftia acidovoran and Agrobacterium tumefaciens unusually highly detected in these two metagenomes (Figure 36). Aeromonas species are facultative anaerobic chemo-organotrophs found in aquatic environments included heavily polluted waters. They possess various virulence determinants [327] and are known to cause infections in both vertebrates and invertebrates [328]. This genus is highly studied due to its risk for human health and some genomes were sequenced (e.g. Aeromonas hydrophila; [329]). Interestingly, an Aeromonas hydrophila strain was isolated from an anode microbial fuel cell [330], emphasizing electrochemical activities for this species.

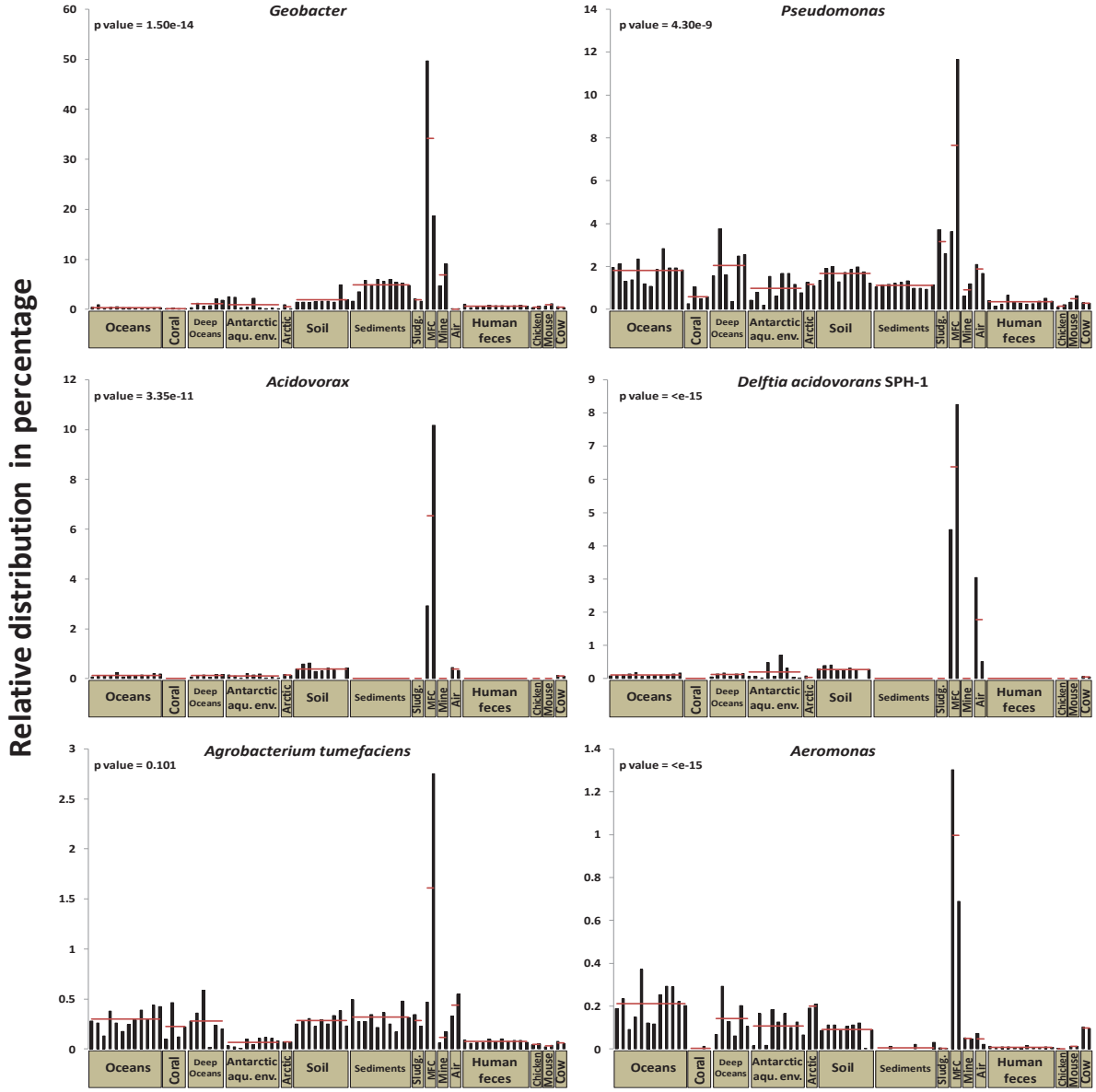


Figure 36. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Agrobacterium tumefaciens is plant pathogen species causing Crown Gall disease [331] by transferring and integrating a part of its DNA into the plant (see [332] as a review). This species was highly studied due to this particularity [333] providing a unique biological tool for the production of transgenic plants [334,335]. Interestingly, although this strain is considered as an aerobe, it possesses denitrification genes and in anaerobic conditions has the potential to use nitrate as an electron acceptor [336]. Delftia acidovorans is an aerobic gram negative bacillus known to be present in soils and water environments and to possess specific capacities to degrade pollutants and in particular pesticides [337,338]. Interestingly, this strain possesses a particular outer membrane, named Omp32 [339,340] of which

channel is strongly anion-selective [341]. This asymmetric conductance is clearly unusual in porins, and could play a role in the MFC process.

Geobacter are anaerobic species possessing the particularity to oxidize organic compounds to CO2 using iron oxides as electron acceptors [342]. In addition, Geobacter species possess important electron transfer capacities and are known to occur near anode surfaces [343-346] where they provide energy when using microbial fuel cells. The prevalence of sequences related to Geobacter species in these two metagenomes confirm they particular role in MFC electricity production.

Pseudomonas related species are rod shaped Gram-negative aerobic bacterium known to possess one or more polar flagella and to occur in water environments. Pseudomonas aeruginosa is one of most detected species related to this genus in these metagenomes. It is well known as a human pathogen, and appears to acquire quickly antibiotic resistance genes (e.g. [347] so increasing its disease potential. In addition, this strain can grow anaerobically into biofilms by using specific outer membrane proteins [348]. Finally, this strain and other species related to the genus Pseudomonas were isolated from the anode of microbial fuel cells [349] and appear to play a role in electrochemical activities by secreting a redox mediator, the pyocyanin.

Two species of the Acidovorax genus have been sequenced and are highly detected in these two metagenomes. Althrough these species have the same distribution among the 77 metagenomes, they are known to possess two distinct capacities. Acidovorax avenae citrulli AAC00-1 is a plant pathogen (e.g. [350]) and Acidovorax sp. JS42 is capable of degrading toxic nitroaromatic compounds [351]. Earlier, a third Acidovorax strain was sequenced. This species, named Acidovorax ebreus is the first sequenced anaerobic nitrate-dependent Fe(II) oxidizer [352] but was not present in the MG RAST database during the annotation of the 77 metagenomes. So it is possible that this species also is highly present in the MFC metagenomes. Interestingly, Acidovorax 16S rRNA genes were detected in the cathodic microbial community of a MFC [353] but this genus is not known to occur in anode biofilms.

It is important to note that sequences related to Rhodoferax ferrireducens, known to possess electro-catalytic properties (see [354] as a review) are not more present in these MFC than in activated sludges (see figure 29). So only some organisms are selected in this anode biofilms to generate electricity. More temporal studies need to be performed to understand the colonization of this environment and mechanisms occurring to improve the production of electricity using microorganisms.

In spite of an unusual distribution of sequences related to these different functions and species, adaptation processes and evolution are limited and can take considerable times. An alternative could be to create synthetic microorganisms [306] possessing considerable number of genes involved in the generation of electricity to optimize energetic production capacities and wastewater treatment. The functions described here provide unique

information about specific mechanisms involved in anode biofilms to generate electricity and should be more represented in future MFC synthetic bacteria.

Acid mine drainage biofilms:

Extreme environments generally possess a limited biodiversity. Thus one of the major advantages of studying these environments using the actual metagenomic tools is that the assembly of predominant genomes is possible [37]. One of the best examples is probably the sequencing and partially assembly of two acid mine drainage biofilm metagenomes (extreme environments, pH nears 0, [37]. Due to a very specific phylogeny (see figures 3, 4 and table 1), it is particularly interesting to compare these genetically simple metagenomes to other at the functional level. But unexpectedly, in these biofilms only sequences related to trehalose biosynthesis, mercuric reductase, beta-lactamases, and formate hydrogenase (also present in the same distribution in MFC, see figure 34) appear to be clearly more represented in comparison to the other environments (Figure 37).

Trehalose is a non reducing disaccharide present in a wide variety of organisms where it may serve as a source of energy and carbon [355]. However, this sugar can also be involved in stress responses. Trehalose can protect proteins and cellular membranes from various stresses like desiccation, dehydratation, heat, cold and oxidation [356]. Because the pH in this environment is near 0, we propose that this sugar is also possibly involved in low pH protection. The considerable presence of sequences related to transposons and mercuric reductase emphasize also a particular adaptation of this biofilms to an extreme environment. Formate hydrogenase related sequences help the understanding of energy production by oxidinzing formate to hydrogen and carbon dioxide [312].

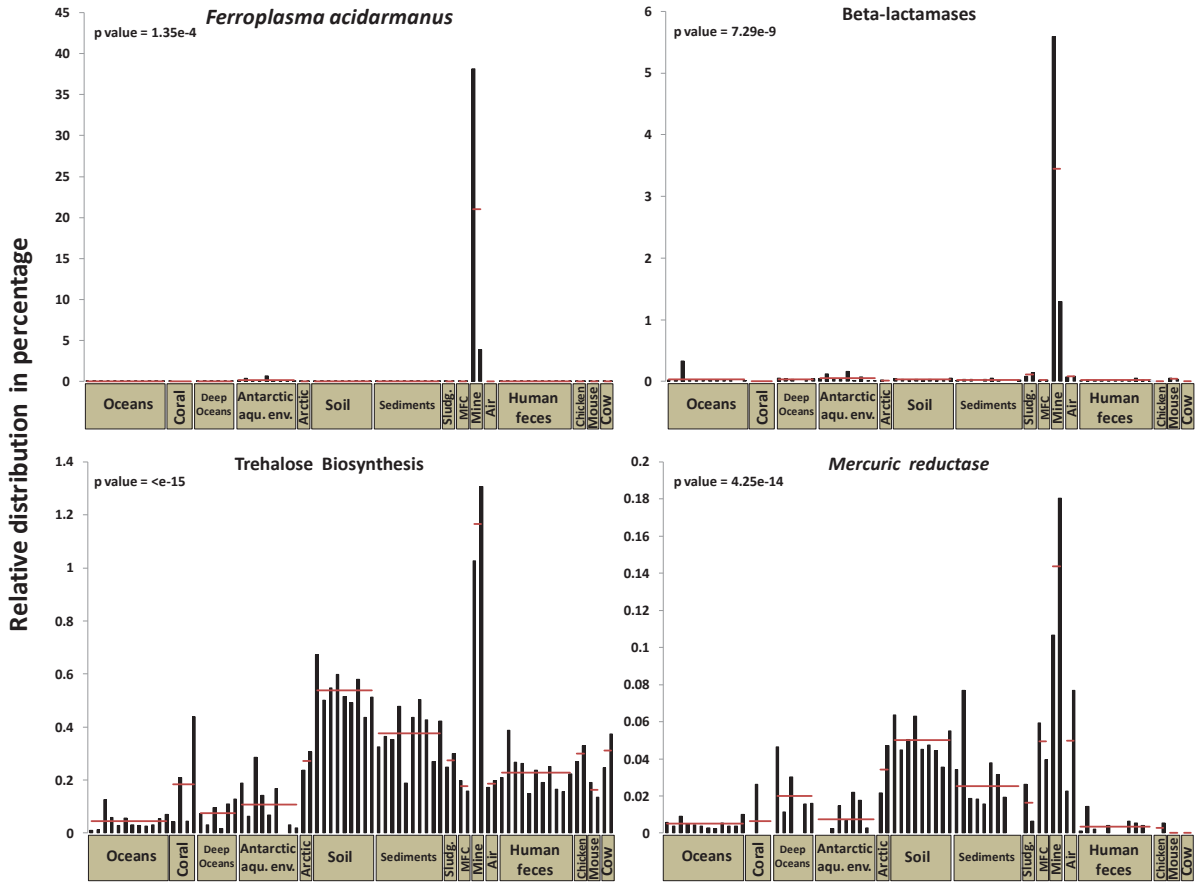


Figure 37. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

In addition, it is important to note that sequences related to orphan regulatory proteins (involved in quorum sensing process) and osmoregulatory periplasmic glucans (role in extreme environments) are unusually lowly represented in these biofilms.

Ferroplasma acidarmanus was isolated from an acid mine drainage biofilm [357] where it is capable of growing at pH 0 and constitutes the major part of the microbial community. This strain was highly studied as an extremophile model and mechanisms involved in Arsenic and copper resistance [358,359], electron transport for iron oxidation [360] and biofilm development [361] were investigated.

Other taxa appear to be more represented in this environment (figure 38). The strain Magnetococcus MC-1 was isolated from an Estuary [362] and due to its magnetotactic capacities produce unique crystal morphology: an elongated pseudo hexahedral prism of magnetite. This process appears to be highly genetically controlled. This species is not known to occur in acid mine drainage biofilms. Of course, sequences matching with this genome

can belong to another species, but this observation can generate new hypotheses about acid mine drainage biofilm characteristics and adaptations.

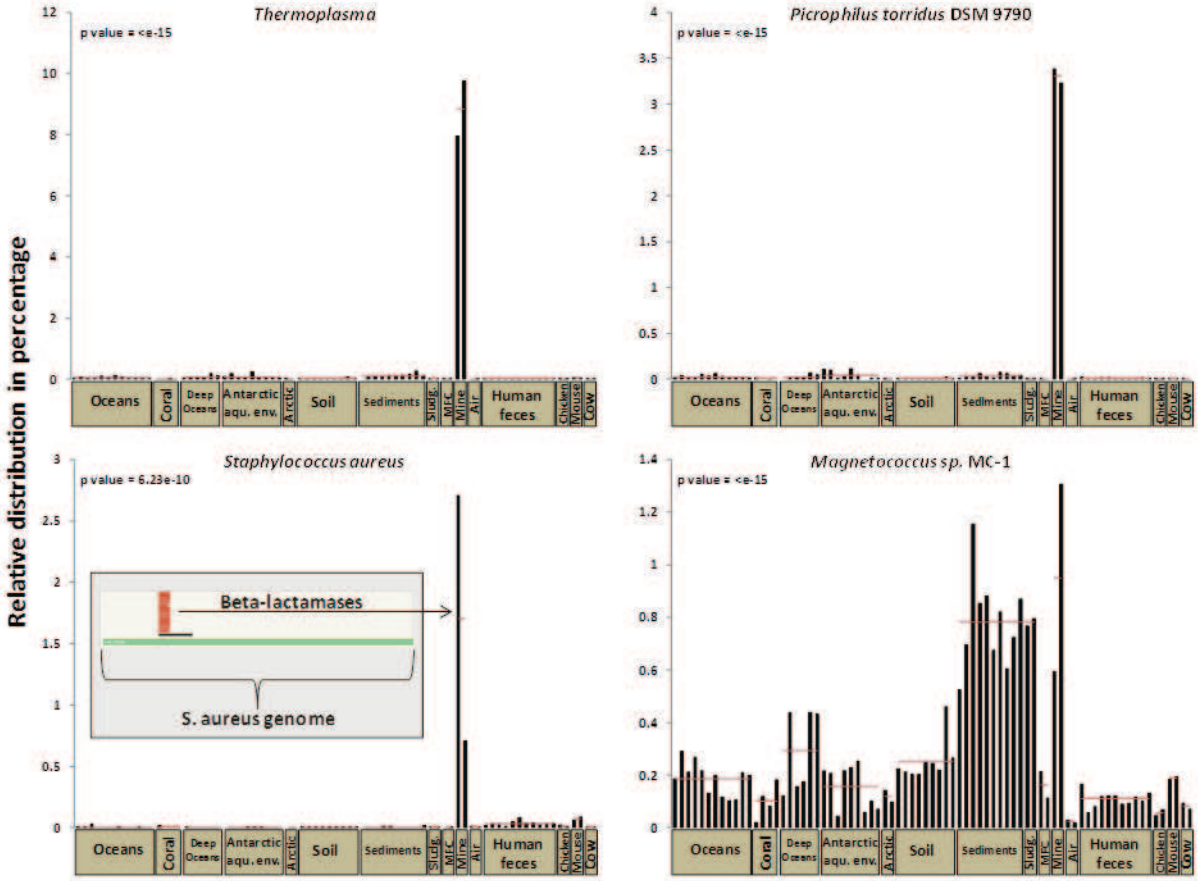


Figure 38. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Picrophilus torridus was isolated from a dry solfataric environment [363] and sequenced to study its genetic adaptation to an extreme environment [364]. This strain grows optimally at pH 0.7 and is adapted to high sulfuric acid concentrations (until 1.2 M) and high temperatures. In addition, in contRAST to the other known thermoacidophilic microorganisms, P. torridus possesses an unusually low intracellular pH (pH 4.6, [365]). Thermoplasma acidophilum and volcanium are two thermoacidophilic species that grow near 60°C and pH 2. Both of them were sequenced [366,367] and provide unique insights into thermophilic archaea metabolic machinery and evolution [368]. They are detected at the same level in these two metagenomes.

Staphylococcus aureus is a Gram-positive coccus facultatively anaerobic species well known and studied due to its common cause in staphylococcocal infection. Approximately 30% of the population is probably colonized with Staphylococcus aureus either chronically or

intermittently [369] and in some cases this colonization become an infection. The causes of transition between the two states are largely unknown (see [370] as a review). All sequences related to this species correspond to the beta-lactamase subsystem. So only a specific part of its genome is detected in these metagenomes and probably belong to the genome or mobilome of other species.

Polluted air:

To study bacterial and archaeal communities present in the atmosphere and their potentials to impact on human health, microorganisms from two densely populated urban building indoor air were sequenced [52]. One of the principal conclusions was that the major stresses occurring in air microbiota may be iron limitation, oxidative damage and desiccation. Using the 75 other metagenomes for distribution comparisons, sequences related to ton and tol transport system, toxin-antitoxin systems, arsenic resistance (see figure 14), fatty acid degradation regulon, phenylpropanoids general biosynthesis, and widespread colonization island appear to be more represented in the two polluted indoor air metagenomes (Figures 39 and 40) and can provide substantial information about this unique life style.

The gyrase enzyme modules bacterial chromosomal topology by introducing negative DNA supercoils [371]. This enzyme is composed of two subunits named GyrA and GyrB [372] and is involved in primordial bacterial processes: DNA repear, replication and transcription, but also recombination and decatenation [373-375]. Due to its role in bacterial life, Gyrase is the target of quinolones and other antibacterial agents [376]. By interacting with the GyrA subunit, inhibitors alter the enzyme activity which will create double stranded DNA lesions (e.g. [377]), so inducing oxidative damages [378], cell division arrest and bacteriostasis state [379,380] due to the inhibition of both replication and transcription mechanisms [381,382]. Finally, gyrase inhibitors can induce cell death (e.g. [383]).

In addition to antibacterial agents, some bacterial toxins also target gyrase. In particular, the CcdB toxin highly studied from the E.coli plasmid F ([384,385]) binds the open formation of GyrA [386]. But another protein encoded from the same operon and named CcdA prevents the complex formation of CccB and GyrA. This protein is also capable of reversing the CcdB toxin effect on gyrase enzyme. This process between CcdA and CcdB is called a toxin-antitoxin system (see [387,388] as reviews) and generally ensures the persistence of plasmids encoding them during replication by decreasing the proportion of cells that survive after losing the plasmid. Interestingly, in addition to the high distribution of this subsystem in these two metagenomes, more than 99% of the sequences matching correspond to the ccdB toxic gene, and the ccdA gene is not detected. So it is possible that a majority of bacterial cells in air are in a bacteriostasis state, with a low rate of replication and transcription due to a complex between CcdB and GyrA, waiting to colonize an area of interest. The different functions of toxin-antitoxin systems are still debatable [389] and the

unusual distribution observation of genes related to CcdB in these two metagenomes could help understanding its role in this particular environment. However, the mechanisms that liberate GyrA are not understood based on these metagenomic comparisons and additional studies have to be performed.

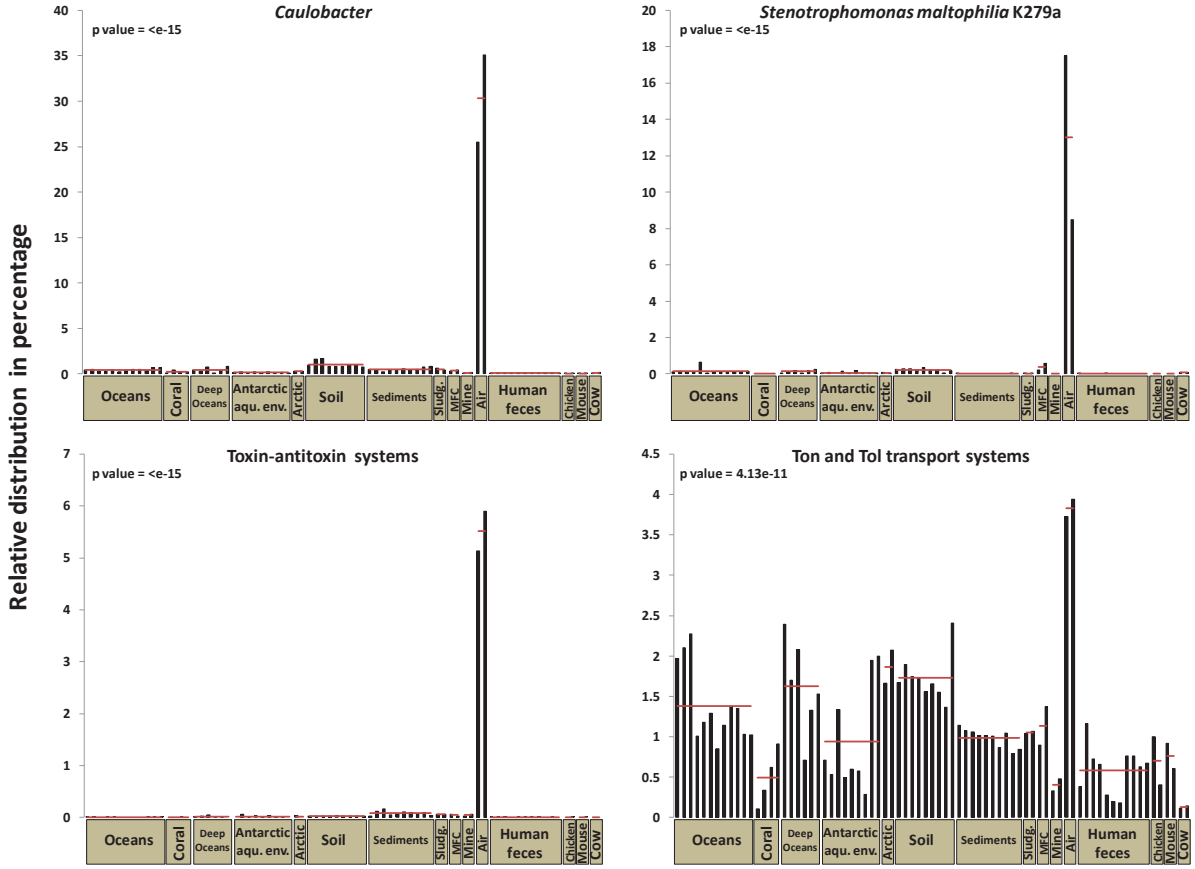


Figure 39. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Iron is an essential element for most organisms [390] but is often a limiting factor of life (e.g. in oceans, [391]) due to its insolubility in aerobic environments when the pH is near 7. In response to this stress, some bacteria, including a majority of pathogens, possess high-affinity transport systems and generate high-affinity siderophores that complex extracellular iron [392] to optimize the acquisition of this element. Because ferric siderophores are too large to across classical porrins into the outer membrane, an energy dependant cell envelope protein, named TonB, is used (see [393] as a review). In these two metagenomes, the majority of sequences matching with the ton and tol transport system correspond to tonB-dependant receptors. The unusually distribution of sequences related to genes coding these receptors emphasizes a possible adaptation of air microbial populations to an iron limited environment. However, it is important to note that these receptors are also known

to play a role in transport of some bacteriophages [394], group B colicin [395] and vitamin B12 [396]. But whatever the predominant role of these receptors, its architecture suggests a membrane surveillance and detection of peptidoglycan-associated outer membrane proteins [397] and it is apparent that the prevalence of these receptors are the consequence of an important lack of one or several crucial elements for microbial life.

Sequences related to Caulobacter species (Caulobacter crescentus CB15 and Caulobacter sp. K31 which are equally represented) and Stenotrophomonas maltophilia represent a considerable part of these metagenomes. While genomes present in these datasets are not already sequenced they are closes to these phylotypes and so could represent additional species among the genera Caulobacter and Stenotrophomonas.

Caulobacter crescentus possess particular mechanisms to survive in low nutrient environments and is highly studied due to an asymmetric cellular division (e.g., [398]). Caulobacter crescentus cell division results in the formation of two distinct cell types: a swarmer cell possessing a pili and a stalked cell. While the swarmer cell aims to discover a favorable habitat (e.g., iron concentration), the stalk possess a considerable strength potential for surface attachment [399]. Stalks contain outer membrane and periplasmic proteins [400,401]. This stalk possesses several TonB-dependent receptors and indicates competence for nutrient uptake [400]. In addition, TonB-dependent receptors are activated during carbon starvation in Caulobacter crescentus [402]. While the majority of sequences matching with the Caulobacter crescentus genome correspond in reality to other genomes, some of them are highly similar to Caulobacter crescentus genes (percentage of identity superior to 90%). Interestingly, it is the case of genes coding for the Ftsh protease which are known to be involved in the development and stress response control in this species and to control the timing of cell differentiation [403]. Genes corresponding to TolA and highly similar to those present in the genome of Caulobacter crescentus (e-value<10^{-60} and percentage of identity superior to 90%) are also detected. TolA is part of the Tol-Pal Complex, an essential component for outer membrane integrity and the positioning of a polar localization factor in Caulobacter crescentus [404]. So it is possible that the predominant species living in this indoor air possess an asymmetric cellular division to survive in this particular environment, probably poor in crucial nutrients like iron [52].

S. maltophilia is the unique member of the genus Stenotrophomona, and its genome was sequenced [405] due to an important role in nosocomial diseases [406]. S. maltophilia is found in a wide variety of environments and geographical regions and is frequently isolated from humans with chronic respiratory disease (see [407] as a review). S. maltophilia was recently associated with chronic lower airway disease in the horse [408]. Based on its distribution in the two indoor air related metagenomes, this species could be highly transferable by air which could explain some reported cases of person-to-person transmissions and its responsibility in nosocomial diseases.

In addition to Caulobacter and Stenotrophomona, other species are unusually detected in these two metagenomes. It is the case of Methylobacterium extorquens, Sphingomonas wittichii, Streptococcus equi (subsp. Zooepidemicus) and Streptococcus pneumoniae (Figure 40). Methylobacterium extorquens is a facultative methylotroph ubiquitous in soils, water and air environments [409]. Its distribution among the 77 metagenomes confirms this information. Methylobacterium species have the particularity to grow on reduced one carbon compounds other than methane (e.g., the methanol emitted by the vegetation, [410]. This metabolism could represent an important way to generate energy in this environment. Sphingomonas wittichii is a dioxin-mineralizing bacterium that metabolize chlorinated aromatic hydrocarbons [411] [412] [413]. Its distribution in these metagenomes could be due to a pollution by aromatic compounds in the indoor air studied. Streptococcus equi are important pathogens in equine respiratory disease. In particular, S. zooepidemicus is a normal commensal organism and an opportunistic pathogen in the equids [414]. The persistence of this species is known to be limited in outdoor environments and negatively affected by the sunlight [415]. However, this species is more detected in air than anywhere else in the compared environments, and could represent a survival capacity in the atmosphere. Streptococcus pneumoniae (strain OXC141), a significant human pathogenic microorganism, is also unusually detected in these metagenomes. The fact that urban particulate matters increase adhesion of Streptococcus pneumoniae to human airway epithelial cells [416] can in part explain these observations.

Finally, even if these functions are also highly detected in other metagenomes, sequences related to widespread colonization island and fatty acid degradation regulons are in average more represented in this indoor air than in the other environments and can provide additional information about the life style of related microbial populations. Obviously, these indoor-air metagenomes represent unique microbial communities that are probably highly adapted to particular environmental characteristics. However, additional sequencing efforts have to be done to assemble the predominant species present (related to the Caulobacter genus). The study of complete genomes should help elucidating the unexpected distribution observation of sequences related to CcdB and their role for the survival of atmospheric cells.

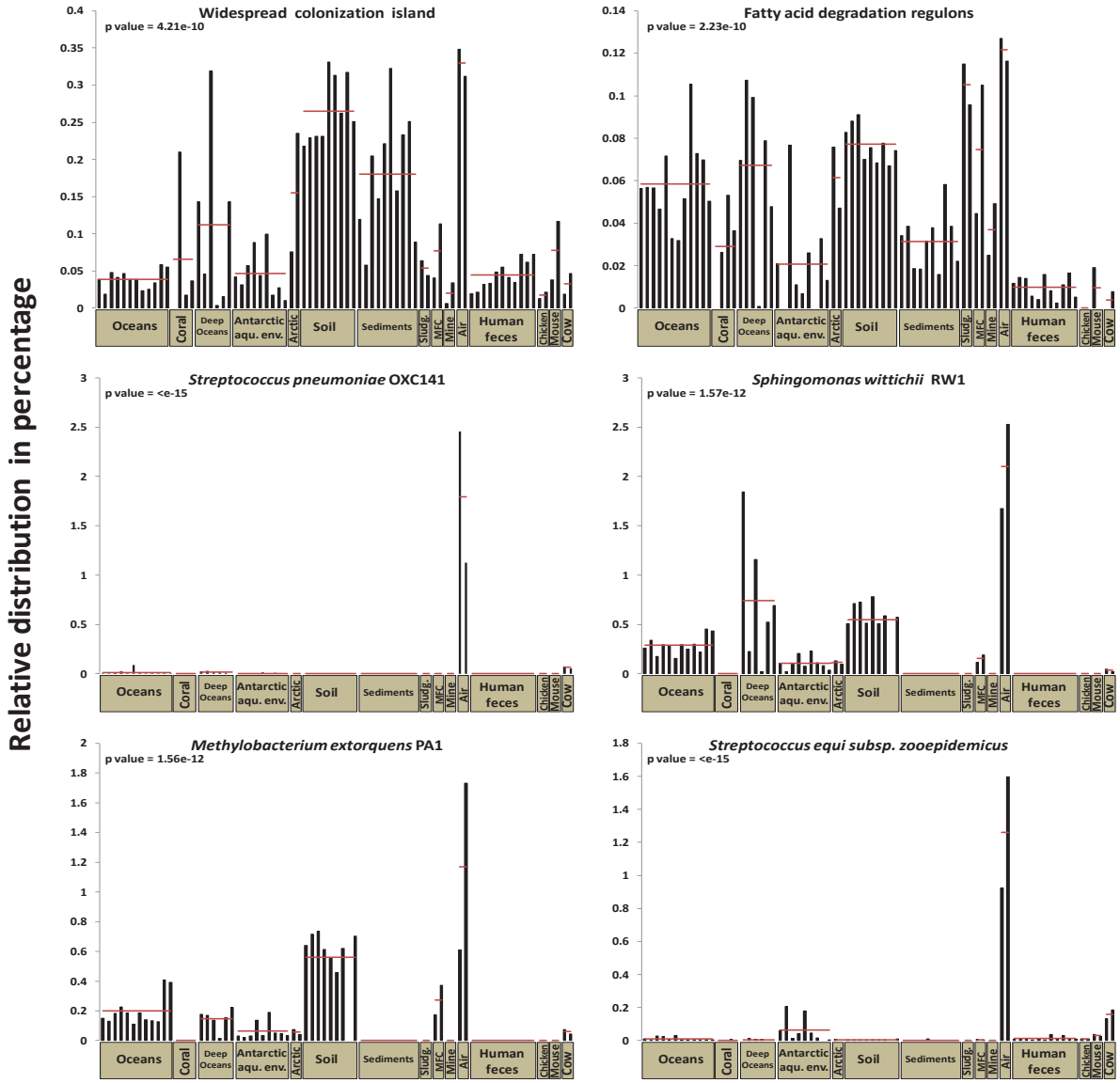


Figure 40. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Animals:

Animals emerged and evolved in an environment already entirely colonized by Bacteria and Archaea. As a consequence, they are probably more adapted to these microorganisms than the inverse. However, due to an important co-evolution, microorganisms living into animals (the microbiotas) appear to represent unique communities on Earth (see figure 11 and table 1) and are highly adapted to animal environments. Moreover, these microorganisms represent a considerable genetic part of animals (the "second genome"), and as an example,

there are ten times more of microorganisms than human cells in our body [417]. Most of these cells live in the intestinal tract, where their density can attain 1012 organisms per gram of intestinal contents [3], and correspond to a majority of Firmicutes and Bacteroidetes [418]. These microorganisms generate simplified vitamins, carbohydrates and amino acids and have the ability to extract energy from animal diet and to transform it into fat [419]. As an example, carbohydrate fermentation by intestinal bacteria produces volatile fatty acids that are used as a considerable source of energy by the host [420]. Consequently, microorganisms appear to limit food consumption required to live but in the same time to increase fat body as was shown with mice [421].

While these mechanisms represent a substantial energy complement for animals and were probably a considerable benefit for human survival during its evolution, they are now actively studied to understand their role in term of obesity in countries where the lack of food is not a major problem, but where diseases associated to obesity are. Based on several studies, Bacteroidetes and Firmicutes appear to be linked to obesity due to carbohydrate metabolism differences [4,41,422,423]. Firmicutes distribution is more important in obese individuals and this ratio modification could be due to particular energy extraction capacities [4]. In addition, microorganisms are highly studied to understand their response during antibiotic treatments and their resilience capacities (see [424] as a review). These molecules can modify microorganisms for long periods, and so represent a risk for human health. Finally, efforts are made to comprehend interactions between gut flora and related host, and as an example, discoveries were done to understand how a specific bacteria (Bacteroides fraglis) can positively stimulate human immune system and prevent intestinal inflammation in animal models of colitis [425,426].

Metagenomes corresponding to four distinct animal microbial populations were compared to visualize their functional specificities based on general functional subsystems (Figure 41). Each animal appears to possess unique communities. Interestingly, the sampling localization appears to not be a crucial parameter in term of functional distributions. Chicken ceacum communities are closer to those from cow rumen than to mouse ceacum communities. Sequences related to nucleosides and nucleotides, RNA metabolism and protein metabolism are more detected in chicken ceacum and cow rumen. In contRAST, sequences related to regulation and cell signaling, cell division and cell cycle, potassium metabolism and virulence are more detected in human feces and mouse ceacum. In addition, sequences related to photosynthesis, motility and chemotaxis, phosphorus metabolism, cell wall and cell capsule, and prophage subsystems are more represented in mouse ceacum populations. Finally, sequences related to dormancy and sporulation, sulfure metabolism, secondary metabolism, stress response, miscellaneous, “cofactors vitamins prosthetic groups pigments”, fatty acids and lipids, carbohydrates, nitrogen metabolism, and membrane transport appear to be more represented in human feces microbial populations in comparison to the three other communities.

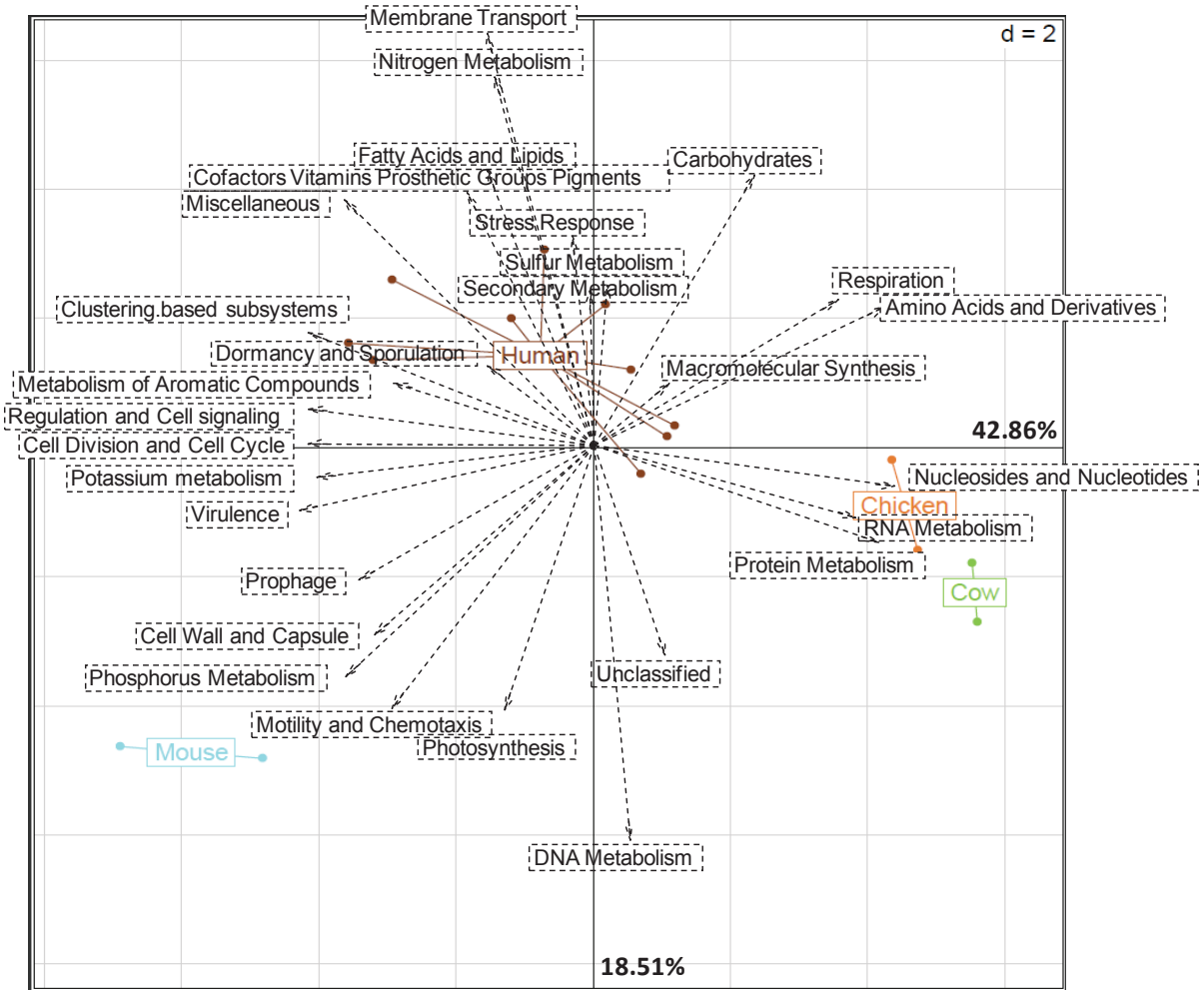


Figure 41. Principal component analysis based on the relative distribution (in percentage of annotated reads) of general functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for datasets corresponding to four distinct animal microbial communities. The percentages of the illustrated two major axes correspond to the fraction of the total variance that they represent.

Even if these observations can provide global information about these communities, more specific functional levels have to be used to compare them. In addition, even if considerable studies were done on animal (and especially human) microbiotas, to compare 15 environments of which four correspond to different animals provides interpretations about these communities' specificities, both in term of functions and taxonomy, which are for some of them difficult to detect when studying only microorganisms. Some genera appear to be clearly more represented in animals than anywhere else. It is especially the case of Bacteroides and Clostridum which represent a considerable part of these microbiotas, but also Desulfitobacterium, Lactobacillus, Enterococcus, Streptococcus, Fusobacterium, Listeria, Lactococcus, and Oenococcus (Figure 42 and 43). Most of these genera were already known to be prevalent in the gastrointestinal tract [427].

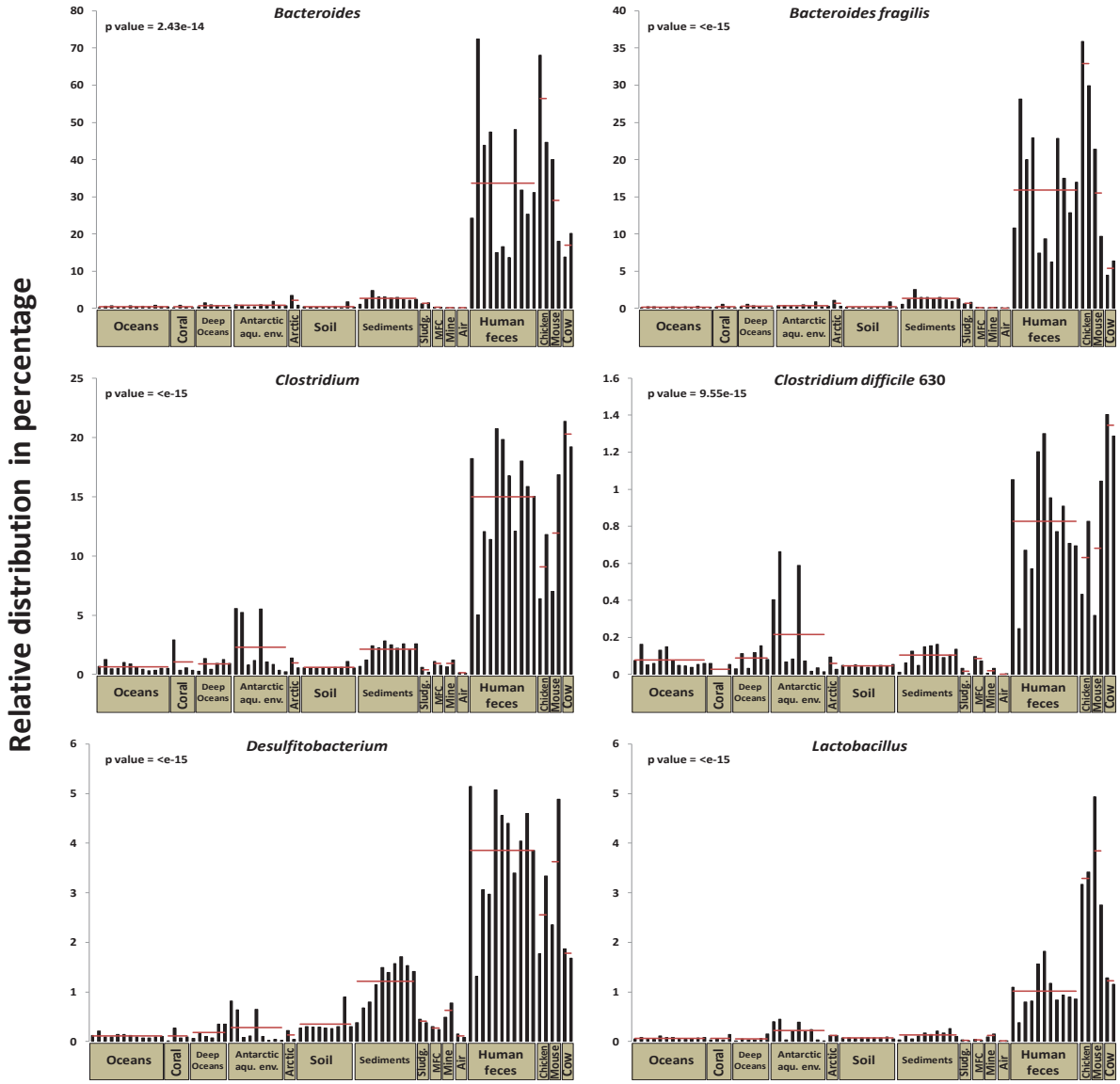


Figure 42. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Bacteroides are gram-negative, bile-resistant and non endospore-forming anaerobes rods (see ([428] as an overview of this genus and its role in human bodies). They are known to be the most predominant anaerobes in the human gut and to play a considerable role in term of immunity, digestion, but also protection against diseases (e.g., [429]). As expected, this genus is highly detected in animals and represents the most important group in these specific environments. However, some Bacteroides species are virulent and represent a risk for human health. In particular, Bacteroides fragilis is the most frequent isolate from clinical specimens and is considered as the most virulent Bacteroides species [428] even if paradoxically it is also studied due to its beneficial effects on human immune system (e.g.,

[430]). Based on its distribution using SEED annotation, this species represents 15 percent of human feces and mouse ceacum, and more than 30 percent of chicken ceacum.

Clostridium are obligate anaerobes Gram positive and spore-forming bacillus. They belong to the Firmicute phylum and are capable of producing endospores. They are known to be highly represented into human microbial populations (e.g., [427]) and to cause in rare cases opportunistic infections (see ([431] as a review). Species related to this genus have distributions varying considerably between the four animals (supplement data). Additional comparisons can be made by specialized microbiologists to descript more in detail these differences. As an example, Clostridium difficile has the capacity to develop when other bacteria are killed during antibiotic treatments. As a consequence, it is the most serious cause of antibiotic-associated diarrhea and can lead to pseudomembranous colitis, a severe infection of the colon [432]. Clostridium difficile represents up to 25% of nosocomial antibiotic associated diarrhea [433]. As a beneficial effect for human health, it is important to note that some specific species of Bacteroides appear to have the potential to prevent Clostridium difficile infections [434,435]. Both Clostridium genus and Clostridium difficile species distributions among the 77 metagenomes are represented in the figure 42. They are clearly more detected in animals than in the environment, and appear to be more represented in cow rumens (more than 20 percent for the genus) than in the other sequenced animal microbial populations.

Desulfitobacterium related species are strictly anaerobic bacteria that were first isolated from environments contaminated by halogenated organic compounds (see [436] as a review). Desulfitobacterium genus appears to be highly represented (even if lower than Bacteroides and Clostridium genera) in animal microbial populations (Figure 42). Interestingly, this genus is not particularly known to colonize animal bodies. So if this observation corresponds to reality, it represents a novelty in the field. However, annotation limits (due to a restricted number of already sequenced genomes) could overestimate the representation of this genus in these environments. Additional experiments have to be done to confirm these results.

Lactobacillus (a well studied probiotic genera), Streptococcus, Enterococcus, Lactococcus and Oenococcus appear to be more represented in animals than in the environment (Figures 43). These genera belong to the Firmicute Phyla and have the particularity to be members of the lactic acid bacteria (LAB). These members are naturally associated with mucosal surfaces (and especially the gastrointestinal tract) but are also present in several foods. They have in common the ability to ferment hexose sugars and lactic acid.

Fusobacteria are obligate anaerobes, non spore forming Gram negative bacilli related to the Bacteroidaceae family. Species related to this genus are known as human pathogens (especially in mouth). For example, Fusobacterium nucleatum has a causative role in infectious diseases of the oral cavity [437]. Fusobacterium necrophorum is a prevalent pathogen in peritonsillar abscess [438] and known as the most common pathogen in

Lemierre's Syndrome (e.g., [439]). Even if this genus is detected in all the compared environments, it is clearly more represented in animals.

Listeria genus contains pathogen species which for some of them are highly studied for their role in human health. In particular, Listeria monocytogenes, an aerobic and facultatively anaerobic gram-positive bacillus, is known as the causative agent of listeriosis (e.g., [440]), but other species have other roles (e.g., Listeria ivanovii is a pathogen of ruminants). This genus is highly detected in animals in comparison to the environment, but appears to be more common to human and mouse microbial populations.

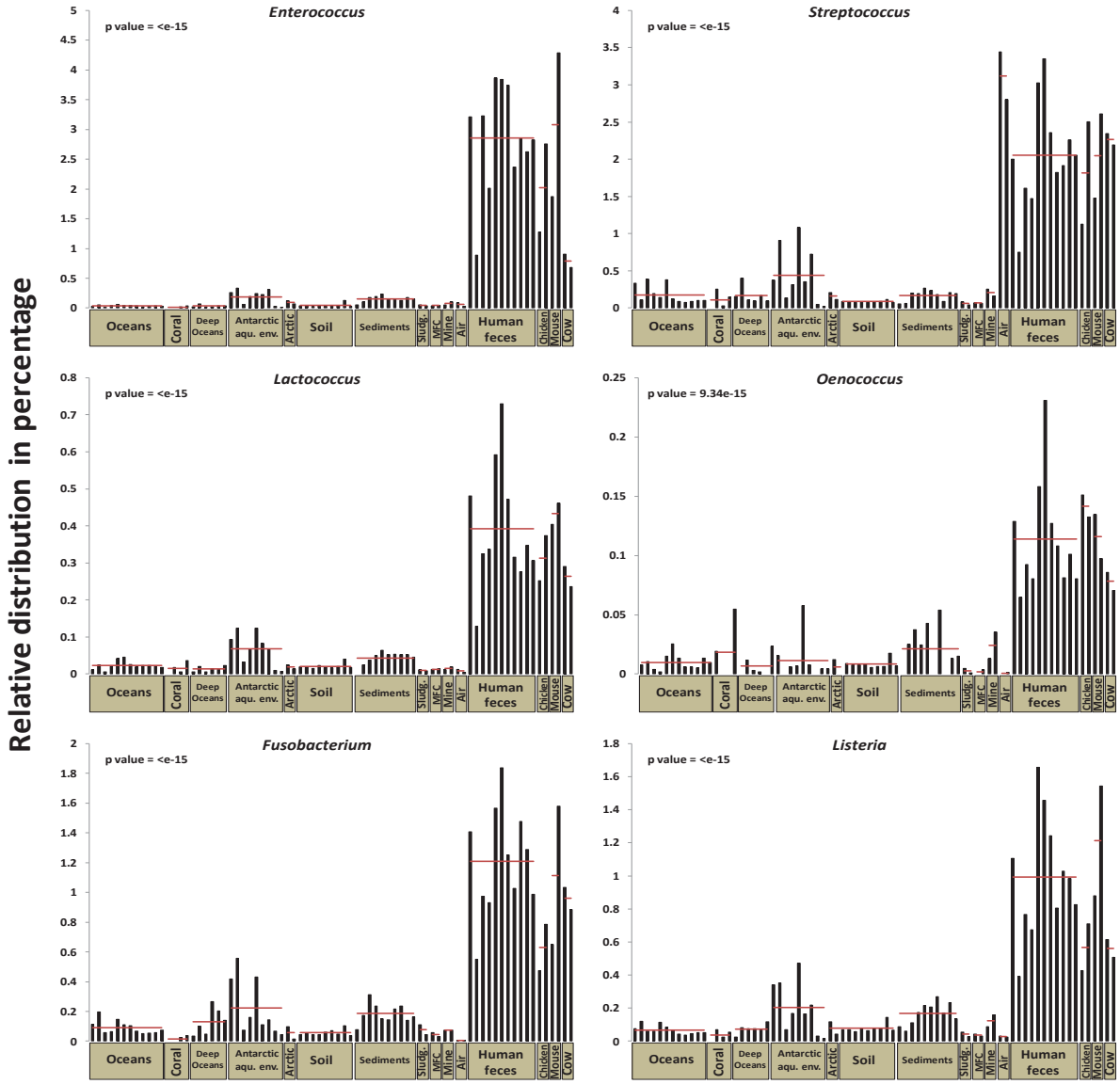


Figure 43. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

In addition to considerable taxonomical specificities in animals, some functions and genetical structures appear to be unusually highly represented in these populations. It is in particular the case of subsystems related to acetone, butanol and ethanol synthesis, Fe-S cluster assembly, L-arabinose ctilization, maltose and maltodextrin utilization, spore coat, and a conjugative transposon related to Bacteroidales (Figure 44). The anaerobic production of acetone, butanol and ethanol was a considerable biotechnological industry before the 1960's and the development of the petrochemical industry [441,442]. The synthesis of these compounds is known to be realized by solventogenic clostridia and Clostridium acetobutylicum in particular [443]. Interestingly, Clostridium acetobutylicum ATCC 824 is more represented in animals (supplement data) and is positively correlated to the subsystem corresponding to acetone, butanol and ethanol synthesis (R2=0.566), so corroborating by a metagenomic approach its role in this process.

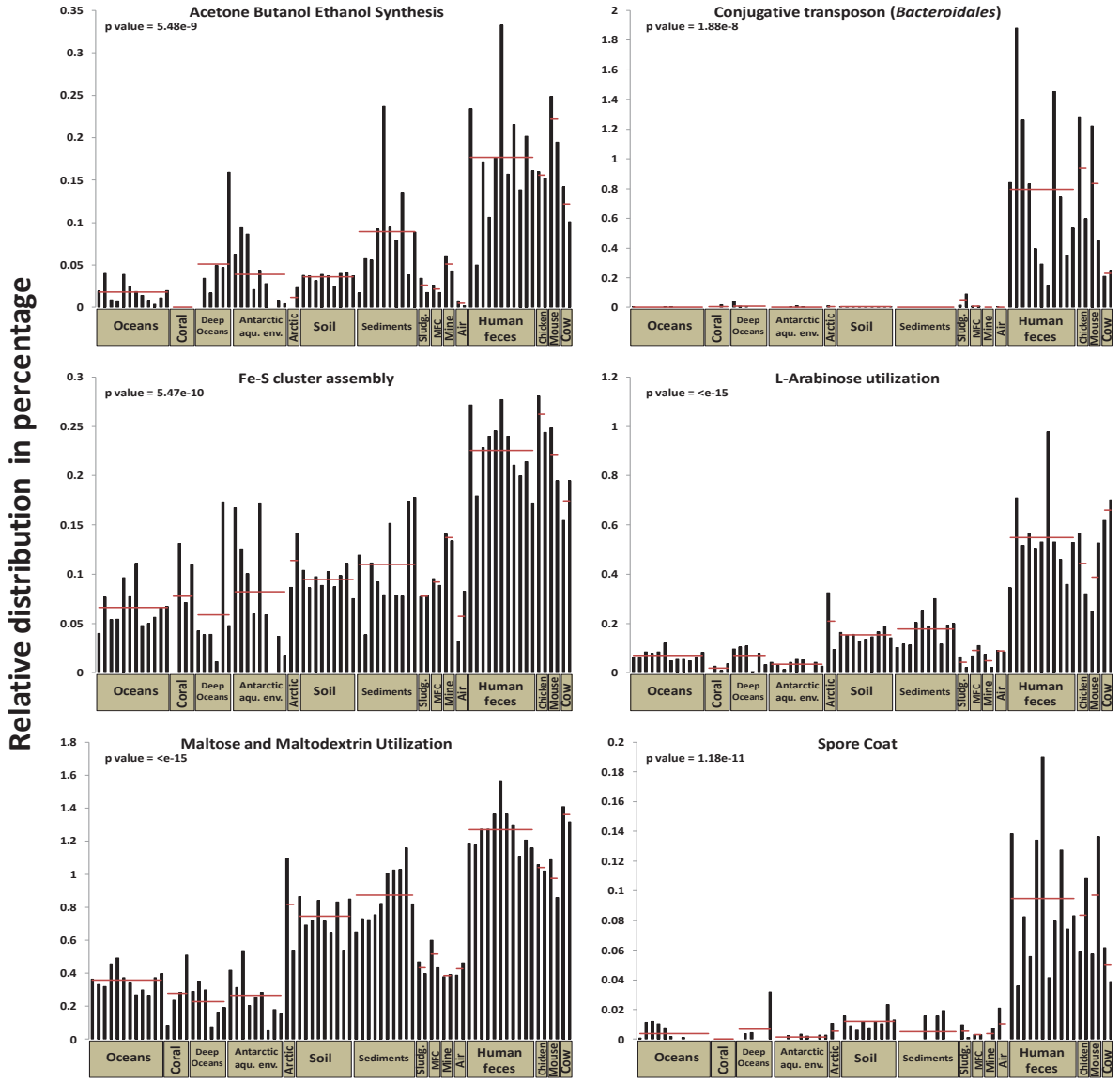


Figure 44. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic

datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Mobile genetic elements related to Bacteroides comprise plasmids, bacteriophages, transposons and conjugative transposons [444]. Conjugative transposons are detected at least one time in a large majority of Bacteroides strains [445]. These elements have a specific strategy of excision and integration [446] and are capable of insertion in chromosomes but also in plasmids. In this case, they can stimulate the transfer of the plasmid-conjugative transposons in other cells [447]. In addition, conjugative transposons can be cumulated in strains and in this case are suggested to stimulate transposition [446]. These elements are highly detected in animals even if important variations can be observed (especially in human feces, see figure 44). As expected, the same variations can be observed in the distribution of Bacteroides (Figure 42), and the correlation between the structure and the genus is important based in the two distributions among the 77 metagenomes (Figure 45), so confirming their close affiliation.

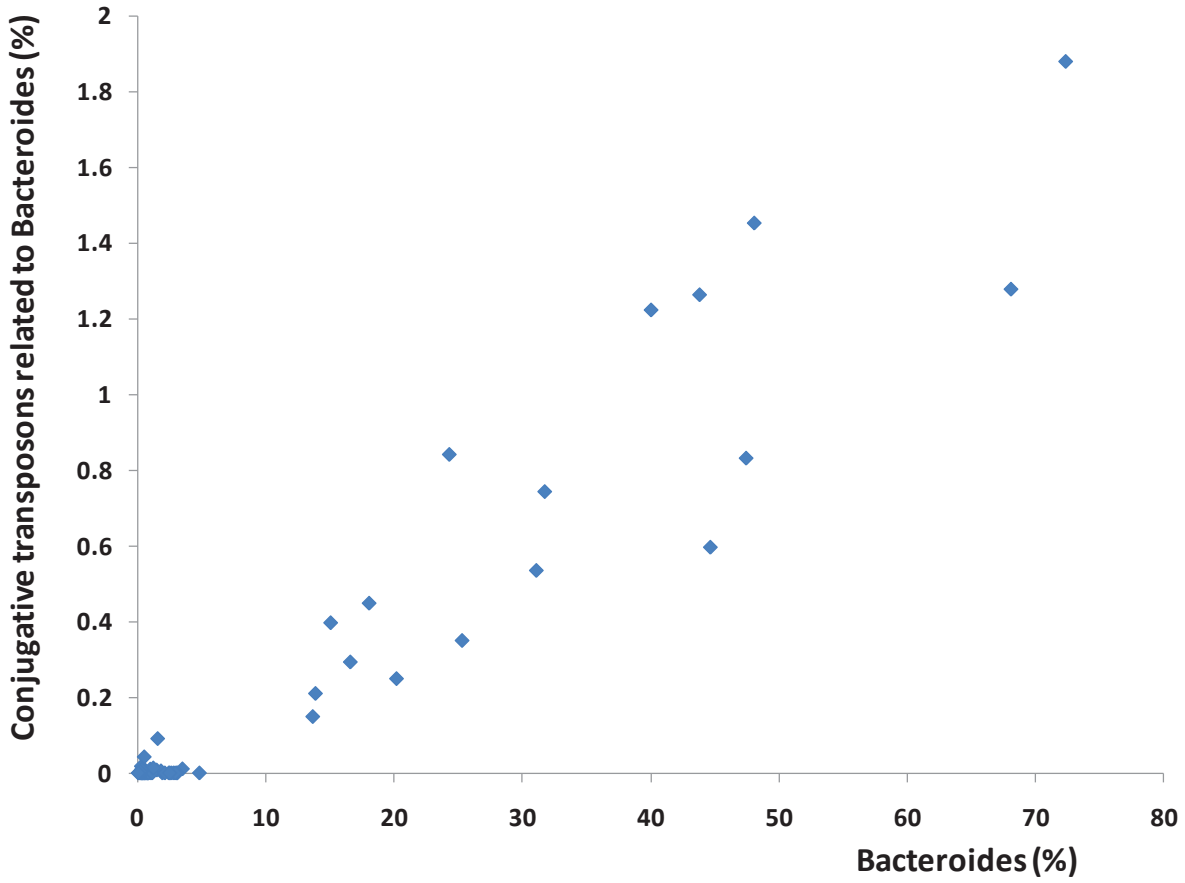


Figure 45. Relative distribution correlation between a phylogenetic group and a functional subsystem (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets.

Sequences related to Fe-S cluster assembly appear to be more represented in animals even if they are also detected in all metagenomes (Figure 44). Iron-sulfur clusters are stable at several oxidation states (redox potentials from −500 to 150 mV) [448]. As a consequence, Iron-sulfur (Fe-S) proteins are ubiquitous in both organisms and microorganisms and play an essential role in metalloenzyme catalysis and electron transport [449]. Iron-sulfur (Fe-S) clusters are necessary for respiration, photosynthesis, and nitrogen fixation. Three distinct pathways of Fe-S cluster assembly were identified and called Nif (nitrogen fixation), Isc (iron sulfur cluster) and Suf (sulfur formation) [450-452]. Ayala-Castro and colleagues suggested that due to specific functions these pathways can be divided into three roles: housekeeping cluster assembly for Isc, response to stress conditions for Suf, and finally assembly of complex or specialized clusters for specific enzymes [453 table of contents].

In all the metagenomes corresponding to animals, the majority of sequences related to this subsystem correspond to the Sul pathway and cysteine desulfurase enzymes which are required to liberate sulfur atoms from free cysteine for further cluster assemblies [454]. Of course and because only a tiny part of these metagenomes were sequenced, the other pathways are probably also present, but in lower proportion.

Genes involved in L-arabinose utilization were highly studied in molecular biology, and in particular in Bacillus subtilis which is able to grow on L-arabinose as the sole carbon and energy source (see [455]) and in Escherichia coli (see [456] as a review). L-arabinose utilization subsystem is detected in all the environments (Figure 44). However, it is detected more in the four animal microbial populations and could represent there an important mechanism to generate energy. In addition, the maltose and maltodextrin utilization subsystem is more represented in animals even if this difference is probably less visible than for L-arabinose (Figure 44). Bacillus subtilis and Escherichia coli are also models to study the transport, metabolism and regulation of the maltose/maltodextrin system [457,458], but these species are not correlated to these subsystems, emphasizing that other organisms use these molecules in the environment. These unusual distributions probably reflect an important density of sugar accessible for animal microbial populations.

During environmental stresses, some microorganisms have the capacity to produce a dormant cell called a spore which possesses high levels of biological resistance. Several bacilli and clostridia related species are known to produce a spore in response to starvation, and as an example, the sporulation in Bacillus subtilis is a physiological response to nutritional stress. During the cellular transformation, a proteinacious shell named the coat is assembled and aims to protect the spore (see [459,460] as a review). Genes related to the formation of a spore coat are clearly more represented in animals than in the other environments (Figure 44), and this observation provides insights about the life style strategies and resistance capacities of these microbial communities. This capacity can impact strongly antibiotic treatment success for example, and could stimulate the survival of a part of the community before the wide spray of antibiotic resistance genes to by-pass this stress.

Human feces microbial populations' specificities:

Some taxonomical and functional subsystems appear to be specific to human microbial populations. It is in particular the case of sequences related to Bifidobacterium, Lact-N-Biose I and Galacto-N-Biose metabolic pathway, sucrose utilization, beta-glucoside metabolism, ECF class transporters and nitrosative stress which are more detected in human feces (Figure 46). Bifidobacteria are ubiquitous commensal bacteria in the gastrointestinal tract of animals [461]. They use a particular mechanism that utilize both milk disaccharides and host glycoconjugates to generate energy. This mechanism was named the Lact-N-Biose I and Galacto-N-Biose metabolic pathway [462]. The important detection of sequences related to Bifidobacterium and to the Lact-N-Biose I and Galacto-N-Biose metabolic pathway in some human feces is probably due to the utilization of milk by human populations even after child weaning. In addition, the fact that sequences related to sucrose utilization and beta-glucoside metabolism are more present in human microbial populations than in the other animals is probably due to an unusual food supply of their hosts.

Energy-coupling factor (ECF) transporters import micronutrients in bacteria and archaea [463-465]. This system (especially the subclass II) is known to be widespread in Firmicutes and some archaea [466]. The relatively important distribution of this subsystem in human feces reflects probably a particular strategy to uptake some micronutrients available to human microbial populations.

Anaerobic microbial populations that use nitrate and nitrite as terminal electron acceptors produce a low amount of nitric oxide [467]. However, high concentration of nitric oxide is toxic for microorganisms and the mammalian hosts use this molecule into macrophages to eradicate inopportune species [468] [469]. To respond to this toxicity, microbial cells possess different mechanisms to react quickly during a nitrosative stress and to protect themselves to its damages [469-471]. The unusual distribution of sequences related to nitrosative stress in human feces probably reflects a conflict between some microorganisms and their host. However, the fact that this distribution is more important in human feces than in other animals have to be defines, but could be in a simple way due to the sampling localization of these metagenomes.

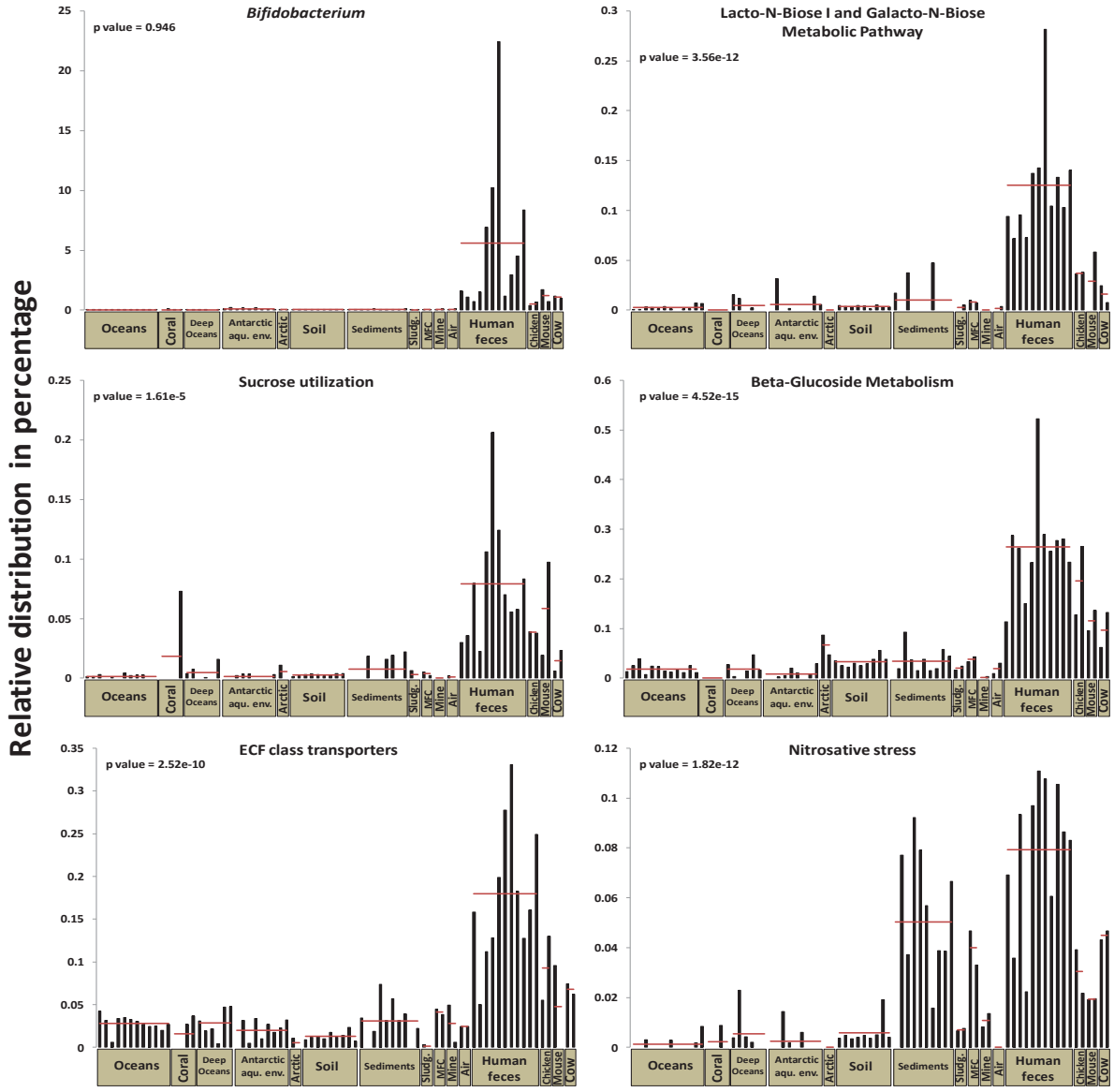


Figure 46. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Because faecal microorganisms appear to be representative to individual differences in the distal gut [418], further sequencing efforts based on human populations living close and far away from the civilization could provide considerable information about microbial communities' adaptation during the actual evolution of human food supply. In addition, it is possible that a considerable number of species and functions are correlated to some nutrients but also environmental stresses and to analyze patient metagenomes could help to define unique treatments. But in a general way, to compare human metagenomes corresponding to people living in a rural or urban environment will help defining the impact of the civilization on our second genome and so in our health. Give me your metagenome, I

will tell you what you eat and how you live based on specific functional and taxonomical distributions could also become a classical clinical approach in the future.

Mouse ceacum microbial populations' specificities:

Two metagenomes corresponding to a lean and an obese mouse ceacum were generated [4] and provided information about microorganisms metabolic distributions in obese individuals. Interestingly, results emphasized the gut microorganisms contribution to pathophysiology of obesity. In addition, based on this global metagenomic comparison, these metagenomes appear to possess unusual distributions.

In particular, sequences related to sucrose metabolism subsystem are more distributed in animals in general but are highly detected in the obese mouse in comparison to the other metagenomes (including the lean mouse, see figure 47). This subsystem reflects the host corpulence and could be used as a biomarker of obesity.

While the majority of nutrient phosphorous are in a high redox state (+5) in the nature and are used by microorganisms in this condition, some microorganisms can metabolize phosphate at a lower redox state (see [472] as a review). In particular, Phosphonates can represent an important part of the available phosphate in some environments even if the C-P bonds in these compounds are highly stable in contRAST to C-O-P bonds. Some microorganism can cleave carbon-phosphorous bonds (C-P lyase enzyme, [473] and so are able to degrade alkylphosphonates [474]. Sequences related to alkylphosphonate utilization are overrepresented in mouse microbial populations (Figure 47), and could reflect an important mechanism to uptake phosphate in a low redox state in this specific environment. However, when focusing on what is present inside this subsystem in these two metagenomes, all sequences are related to chloramphenicol acetyltransferase (chloramphenicol resistance in bacteria). Thus these mouse microbiomes probably have a particular potential to resist to chloramphenicol instead of taking up phosphate. This annotation error reflects the presence of occasional bugs in the MG-RAST annotation system. However, the majority of checked subsystems appeared to be correct.

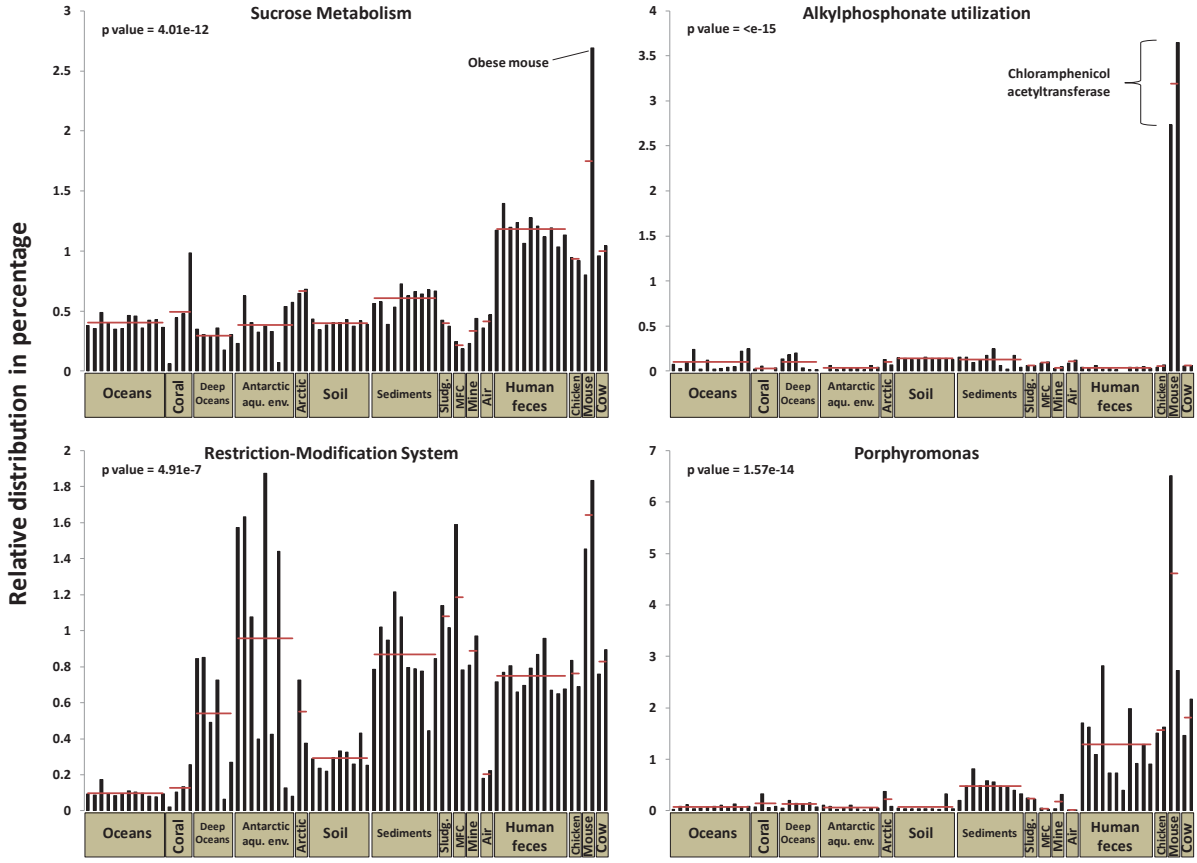


Figure 47. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Restriction-modification systems are considered as a bacterial immune system due to its capacity to protect from DNA invasion and in particular from bacteriophages [475]. The distribution of the related subsystem in mouse microbial populations (Figure 47) could reflect a particular strategy to protect genomes. However, this subsystem is also highly detected in other environments and so is not specific to these communities.

Porphyromonas gingivalis is one of the most frequenctly isolated bacteria from periodontitis lesions (Periodontitis is a chronic infectious disease of the tooth) and the genome of the strain W83 was sequenced [476]. This genome is the only representing this genus in the actual SEED database, and so sequences related to this strain could in reality represent other species. However, this genus appears to be more represented in animals than in the environment (Figure 47), and is especially highly detected in the lean mouse were it could have a particular role for microbial communities.

Cow rumen microbial populations' specificities:

In the two cow rumen metagenomes generated, the most unexpected observation was probably the detection of sequences matching from 97% to 99% with the human immunodeficiency virus 1 (Figure 48). Of course, because these metagenomic sequences have a length of approximately 100bp, this detection is not sufficient to reveal if the entire structure (>9 kbp) is present and if this virus is active there. However, it is interesting from our point of view to detect sequences from one of the most studied viruses in these metagenomes. Additional experiments should be performed to study the entire structure corresponding to these sequences in this specific environment. Some genera and species are also unusually highly detected in these metagenomes. It is the case of Parabacteroides, Alkaliphilus, Actinobacillus succinogenes, Methanobrevibacter and Clostridium phytofermentans ISDg (Figure 48).

Parabacteroides strains can be isolated from feces [477] and possess a considerable rate of resistance to various antibiotics [478]. In particular, 87.5% of the tested strains were resistant to tetracycline. The important distribution of this genus in cow rumens (between 4 and 6 percent of the annotated sequences) could explain in part the unusual distribution of genes related to tetracycline resistance in this environment (see figure 15).

Alkaliphilus species are present in various environments. As examples, Alkaliphilus halophilus was isolated from a saline lake [479], Alkaliphilus transvaalensis from a deep South African gold mine [480], and Alkaliphilus transvaalensis from a methanogenic environment [481]. The two sequenced and detected species, Alkaliphilus metalliredigens QYMF and Alkaliphilus oremlandi oremlandii OhILAs are equally detected in the 77 metagenomes. It is probable that Alkaliphilus species living in cow rumens correspond to not yet sequenced organisms and Alkaliphilus transvaalensis, which is known to live in methanogenic environments could be one of them.

Actinobacillus succinogenes 130Z was isolated from the bovine rumen [482,483]. Microorganisms from rumen are known to transform plant carbohydrates to fatty acids and in particular to succinic acid. However this strain can accumulate unusually high concentrations of succinic acid [484] and probably plays an important symbiotic role in the rumen [483]. The distribution of this species among environments (undetected in other animals for example) is a typical example of correlation between metagenomic and pure culture approaches.

Methanobrevibacter affiliated species are known to represent the major part of archaea in the cow rumen [485]. They represent also the majority of methanogens in rumen microbial populations (e.g., [486] [487] [488]. Because methanogens produce non negligible quantities of methane and impact the global warming, Methanobrevibacter species are of interest for scientists (e.g., by sequencing the genome of Methanobrevibacter ruminantium, [489]).

Metagenomic approaches confirm the particular role of this genus in cow rumens (Figure 48).

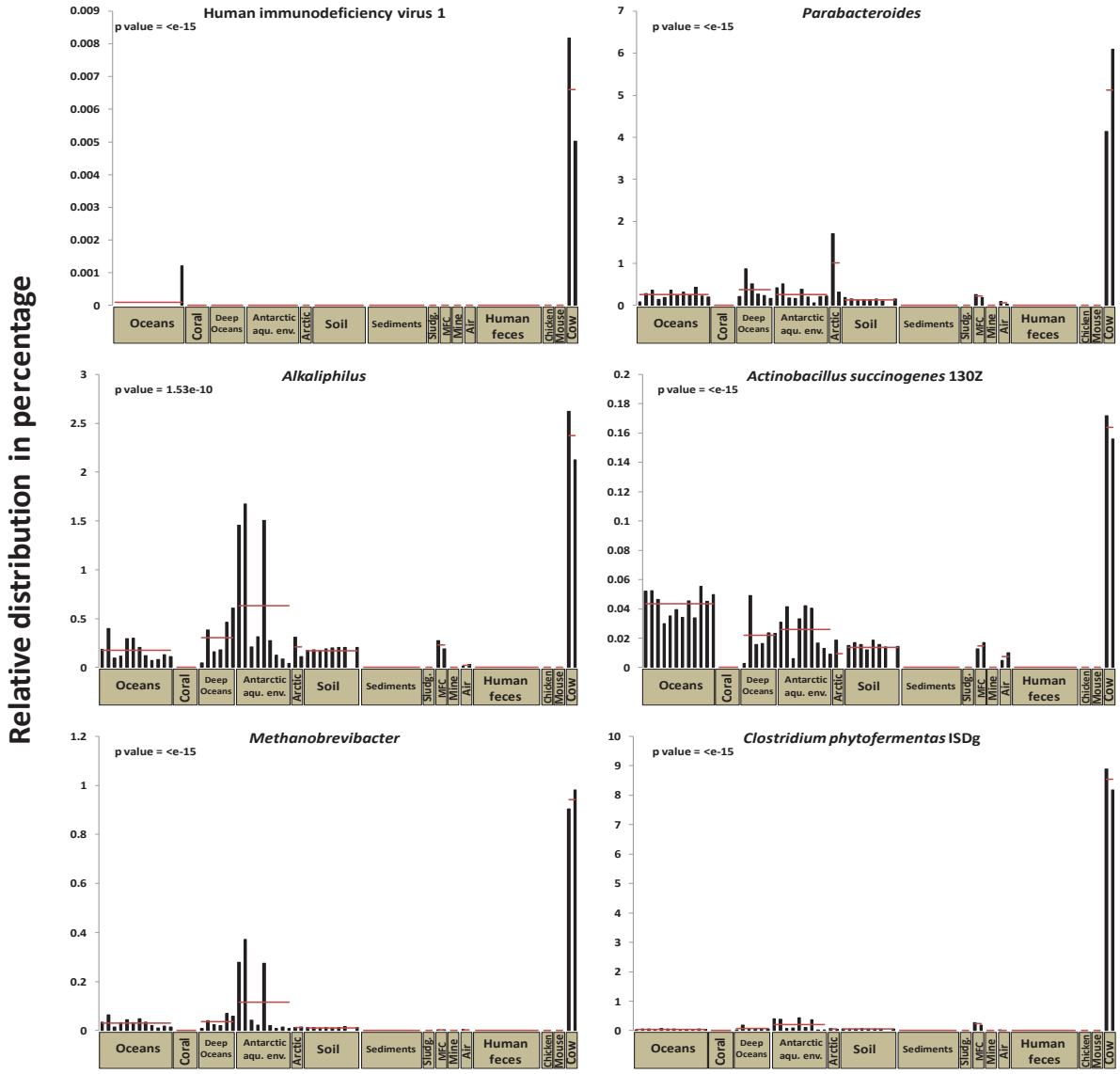


Figure 48. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Clostridium phytofermentans ISDg is a cellulolytic species isolated from a forest soil [490]. Interestingly, its optimum temperature is 35-37°C which is inconsistent with soil environments. Its distribution among the 77 metagenomes highlights a more suitable environment for this species or at least for species possessing a similar genome (the cow rumen where it is detected at more than 8%) and a particular roles (plant-derived polysaccharides degradation). In this case and in contRAST to Actinobacillus succinogenes

130Z, metagenomic approaches provide additional (and in part contradictory) information to culture based methods that are crucial to understand the ecological role of the strain.

Human and Mouse microbial population specificities:

Among the functions and genera detected, human and mouse microbial populations appear to possess some distribution specificities. It is in particular the case of sequences related to the Bacillus genus, multidrug Resistance Efflux Pumps (Figure 15), cellulosome, ATP-dependant Nuclease, Fructoologosaccharides (FOS) and Raffinose utilization, Xylose utilization and sortase subsystems which are in average more represented in human feces and mouse ceacum than in the other compared environments (Figure 49). Bacillus is a Gram positive rod-shaped genus member of the Firmicute phylum. Bacillus related sequences are detected in all the metagenomes, but its distribution increases in human feces and mouse ceacum. Some of these species are pathogen for human (e.g., Bacillus anthracis, see [491] as a review), others are probiotics (e.g., [492] and possess some beneficial potentials. To understand the role of these species in human microbial populations will provide strategies to favorably change the ratio pathogen/probiotic bacillus in the human microbiota.

The cellulosome corresponds to multienzyme complexes that can degrade plant cell wall. These enzymes are highly studied and looked for due to an important industrial interest (e.g., biomass degradation; production of butanol, ethanol and amino acids from sugars). They are produced by anaerobic bacteria and were firstly characterized in specific clostridium species [493,494] and fungi (Teunissen and Op den Camp, 1993). While ruminant rumens are known to be highly efficient environments for the degradation of plant biomass by cellulosome enzymes (e.g., [495] [496] [84], this function is not particularly highly detected in the two cow rumens but is more represented in human feces and mouse ceacum. The distribution of this subsystem fluctuates among human feces related metagenomes and could be due to food supply differences between individuals. In addition, the capacity of microbial populations to convert plant biomass into sugars provides an additional source of energy to the host.

The ATP-dependent nuclease genes addA and addB present in Bacillus subtilis [497] are the analogs of the highly studied Escherichia coli RecBCD genes (e.g., [498] [499] [500]). These genes possess helicase and nuclease activities. They repair double-stranded DNA breaks in Bacillus subtilis and play a crucial role in homologous transformation [501]. In addition, this ATP-dependent nuclease is required for the stability of several plasmids [502] [503] [504] [505]. In Helicobacter pylori, addA and addB are also required for mouse infectivity [506]. The ATP-dependent nuclease SEED subsystem represents to addA and addB genes, and its distribution is correlated with the Bacillus genus ($R^2>0.78$), confirming their association.

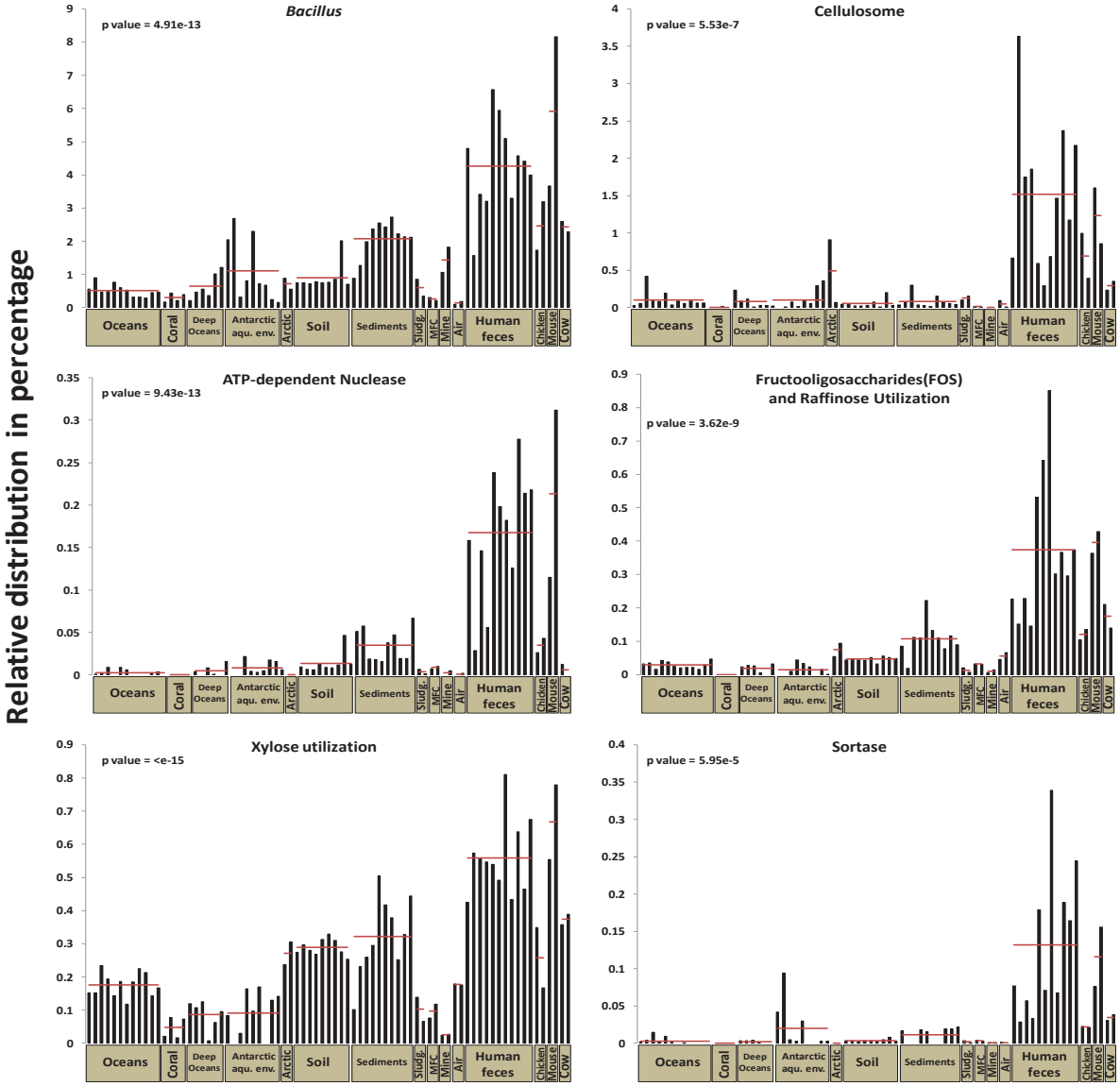


Figure 49. Relative distribution (in percentage of annotated reads) of different microbial phylogenetic groups and functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

Fructooligosaccharides are prebiotic carbohydrates that have a benefit effect on the growth and activity of the probiotic bacteria Bifdobacterium and Lactobacillus (e.g., [507] [508]). The utilization of Fructooligosaccharides by probiotic bacteria can limit the proportion of pathogenic microorgansisms in animal microbiotas (e.g., [509] [510]). Raffinose can be used by probiotic bacteria, but is also used in other species, like Enterococcus faecium where the corresponding genetic element is present in a megaplasmid [511], underlining lateral gene transfert cappacities of this function among microbiota populations. While the subsystem corresponding to Fructooligosaccharides and Raffinose utilization is more detected in human

and mouse microbiotas, its distribution varies between individuals. Interestingly, this distribution is positively correlated to the sum of Bifidobacterium and Lactobacillus genera (R2>0.82) among the 77 metagenomes (supplement data).

Xylose is degraded from Xylan and xyloglucan, two major plant cell wall components providing energy to several organisms [512] [513] [514]. Xylose is then transformed to xylulose 5-phosphate, a metabolic intermediaire. This pathway is of industrial interest for the transformation of Lignocellulosic feedstocks into ethanol (e.g., [515]). Based on its distribution among the 77 metagenomes, xylose appears to be more metabolized in human and mouse microbiotas than in the other animals and environments. Sortase transpeptidase enzymes are synthetized by Gram positive bacteria to covalently attach the host cell wall [516]. They polymerize pili proteins for bacterial adhesion ([517,518]) and contribute largely to the virulence of several pathogens (e.g., in Staphylococcus aureus, [516] [519]). Efforts are now done to inhibit sortase proteins to limit pathogenic species virulence [520]. In addition, Sortase enzymes are studied for protein engineering due to an important protein ligation potential (e.g., [521]). This function is detected in several environments, but is unusually highly represented in some human and mouse microbiotas. The important distribution fluctuation of this subsystem in human feces metagenomes represents different virulence potentials of the related microbial communities.

Chicken and Cow microbial population specificities:

Cow rumen and chicken ceacum microbial communities appear to be globally relatively similar (see figure 11 and 41). Some functional distributions appear to be unusually distributed in these two distinct environments and can in part explain this similarity visualized in the different PCA. If the distribution of the genes related to cell division are correlated to communities' activity, microorganisms from animals in general and cow rumen and chicken ceacum in particular are highly active. In comparison to some environments like in the atmosphere where the nutrient (especially iron) availability is probably limited [52], animals apparently provide some opportunities for communities to develop themselves quickly. This possibly unusual activity is consistant with the distribution of sequences related to the bacterial RNA polymerase subsystem, which like for the cell division cluster is highly detected in the cow rumen and the chicken ceacum in comparison to the other environments (Figure 50). RNA polymerase distribution is known to be coupled to dynamique structure of the nucleoid [522] which responds to environmental conditions. However, sequences related to some other subsystems related to bacterial activity (e.g., bacterial transcription initiation sigma factors) are not particularly highly detected in these environments (supplement data).

The stringent response is a regulatory reaction in response to amino acid starvation or other crucial element limitations (see [523] [524]) as reviews). During starvation, guanosine tetra-

and pentaphosphates are accumulated and bind to the bacterial RNA polymerase and regulate the transcription of specific genes (e.g., induction of stress response related genes and inhibition of transcriptional factors). However, the different roles of these molecules are still debatable (see [524]) but appear to impact several aspects of cell biology and to provide particular abilities for bacteria to survive under hostile conditions. Because animal microbiotas live in an environment unstable in term of nutrient availability, they appear to have particular potentials to develop quickly in certain period and to stop their activity during starvation phases.

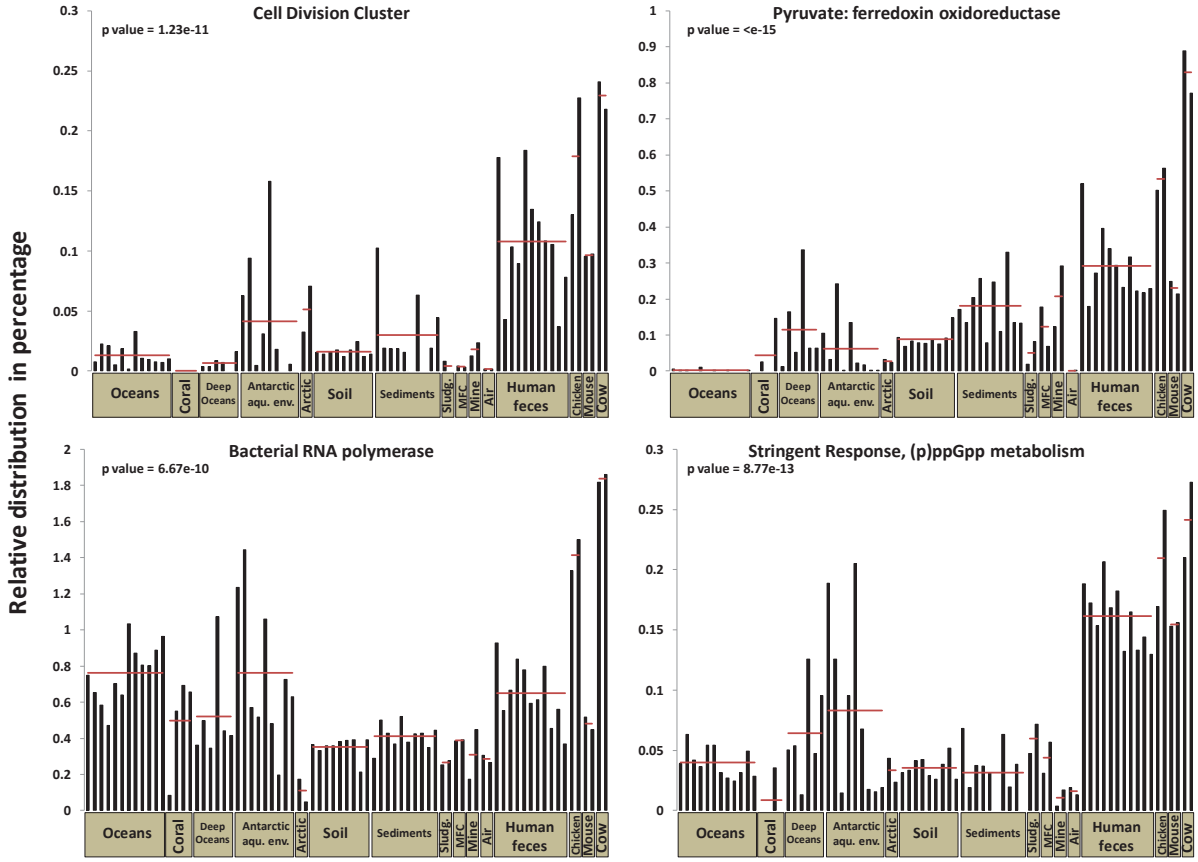


Figure 50. Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

The pyruvate:ferredoxin oxidoreductase enzyme catalyses the decarboxylation of pyruvate to acetyl CoA and CO2 [525], an important metabolism for anaerobic microorganisms to generate energy without oxidizing pyruvate. Interestingly, this subsystem appears to be a good marker to track anaerobic environments. Its distribution, which is limited on the surface of oceans and in the atmosphere, is greater in deep oceans, sediments, soils and animals. Its distribution in chicken ceacum and cow rumen highlights particular microbial populations, dominated by anaerobic microorganisms (see table 1).

Discussion:

Microorganisms are the first forms of life to colonize our planet. They have evolved to adapt to increasing diversity of environmental niches and while it is difficult to study ancient forms of microorganisms due to a lack of stored samples, current bacterial, archaeal, viral and fungal communities living in various environments across the planet are accessible. Recently, new technologies and ambitious projects have generated several environmental genomic datasets (metagenomes) corresponding to microbial populations extracted from different environments on Earth (e.g., [38] [90]).

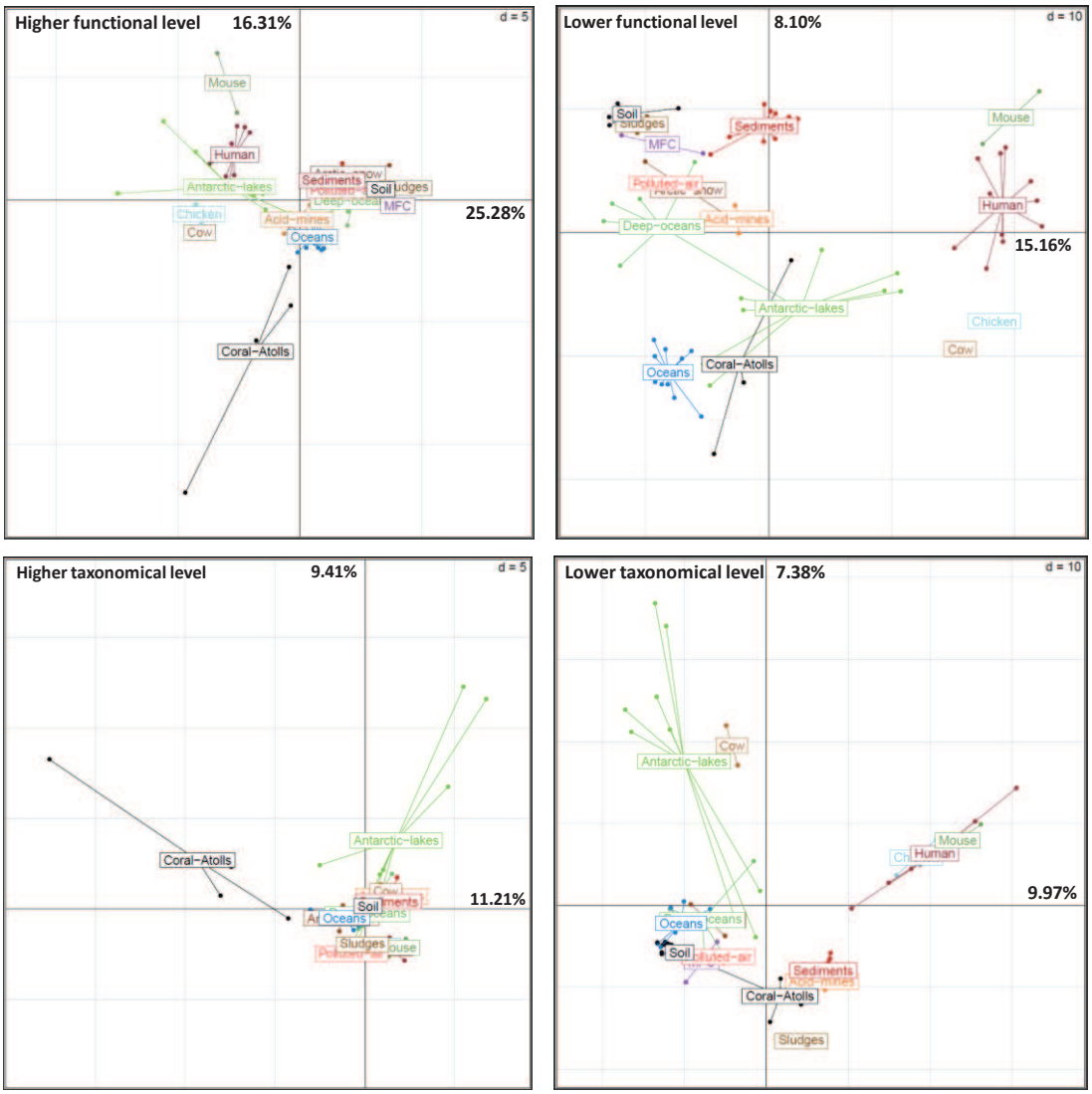


Figure 51. Principal component analysis based on the relative distribution (in percentage of annotated reads) of different taxonomical and functional levels (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. The percentages of the illustrated two major axes correspond to the fraction of the total variance that they represent.

Comparing these entire datasets (e.g., distribution of all detected functions) provides general information about similarities and differences between metagenomes (and by

extrapolation their corresponding environments) (Figure 51). For example, animal microbiomes are grouped together and separated to the other environments at the lower functional subsystem level but not at the higher taxonomical level (phyla). However, the location of the 77 metagenomes in these principal component analyses is due to the distribution of hundreds of subsystems, and to appreciate metagenomes/environments specificities, the distribution of these subsystems have to be compared individually [92].

The aim of this study was to emphasize microbial community peculiarities by confronting 77 available metagenomes corresponding to 15 distinct environments. Interestingly, each environment appears to possess a unique nucleic richness at both functional and taxonomical levels (Tables 1 and 2). However, the principal genera are more divergent in the different metagenomes than the principal functions, so underlining the fact that some functions are indispensable in high proportion for microbial cells processes whatever the environment and community structure.

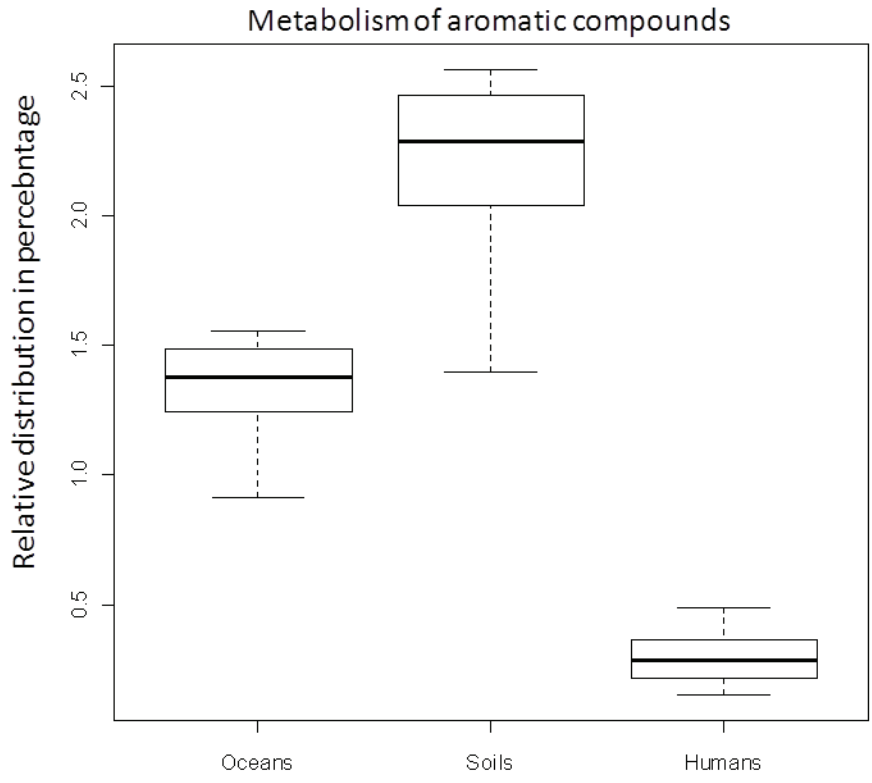


Figure 52. Box plot representing the relative distribution (in percentage of annotated reads) of a functional subsystem (based on SEED assignments of sequenced genomes in the MG-RAST program) for datasets related to three distinct environments: Oceans, Soils and Humans.

A minimum of two metagenomes per environment is critical and additional datasets need to be generated to improve comparisons. As an example, if selecting only metagenomes from

soil, ocean and human microbial populations, statistical analyses can be performed for inter-environmental comparisons (e.g., box plots in figure 52). However, to compare only these environments limits the possible comparisons at this moment, so we selected all environments where at least two metagenomes were available. Given the danger of assuming functional distributions, statistical analyses by non-parametric approaches should be performed after compiling a minimum number of metagenomes for robust statistical comparisons (e.g., Kruskal-Wallis H-test which needs a minimum of five datasets per environment).

Over 800 functional and almost 1600 taxonomical subsystems (SEED) were detected using the MG-RAST annotation server [75] and provide information about the different metagenomes. Not all of these differences have been shown and discussed here. While the relative distribution of particular species and functions were used to separate different ecosystems in principal component analyses, only select species and functions were illustrated to emphasize the interest of global metagenomic comparisons (e.g., genes involved in resistance to metals or antibiotics). This choice is based on our appreciation of possible interests and the absence of other comparisons does not mean they cannot be made.

Even if only partial bioinformatics and statistical analyses were applied to these datasets here, these simple observations provide/confirm information about how life evolved and adapted to specific environmental physical and chemical characteristics, such as oxygen, temperature, pH, salinity, UV radiation. The observations presented here can be used to confirm results but also to create new hypotheses about how microorganisms have adapted. For example, microorganisms have developed different strategies to optimize their function under varying nutrient conditions (e.g. nutrients uptake in animals, photosynthesis in oceans and coral atolls) and resist potentially harmful environmental conditions (e.g., temperature in Arctic snow, salinity in hypersaline sediments). The study of different microbial communities among the different ecosystems can provide information about how to improve human heath by studying animal microbial population specificities in general (see animal section) and antibiotic resistance gene distributions in particular.

In addition, some unexpected observations or correlations can be discovered requiring additional experimental work to uncover the hidden meaning. As an example, the distribution of sequences related to 16S rRNA and 23S rRNA genes are globally correlated, but in two metagenomes related to ocean and in the two indoor polluted air the 16S rRNA gene is unusually highly detected (more than 4% of the annotated sequences; figure 53). If this distribution is not due to some contamination or other error, it might reflect a particular life style of the related bacterial and archaeal communities. Another example is the distribution of sequences related to human genome which is unexpectedly more highly represented in coral atolls than in human feces (Figure 53). This could be of course a sign of sample contamination.

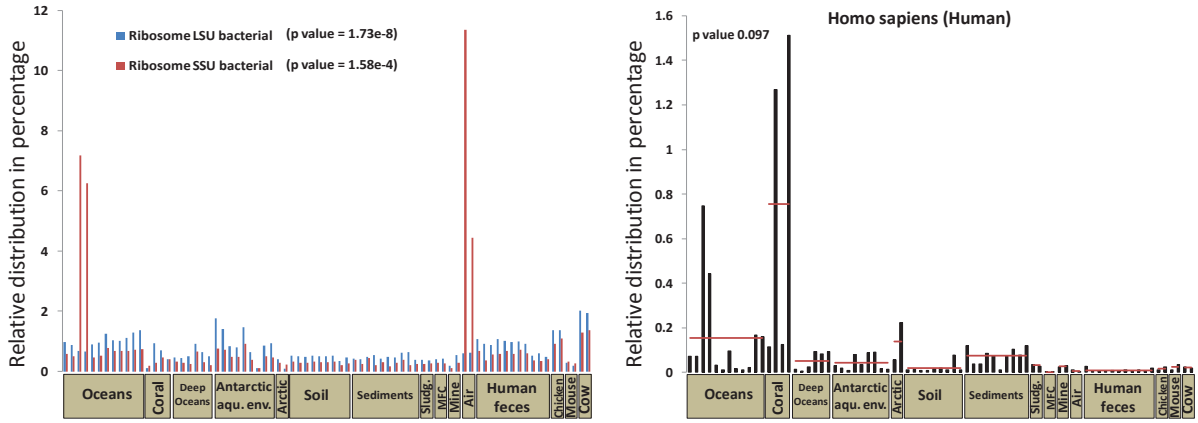


Figure 53. Relative distribution (in percentage of annotated reads) of the Human phylogenetic groups and two functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for the 77 metagenomic datasets. Horizontal lines are the mathematical averages for the metagenomic datasets from each ecosystem. The p-values are the likelihoods that the distribution differences observed between environments are random.

The majority of observations is generally coherent with our current knowledge about the 10^{30} Bacteria and Archaea estimated to live on Earth. An important limit of metagenomic approaches is that annotation is dependent of the discovery of functions based on known microorganisms. Current exploration of sequences not related to known functions will expand the possible annotation. Yet even now, global metagenomic comparisons can be used to test correlation between functions and species (e.g., figure 45) as was schematized in the figure 54. As a consequence, it is possible using these comparisons to corroborate discoveries based on pure culture approaches.

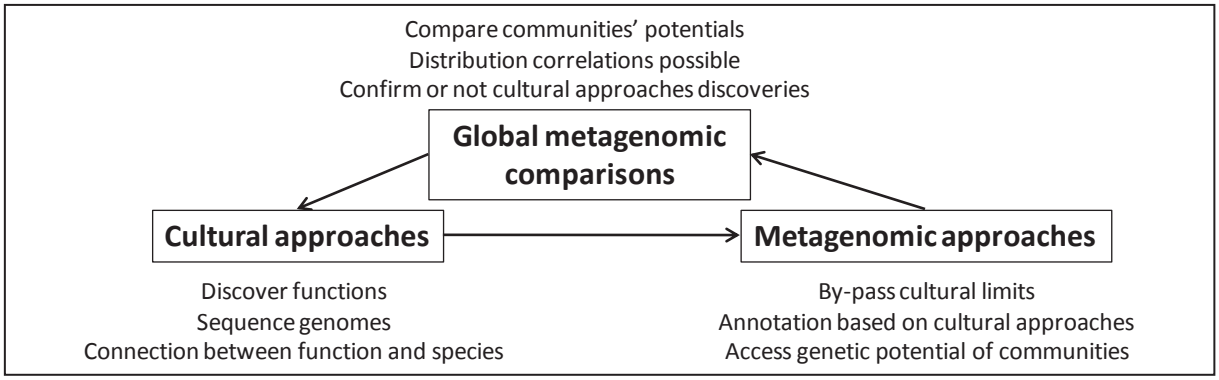


Figure 54. Schema representing interactions between cultural, metagenomic and dataset comparison approaches to study microorganisms.

As metagenomics and high throughput sequencing are still relatively new and undeveloped, a gap between metagenomic and cellular culture approaches exists. The species known to be involved in the different process presented (e.g. resistance to heavy metals and aromatic compounds) are in several cases not correlated to their functions. This result can easily be

explained by the important limits of the two approaches [62,526]. The crux of the problem will be determining whether the presence of the functions (and associated genes) is much more variable than that supposed by pure culture studies (i.e., not always in organisms associated with the function and found in those not associated with the function). So there is a need to improve our access to more complete genomic data from microbial communities. However, in some cases, pure culture and metagenomic approaches have provided very coherent results emphasizing that our perception of microorganisms might not be completely biased.

Conclusion:

This metagenomic comparison decrypts microbial community peculiarities by confronting datasets generated by various research groups and corresponding to 15 distinct environments. Due to the complexity of microorganisms at the planet level, the task of discovering and representing all specificities of these communities is considerable. As a consequence and in spite of a particular effort to present observations without any a priori conclusions, this study is more an assessment of what might be derived from future more detailed metagenomic studies. While some hypotheses were proposed here, they cannot be considered anything other than apparent correlations and not mechanistic demonstrations, which need to be done by appropriate specialists of the given ecosystems. Available datasets need to be exploited and observations presented here aim to stimulate the interest of global metagenomic comparisons. One interesting point highlighted by this study is that pure culture and metagenomic approaches can in some cases be in ad equation. Thus inter-environmental comparisons could be the missing link between the two approaches due to the possibility to perform correlation tests between function and species.

In the future and to improve the study of microbial communities, additional microbial genomes from various environments need to be sequenced and datasets re-annotated. In addition, the evolution of annotation pipelines will stimulate the study and comparison of metagenomes. But perhaps more important, statistical tools applied to metagenomes should evolve to integrate the comparison of several environments represented by two or more metagenomes as was done here.

Acknowledgments:

We want to thanks Catherine Larose and Jean-Michel Monier for providing metagenomes of high scientific interest from Arctic snow and Microbial fuel cells and for their help editing these sections.

References

1. Altermann W, Kazmierczak J (2003) Archean microfossils: a reappraisal of early life on Earth. Res Microbiol 154: 611-617.

2. Siefert JL (2009) Defining the mobilome. Methods Mol Biol 532: 13-27.

3. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A 95: 6578-6583.

4. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.

5. Falkowski PG, Rosenthal Y (2001) Biological diversity and resource plunder in the geological record: casual correlations or causal relationships? Proc Natl Acad Sci U S A 98: 4290-4292.

6. Porter JR (1976) Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. Bacteriol Rev 40: 260-269.

7. Sugihara TF, Kline L, Miller MW (1971) Microorganisms of the San Francisco sour dough bread process. I. Yeasts responsible for the leavening action. Appl Microbiol 21: 456-458.

8. Thomas SB (1969) Methods of assessing the psychrotrophic bacterial content of milk. J Appl Bacteriol 32: 269-296.

9. Bryant MP (1974) Nutritional features and ecology of predominant anaerobic bacteria of the intestinal tract. Am J Clin Nutr 27: 1313-1319.

10. Ward DM, Ferris MJ, Nold SC, Bateson MM (1998) A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. Microbiol Mol Biol Rev 62: 1353-1370.

11. Zobell CE (1946) Action of microorganisms on hydrocarbons. Bacteriol Rev 10: 1-49.

12. Iwai Y, Omura S (1982) Culture conditions for screening of new antibiotics. J Antibiot (Tokyo) 35: 123-141.

13. Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev 59: 143-169.

14. Hobbie JE, Daley RJ, Jasper S (1977) Use of nuclepore filters for counting bacteria by fluorescence microscopy. Appl Environ Microbiol 33: 1225-1228.

15. Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103: 12115-12120.

16. Morris CE, Bardin M, Berge O, Frey-Klett P, Fromin N, et al. (2002) Microbial biodiversity: approaches to experimental design and hypothesis testing in primary scientific literature from 1975 to 1999. Microbiol Mol Biol Rev 66: 592-616, table of contents.

17. Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87: 4576-4579.

18. Fromin N, Hamelin J, Tarnawski S, Roesti D, Jourdain-Miserez K, et al. (2002) Statistical analysis of denaturing gel electrophoresis (DGE) fingerprinting patterns. Environ Microbiol 4: 634-643.

19. Pringault O, Viret H, Duran R (2011) Interactions between Zn and bacteria in marine tropical coastal sediments. Environ Sci Pollut Res Int.

20. Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol 40: 337-365.

21. Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, et al. (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. Proc Natl Acad Sci U S A 103: 13104-13109.

22. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, et al. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1: 283-290.

23. Mendum TA, Chilima BZ, Hirsch PR (2000) The PCR amplification of non-tuberculous mycobacterial 16S rRNA sequences from soil. FEMS Microbiol Lett 185: 189-192.

24. Yergeau E, Bokhorst S, Kang S, Zhou J, Greer CW, et al. (2011) Shifts in soil microorganisms in response to warming are consistent across a range of Antarctic environments. ISME J.

25. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245-249.

26. Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, et al. (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66: 2541-2547.

27. Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, et al. (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol 68: 4301-4306.

28. Knietsch A, Waschkowitz T, Bowien S, Henne A, Daniel R (2003) Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on Escherichia coli. Appl Environ Microbiol 69: 1408-1416.

29. Sebat JL, Colwell FS, Crawford RL (2003) Metagenomic profiling: microarray analysis of an environmental genomic library. Appl Environ Microbiol 69: 4927-4934.

30. Bertrand H, Poly F, Van VT, Lombard N, Nalin R, et al. (2005) High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. J Microbiol Methods 62: 1-11.

31. Nacke H, Will C, Herzog S, Nowka B, Engelhaupt M, et al. Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German Biodiversity Exploratories. FEMS Microbiol Ecol.

32. Hu Y, Fu C, Huang Y, Yin Y, Cheng G, et al. Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment. FEMS Microbiol Ecol 72: 228-237.

33. Torres-Cortes G, Millan V, Ramirez-Saad HC, Nisa-Martinez R, Toro N, et al. Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. Environ Microbiol 13: 1101-1114.

34. Sharma S, Khan FG, Qazi GN Molecular cloning and characterization of amylase from soil metagenomic library derived from Northwestern Himalayas. Appl Microbiol Biotechnol 86: 1821-1828.

35. Meilleur C, Hupe JF, Juteau P, Shareck F (2009) Isolation and characterization of a new alkali-thermostable lipase cloned from a metagenomic library. J Ind Microbiol Biotechnol 36: 853-861.

36. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.

37. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.

38. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.

39. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, et al. (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311: 392-394.

40. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, et al. (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113-1118.

41. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.

42. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

43. Boubakri H, Beuf M, Simonet P, Vogel TM (2006) Development of metagenomic DNA shuffling for the construction of a xenobiotic gene. Gene 375: 87-94.

44. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5: e16.

45. Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, et al. (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS One 3: e1456.

46. Martin-Cuadrado AB, Lopez-Garcia P, Alba JC, Moreira D, Monticelli L, et al. (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. PLoS One 2: e914.

47. Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, et al. (2008) Microbial ecology of four coral atolls in the Northern Line Islands. PLoS One 3: e1584.

48. Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Natl Acad Sci U S A 105: 10583-10588.

49. Konstantinidis KT, Braff J, Karl DM, DeLong EF (2009) Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. Appl Environ Microbiol 75: 5345-5355.

50. Kunin V, Raes J, Harris JK, Spear JR, Walker JJ, et al. (2008) Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. Mol Syst Biol 4: 198.

51. Larose C, Berger S, Ferrari C, Navarro E, Dommergue A, et al. (2010) Microbial sequences retrieved from environmental samples from seasonal arctic snow and meltwater from Svalbard, Norway. Extremophiles 14: 205-212.

52. Tringe SG, Zhang T, Liu X, Yu Y, Lee WH, et al. (2008) The airborne metagenome in an indoor urban environment. PLoS One 3: e1862.

53. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174-180.

54. Rohwer F, Seguritan V, Choi DH, Segall AM, Azam F (2001) Production of shotgun libraries using random amplification. Biotechniques 31: 108-112, 114-106, 118.

55. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99: 14250-14255.

56. Abulencia CB, Wyborski DL, Garcia JA, Podar M, Chen W, et al. (2006) Environmental whole-genome amplification to access microbial populations in contaminated sediments. Appl Environ Microbiol 72: 3291-3301.

57. Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J 2: 233-241.

58. Yilmaz S, Allgaier M, Hugenholtz P (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Methods 7: 943-944.

59. Woyke T, Xie G, Copeland A, Gonzalez JM, Han C, et al. (2009) Assembling the marine metagenome, one cell at a time. PLoS One 4: e5299.

60. Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, et al. (2010) One bacterial cell, one complete genome. PLoS One 5: e10314.

61. Schloss PD, Handelsman J (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome Biol 6: 229.

62. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, et al. (2011) Accessing the soil metagenome for studies of microbial diversity. Appl Environ Microbiol 77: 1315-1324.

63. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145.

64. Kahvejian A, Quackenbush J, Thompson JF (2008) What would you do if you could sequence everything? Nat Biotechnol 26: 1125-1133.

65. Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 34: 5623-5630.

66. (2007) The Universal Protein Resource (UniProt). Nucleic Acids Res 35: D193-197.

67. (2008) The universal protein resource (UniProt). Nucleic Acids Res 36: D190-195.

68. (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 37: D169-174.

69. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res 38: D142-148.

70. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377-386.

71. Huson DH, Richter DC, Mitra S, Auch AF, Schuster SC (2009) Methods for comparative metagenomics. BMC Bioinformatics 10 Suppl 1: S12.

72. Mitra S, Klar B, Huson DH (2009) Visual and statistical comparison of metagenomes. Bioinformatics 25: 1849-1855.

73. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biol 5: e75.

74. Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, et al. (2006) An experimental metagenome data management and analysis system. Bioinformatics 22: e359-367.

75. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

76. Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26: 715-721.

77. Hu GQ, Guo JT, Liu YC, Zhu H (2009) MetaTISA: Metagenomic Translation Initiation Site Annotator for improving gene start prediction. Bioinformatics 25: 1843-1845.

78. Hoff KJ, Lingner T, Meinicke P, Tech M (2009) Orphelia: predicting genes in metagenomic sequencing reads. Nucleic Acids Res 37: W101-105.

79. Kristiansson E, Hugenholtz P, Dalevi D (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics 25: 2737-2738.

80. Gerlach W, Junemann S, Tille F, Goesmann A, Stoye J (2009) WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. BMC Bioinformatics 10: 430.

81. Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, et al. (2010) METAREP: JCVI metagenomics reports--an open source tool for high-performance comparative metagenomics. Bioinformatics 26: 2631-2632.

82. Nealson KH, Venter JC (2007) Metagenomics and the global ocean survey: what's in it for us, and why should we care? ISME J 1: 185-187.

83. Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, et al. (2010) TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Microbiol. pp. 252.

84. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, et al. (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science 331: 463-467.

85. (2010) A sequence of changes. Nat Rev Microbiol 8: 85.

86. Sleator RD, Shortall C, Hill C (2008) Metagenomics. Lett Appl Microbiol 47: 361-366.

87. Morales SE, Cosart TF, Johnson JV, Holben WE (2009) Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. Appl Environ Microbiol 75: 668-675.

88. Alonso-Saez L, Sanchez O, Gasol JM, Balague V, Pedros-Alio C (2008) Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. Environ Microbiol 10: 2444-2454.

89. Rajendhran J, Gunasekaran P (2008) Strategies for accessing soil metagenome for desired applications. Biotechnol Adv 26: 576-590.

90. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629-632.

91. Willner D, Thurber RV, Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol 11: 1752-1766.

92. Delmont TO, Malandain C, Prestat E, Larose C, Monier JM, et al. Metagenomic mining for microbiologists. ISME J.

93. Zhou J, Wu L, Deng Y, Zhi X, Jiang YH, et al. (2011) Reproducibility and quantitation of amplicon sequencing-based detection. ISME J 5: 1303-1313.

94. Higginbottom J, Bagnall KM, Harris PF, Slater JH, Porter GA (1976) Ultrasound monitoring of fetal movements. A method for assessing fetal development? Lancet 1: 719-721.

95. Thioulouse J (1989) Statistical analysis and graphical display of multivariate data on the Macintosh. Comput Appl Biosci 5: 287-292.

96. Lechevalier HA, Lechevalier MP (1967) Biology of actinomycetes. Annu Rev Microbiol 21: 71-100.

97. Bruggemann H, Baumer S, Fricke WF, Wiezer A, Liesegang H, et al. (2003) The genome sequence of Clostridium tetani, the causative agent of tetanus disease. Proc Natl Acad Sci U S A 100: 1316-1321.

98. Sowers KR, Baron SF, Ferry JG (1984) Methanosarcina acetivorans sp. nov., an Acetotrophic Methane-Producing Bacterium Isolated from Marine Sediments. Appl Environ Microbiol 47: 971-978.

99. Rohlin L, Gunsalus RP (2010) Carbon-dependent control of electron transfer and central carbon pathway genes for methane biosynthesis in the Archaean, Methanosarcina acetivorans strain C2A. BMC Microbiol 10: 62.

100. Krajmalnik-Brown R, Holscher T, Thomson IN, Saunders FM, Ritalahti KM, et al. (2004) Genetic identification of a putative vinyl chloride reductase in Dehalococcoides sp. strain BAV1. Appl Environ Microbiol 70: 6347-6351.

101. Lovley DR, Giovannoni SJ, White DC, Champine JE, Phillips EJ, et al. (1993) Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. Arch Microbiol 159: 336-344.

102. Juarez JF, Zamarro MT, Barragan MJ, Blazquez B, Boll M, et al. (2010) Identification of the Geobacter metallireducens bamVW two-component system, involved in transcriptional regulation of aromatic degradation. Appl Environ Microbiol 76: 383-385.

103. Gauthier MJ, Lafay B, Christen R, Fernandez L, Acquaviva M, et al. (1992) Marinobacter hydrocarbonoclasticus gen. nov., sp. nov., a new, extremely halotolerant, hydrocarbon-degrading marine bacterium. Int J Syst Bacteriol 42: 568-576.

104. Rokyta DR, Burch CL, Caudle SB, Wichman HA (2006) Horizontal gene transfer and the evolution of microvirid coliphage genomes. J Bacteriol 188: 1134-1142.

105. Maniloff J, Ackermann HW (1998) Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales. Arch Virol 143: 2051-2063.

106. Raoult D, Audic S, Robert C, Abergel C, Renesto P, et al. (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306: 1344-1350.

107. Claverie JM, Abergel C (2009) Mimivirus and its virophage. Annu Rev Genet 43: 49-66.

108. Filee J (2009) Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses. J Invertebr Pathol 101: 169-171.

109. Kennaway EL, Hieger I (1930) Carcinogenic Substances and Their Fluorescence Spectra. Br Med J 1: 1044-1046.

110. Mastrangelo G, Fadda E, Marzia V (1996) Polycyclic aromatic hydrocarbons and cancer in man. Environ Health Perspect 104: 1166-1170.

111. Sram RJ, Binkova B, Rossner P, Rubes J, Topinka J, et al. (1999) Adverse reproductive outcomes from exposure to environmental mutagens. Mutat Res 428: 203-215.

112. Phale PS, Basu A, Majhi PD, Deveryshetty J, Vamsee-Krishna C, et al. (2007) Metabolic diversity in bacterial degradation of aromatic compounds. OMICS 11: 252-279.

113. De Rosa CT, Nickle R, Faroon O, Jones DE (2003) The impact of toxicology on public health policy and service: an update. Toxicol Ind Health 19: 115-124.

114. Bruins MR, Kapil S, Oehme FW (2000) Microbial resistance to metals in the environment. Ecotoxicol Environ Saf 45: 198-207.

115. Silver S (1992) Plasmid-determined metal resistance mechanisms: range and overview. Plasmid 27: 1-3.

116. Rouch DA, Lee BT, Morby AP (1995) Understanding cellular responses to toxic agents: a model for mechanism-choice in bacterial metal resistance. J Ind Microbiol 14: 132-141.

117. Messens J, Silver S (2006) Arsenate reduction: thiol cascade chemistry with convergent evolution. J Mol Biol 362: 1-17.

118. Nies D, Mergeay M, Friedrich B, Schlegel HG (1987) Cloning of plasmid genes encoding resistance to cadmium, zinc, and cobalt in Alcaligenes eutrophus CH34. J Bacteriol 169: 4865-4868.

119. Nies DH (1992) Resistance to cadmium, cobalt, zinc, and nickel in microbes. Plasmid 27: 17-28.

120. Nascimento AM, Chartone-Souza E (2003) Operon mer: bacterial resistance to mercury and potential for bioremediation of contaminated environments. Genet Mol Res 2: 92-101.

121. Saunders JR (1984) Genetics and evolution of antibiotic resistance. Br Med Bull 40: 54-60.

122. Hall RM, Collis CM (1995) Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. Mol Microbiol 15: 593-600.

123. Jacoby GA, Munoz-Price LS (2005) The new beta-lactamases. N Engl J Med 352: 380-391.

124. Zscheck KK, Murray BE (1993) Genes involved in the regulation of beta-lactamase production in enterococci and staphylococci. Antimicrob Agents Chemother 37: 1966-1970.

125. Chopra I, Roberts M (2001) Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. Microbiol Mol Biol Rev 65: 232-260 ; second page, table of contents.

126. Connell SR, Tracz DM, Nierhaus KH, Taylor DE (2003) Ribosomal protection proteins and their mechanism of tetracycline resistance. Antimicrob Agents Chemother 47: 3675-3681.

127. Cohen SP, McMurry LM, Hooper DC, Wolfson JS, Levy SB (1989) Cross-resistance to fluoroquinolones in multiple-antibiotic-resistant (Mar) Escherichia coli selected by tetracycline or chloramphenicol: decreased drug accumulation associated with membrane changes in addition to OmpF reduction. Antimicrob Agents Chemother 33: 1318-1325.

128. George AM, Levy SB (1983) Amplifiable resistance to tetracycline, chloramphenicol, and other antibiotics in Escherichia coli: involvement of a non-plasmid-determined efflux of tetracycline. J Bacteriol 155: 531-540.

129. Randall LP, Woodward MJ (2002) The multiple antibiotic resistance (mar) locus and its significance. Res Vet Sci 72: 87-93.

130. Ariza RR, Cohen SP, Bachhawat N, Levy SB, Demple B (1994) Repressor mutations in the marRAB operon that activate oxidative stress genes and multiple antibiotic resistance in Escherichia coli. J Bacteriol 176: 143-148.

131. Greenberg JT, Chou JH, Monach PA, Demple B (1991) Activation of oxidative stress genes by mutations at the soxQ/cfxB/marA locus of Escherichia coli. J Bacteriol 173: 4433-4439.

132. Asako H, Nakajima H, Kobayashi K, Kobayashi M, Aono R (1997) Organic solvent tolerance and antibiotic resistance increased by overexpression of marA in Escherichia coli. Appl Environ Microbiol 63: 1428-1433.

133. Randall LP, Woodward MJ (2001) Multiple antibiotic resistance (mar) locus in Salmonella enterica serovar typhimurium DT104. Appl Environ Microbiol 67: 1190-1197.

134. Moken MC, McMurry LM, Levy SB (1997) Selection of multiple-antibiotic-resistant (mar) mutants of Escherichia coli by using the disinfectant pine oil: roles of the mar and acrAB loci. Antimicrob Agents Chemother 41: 2770-2772.

135. Revsbech NP, Thamdrup B, Dalsgaard T, Canfield DE (2011) Construction of STOX oxygen sensors and their application for determination of O2 concentrations in oxygen minimum zones. Methods Enzymol 486: 325-341.

136. Li XZ, Nikaido H (2004) Efflux-mediated drug resistance in bacteria. Drugs 64: 159-204.

137. Piddock LJ (2006) Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. Clin Microbiol Rev 19: 382-402.

138. Poole K (2007) Efflux pumps as antimicrobial resistance mechanisms. Ann Med 39: 162-176.

139. Martinez JL, Sanchez MB, Martinez-Solano L, Hernandez A, Garmendia L, et al. (2009) Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems. FEMS Microbiol Rev 33: 430-449.

140. Silver S, Phung LT (1996) Bacterial heavy metal resistance: new surprises. Annu Rev Microbiol 50: 753-789.

141. Silver S, Phung le T (2005) A bacterial view of the periodic table: genes and proteins for toxic inorganic ions. J Ind Microbiol Biotechnol 32: 587-605.

142. Ramos JL, Duque E, Gallegos MT, Godoy P, Ramos-Gonzalez MI, et al. (2002) Mechanisms of solvent tolerance in gram-negative bacteria. Annu Rev Microbiol 56: 743-768.

143. Pumbwe L, Skilbeck CA, Wexler HM (2007) Induction of multiple antibiotic resistance in Bacteroides fragilis by benzene and benzene-derived active compounds of commonly used analgesics, antiseptics and cleaning agents. J Antimicrob Chemother 60: 1288-1297.

144. Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. Proc Natl Acad Sci U S A 101: 3160-3165.

145. Fromme P, Grotjohann I (2008) Structure of Photosystems I and II. Results Probl Cell Differ 45: 33-72.

146. Hayes JM, Waldbauer JR (2006) The carbon cycle and associated redox processes through time. Philos Trans R Soc Lond B Biol Sci 361: 931-950.

147. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281: 237-240.

148. Tolbert NE (1997) The C2 Oxidative Photosynthetic Carbon Cycle. Annu Rev Plant Physiol Plant Mol Biol 48: 1-25.

149. Fuhrman JA, Schwalbach MS, Stingl U (2008) Proteorhodopsins: an array of physiological roles? Nat Rev Microbiol 6: 488-494.

150. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289: 1902-1906.

151. Beynon J, Ally A, Cannon M, Cannon F, Jacobson M, et al. (1987) Comparative organization of nitrogen fixation-specific genes from Azotobacter vinelandii and Klebsiella pneumoniae: DNA sequence of the nifUSV genes. J Bacteriol 169: 4024-4029.

152. Postgate JR (1970) Biological nitrogen fixation. Nature 226: 25-27.

153. Halbleib CM, Ludden PW (2000) Regulation of biological nitrogen fixation. J Nutr 130: 1081-1084.

154. Kneip C, Lockhart P, Voss C, Maier UG (2007) Nitrogen fixation in eukaryotes--new models for symbiosis. BMC Evol Biol 7: 55.

155. Tavares P, Pereira AS, Moura JJ, Moura I (2006) Metalloenzymes of the denitrification pathway. J Inorg Biochem 100: 2087-2100.

156. Zumft WG (1997) Cell biology and molecular basis of denitrification. Microbiol Mol Biol Rev 61: 533-616.

157. Lipschultz F, Zafiriou OC, Wofsy SC, McElroy MB, Valois FW, et al. (1981) Production of NO and N2O by soil nitrifying bacteria. Nature 294: 641-643.

158. Klotz MG, Stein LY (2008) Nitrifier genomics and evolution of the nitrogen cycle. FEMS Microbiol Lett 278: 146-156.

159. DeLong EF, Karl DM (2005) Genomic perspectives in microbial oceanography. Nature 437: 336-342.

160. Sunda W, Kieber DJ, Kiene RP, Huntsman S (2002) An antioxidant function for DMSP and DMS in marine algae. Nature 418: 317-320.

161. Charlson RJ, Lovelock JE, Andreae MO, Warren SG (1987) Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. Nature 326: 655-661.

162. Johnston AW, Todd JD, Sun L, Nikolaidou-Katsaridou MN, Curson AR, et al. (2008) Molecular diversity of bacterial production of the climate-changing gas, dimethyl sulphide, a molecule that impinges on local and global symbioses. J Exp Bot 59: 1059-1067.

163. Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, et al. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature 438: 82-85.

164. Biebl H, Allgaier M, Tindall BJ, Koblizek M, Lunsdorf H, et al. (2005) Dinoroseobacter shibae gen. nov., sp. nov., a new aerobic phototrophic bacterium isolated from dinoflagellates. Int J Syst Evol Microbiol 55: 1089-1096.

165. Dickschat JS, Zell C, Brock NL (2010) Pathways and substrate specificity of DMSP catabolism in marine bacteria of the Roseobacter clade. Chembiochem 11: 417-425.

166. Tripp HJ, Kitner JB, Schwalbach MS, Dacey JW, Wilhelm LJ, et al. (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature 452: 741-744.

167. Clokie MR, Mann NH (2006) Marine cyanophages and light. Environ Microbiol 8: 2074-2082.

168. Buchan A, Gonzalez JM, Moran MA (2005) Overview of the marine roseobacter lineage. Appl Environ Microbiol 71: 5665-5677.

169. Schleheck D, Knepper TP, Eichhorn P, Cook AM (2007) Parvibaculum lavamentivorans DS-1T degrades centrally substituted congeners of commercial linear alkylbenzenesulfonate to sulfophenyl carboxylates and sulfophenyl dicarboxylates. Appl Environ Microbiol 73: 4725-4732.

170. Allers E, Niesner C, Wild C, Pernthaler J (2008) Microbes enriched in seawater after addition of coral mucus. Appl Environ Microbiol 74: 3274-3278.

171. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496-503.

172. Grzymski JJ, Murray AE, Campbell BJ, Kaplarevic M, Gao GR, et al. (2008) Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. Proc Natl Acad Sci U S A 105: 17516-17521.

173. Yakimov MM, Golyshin PN, Lang S, Moore ER, Abraham WR, et al. (1998) Alcanivorax borkumensis gen. nov., sp. nov., a new, hydrocarbon-degrading and surfactant-producing marine bacterium. Int J Syst Bacteriol 48 Pt 2: 339-348.

174. Golyshin PN, Martins Dos Santos VA, Kaiser O, Ferrer M, Sabirova YS, et al. (2003) Genome sequence completed of Alcanivorax borkumensis, a hydrocarbon-degrading bacterium that plays a global role in oil removal from marine systems. J Biotechnol 106: 215-220.

175. Hara A, Syutsubo K, Harayama S (2003) Alcanivorax which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation. Environ Microbiol 5: 746-753.

176. Huu NB, Denner EB, Ha DT, Wanner G, Stan-Lotter H (1999) Marinobacter aquaeolei sp. nov., a halophilic bacterium isolated from a Vietnamese oil-producing well. Int J Syst Bacteriol 49 Pt 2: 367-375.

177. Schleheck D, Tindall BJ, Rossello-Mora R, Cook AM (2004) Parvibaculum lavamentivorans gen. nov., sp. nov., a novel heterotroph that initiates catabolism of linear alkylbenzenesulfonate. Int J Syst Evol Microbiol 54: 1489-1497.

178. Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, et al. (2000) The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. Nature 403: 665-668.

179. Badger JH, Hoover TR, Brun YV, Weiner RM, Laub MT, et al. (2006) Comparative genomic evidence for a close relationship between the dimorphic prosthecate bacteria Hyphomonas neptunium and Caulobacter crescentus. J Bacteriol 188: 6841-6850.

180. Ryder C, Byrd M, Wozniak DJ (2007) Role of polysaccharides in Pseudomonas aeruginosa biofilm development. Curr Opin Microbiol 10: 644-648.

181. Jung YH, Yi JY, Jung HJ, Lee YK, Lee HK, et al. (2010) Overexpression of cold shock protein A of Psychromonas arctica KOPRI 22215 confers cold-resistance. Protein J 29: 136-142.

182. Phadtare S (2004) Recent developments in bacterial cold-shock response. Curr Issues Mol Biol 6: 125-136.

183. Bohin JP (2000) Osmoregulated periplasmic glucans in Proteobacteria. FEMS Microbiol Lett 186: 11-19.

184. Lee S, Cho E, Jung S (2009) Periplasmic glucans isolated from Proteobacteria. BMB Rep 42: 769-775.

185. Feller G, Gerday C (2003) Psychrophilic enzymes: hot topics in cold adaptation. Nat Rev Microbiol 1: 200-208.

186. Wilmes B, Hartung A, Lalk M, Liebeke M, Schweder T, et al. (2010) Fed-batch process for the psychrotolerant marine bacterium Pseudoalteromonas haloplanktis. Microb Cell Fact 9: 72.

187. Birolo L, Tutino ML, Fontanella B, Gerday C, Mainolfi K, et al. (2000) Aspartate aminotransferase from the Antarctic bacterium Pseudoalteromonas haloplanktis TAC 125. Cloning, expression, properties, and molecular modelling. Eur J Biochem 267: 2790-2802.

188. Tutino ML, Duilio A, Parrilli R, Remaut E, Sannia G, et al. (2001) A novel replication element from an Antarctic plasmid as a tool for the expression of proteins at low temperature. Extremophiles 5: 257-264.

189. Medigue C, Krin E, Pascal G, Barbe V, Bernsel A, et al. (2005) Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125. Genome Res 15: 1325-1335.

190. Papa R, Parrilli E, Sannia G (2009) Engineered marine Antarctic bacterium Pseudoalteromonas haloplanktis TAC125: a promising micro-organism for the bioremediation of aromatic compounds. J Appl Microbiol 106: 49-56.

191. Xie G, Bruce DC, Challacombe JF, Chertkov O, Detter JC, et al. (2007) Genome sequence of the cellulolytic gliding bacterium Cytophaga hutchinsonii. Appl Environ Microbiol 73: 3536-3546.

192. Borella P, Guerrieri E, Marchesi I, Bondi M, Messi P (2005) Water ecology of Legionella and protozoan: environmental and public health perspectives. Biotechnol Annu Rev 11: 355-380.

193. Atlas RM (1999) Legionella: from environmental habitats to disease pathology, detection and control. Environ Microbiol 1: 283-293.

194. Yoshida T, Sakamoto T (2009) Water-stress induced trehalose accumulation and control of trehalase in the cyanobacterium Nostoc punctiforme IAM M-15. J Gen Appl Microbiol 55: 135-145.

195. Potts M (1994) Desiccation tolerance of prokaryotes. Microbiol Rev 58: 755-805.

196. Bozal N, Montes MJ, Tudela E, Guinea J (2003) Characterization of several Psychrobacter strains isolated from Antarctic environments and description of Psychrobacter luti sp. nov. and Psychrobacter fozii sp. nov. Int J Syst Evol Microbiol 53: 1093-1100.

197. Maruyama A, Honda D, Yamamoto H, Kitamura K, Higashihara T (2000) Phylogenetic analysis of psychrophilic bacteria isolated from the Japan Trench, including a description of the deep-sea species Psychrobacter pacificensis sp. nov. Int J Syst Evol Microbiol 50 Pt 2: 835-846.

198. Romanenko LA, Schumann P, Rohde M, Lysenko AM, Mikhailov VV, et al. (2002) Psychrobacter submarinus sp. nov. and Psychrobacter marincola sp. nov., psychrophilic halophiles from marine environments. Int J Syst Evol Microbiol 52: 1291-1297.

199. Bakermans C, Ayala-del-Rio HL, Ponder MA, Vishnivetskaya T, Gilichinsky D, et al. (2006) Psychrobacter cryohalolentis sp. nov. and Psychrobacter arcticus sp. nov., isolated from Siberian permafrost. Int J Syst Evol Microbiol 56: 1285-1291.

200. Amato P, Christner BC (2009) Energy metabolism response to low-temperature and frozen conditions in Psychrobacter cryohalolentis. Appl Environ Microbiol 75: 711-718.

201. Ayala-del-Rio HL, Chain PS, Grzymski JJ, Ponder MA, Ivanova N, et al. (2010) The genome sequence of Psychrobacter arcticus 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. Appl Environ Microbiol 76: 2304-2312.

202. Breezee J, Cady N, Staley JT (2004) Subfreezing growth of the sea ice bacterium "Psychromonas ingrahamii". Microb Ecol 47: 300-304.

203. Riley M, Staley JT, Danchin A, Wang TZ, Brettin TS, et al. (2008) Genomics of an extreme psychrophile, Psychromonas ingrahamii. BMC Genomics 9: 210.

204. Jakosky BM, Nealson KH, Bakermans C, Ley RE, Mellon MT (2003) Subfreezing activity of microorganisms and the potential habitability of Mars' polar regions. Astrobiology 3: 343-350.

205. Torsvik V, Ovreas L, Thingstad TF (2002) Prokaryotic diversity--magnitude, dynamics, and controlling factors. Science 296: 1064-1066.

206. Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309: 1387-1390.

207. Legendre G, Fay F, Linossier I, Vallee-Rehel K (2011) Evaluation of antibacterial activity against Salmonella Enteritidis. J Microbiol 49: 349-354.

208. Zhou NY, Fuenmayor SL, Williams PA (2001) nag genes of Ralstonia (formerly Pseudomonas) sp. strain U2 encoding enzymes for gentisate catabolism. J Bacteriol 183: 700-708.

209. Kolosova N, Gorenstein N, Kish CM, Dudareva N (2001) Regulation of circadian methyl benzoate emission in diurnally and nocturnally emitting plants. Plant Cell 13: 2333-2347.

210. Kang KH, Jang SK, Kim BK, Park MK (1994) Antibacterial phenylpropanoid glycosides from Paulownia tomentosa Steud. Arch Pharm Res 17: 470-475.

211. Agarwal N, Bishai WR (2009) cAMP signaling in Mycobacterium tuberculosis. Indian J Exp Biol 47: 393-400.

212. Akhter Y, Yellaboina S, Farhana A, Ranjan A, Ahmed N, et al. (2008) Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis. Gene 407: 148-158.

213. Crane BR, Sudhamsu J, Patel BA (2010) Bacterial nitric oxide synthases. Annu Rev Biochem 79: 445-470.

214. Duine JA (1999) The PQQ story. J Biosci Bioeng 88: 231-236.

215. Kay CW, Mennenga B, Gorisch H, Bittl R (2006) Structure of the pyrroloquinoline quinone radical in quinoprotein ethanol dehydrogenase. J Biol Chem 281: 1470-1476.

216. Derrien M, Vaughan EE, Plugge CM, de Vos WM (2004) Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. Int J Syst Evol Microbiol 54: 1469-1476.

217. Derrien M, Collado MC, Ben-Amor K, Salminen S, de Vos WM (2008) The Mucin degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract. Appl Environ Microbiol 74: 1646-1648.

218. Mergeay M, Houba C, Gerits J (1978) Extrachromosomal inheritance controlling resistance to cadmium, cobalt, copper and zinc ions: evidence from curing in a Pseudomonas [proceedings]. Arch Int Physiol Biochim 86: 440-442.

219. Janssen PJ, Van Houdt R, Moors H, Monsieurs P, Morin N, et al. (2010) The complete genome sequence of Cupriavidus metallidurans strain CH34, a master survivalist in harsh and anthropogenic environments. PLoS One 5: e10433.

220. Spaink HP (2000) Root nodulation and infection factors produced by rhizobial bacteria. Annu Rev Microbiol 54: 257-288.

221. Loh J, Stacey G (2003) Nodulation gene regulation in Bradyrhizobium japonicum: a unique integration of global regulatory circuits. Appl Environ Microbiol 69: 10-17.

222. Gonzalez JE, Marketon MM (2003) Quorum sensing in nitrogen-fixing rhizobia. Microbiol Mol Biol Rev 67: 574-592.

223. Benson DR, Silvester WB (1993) Biology of Frankia strains, actinomycete symbionts of actinorhizal plants. Microbiol Rev 57: 293-319.

224. Parsons R, Sunley RJ (2001) Nitrogen nutrition and the role of root-shoot nitrogen signalling particularly in symbiotic systems. J Exp Bot 52: 435-443.

225. Heusser LE, Shackleton NJ (1979) Direct marine-continental correlation: 150,000-year oxygen isotope--pollen record from the north pacific. Science 204: 837-839.

226. Katoch VM (2004) Infections due to non-tuberculous mycobacteria (NTM). Indian J Med Res 120: 290-304.

227. Diaz E (2004) Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility. Int Microbiol 7: 173-180.

228. Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, et al. (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. Nat Biotechnol 22: 55-61.

229. Harrison FH, Harwood CS (2005) The pimFABCDE operon from Rhodopseudomonas palustris mediates dicarboxylic acid degradation and participates in anaerobic benzoate degradation. Microbiology 151: 727-736.

230. Schlesner H, Rensmann C, Tindall BJ, Gade D, Rabus R, et al. (2004) Taxonomic heterogeneity within the Planctomycetales as derived by DNA-DNA hybridization, description of Rhodopirellula

baltica gen. nov., sp. nov., transfer of Pirellula marina to the genus Blastopirellula gen. nov. as Blastopirellula marina comb. nov. and emended description of the genus Pirellula. Int J Syst Evol Microbiol 54: 1567-1580.

231. Ward NL, Challacombe JF, Janssen PH, Henrissat B, Coutinho PM, et al. (2009) Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils. Appl Environ Microbiol 75: 2046-2056.

232. Balkwill DL, Drake GR, Reeves RH, Fredrickson JK, White DC, et al. (1997) Taxonomic study of aromatic-degrading bacteria from deep-terrestrial-subsurface sediments and description of Sphingomonas aromaticivorans sp. nov., Sphingomonas subterranea sp. nov., and Sphingomonas stygia sp. nov. Int J Syst Bacteriol 47: 191-201.

233. Mitsui A, Kumazawa S, Takahashi A, Ikemoto H, Cao S, et al. (1986) Strategy by which nitrogen-fixing unicellular cyanobacteria grow photoautotrophically. Nature 323: 720-722.

234. Mackey SR, Golden SS (2007) Winding up the cyanobacterial circadian clock. Trends Microbiol 15: 381-388.

235. Peterson RB, Wolk CP (1978) High recovery of nitrogenase activity and of Fe-labeled nitrogenase in heterocysts isolated from Anabaena variabilis. Proc Natl Acad Sci U S A 75: 6271-6275.

236. Happe T, Schutz K, Bohme H (2000) Transcriptional and mutational analysis of the uptake hydrogenase of the filamentous cyanobacterium Anabaena variabilis ATCC 29413. J Bacteriol 182: 1624-1631.

237. Borodin VB, Tsygankov AA, Rao KK, Hall DO (2000) Hydrogen production by Anabaena variabilis PK84 under simulated outdoor conditions. Biotechnol Bioeng 69: 478-485.

238. Holtzendorff J, Partensky F, Mella D, Lennon JF, Hess WR, et al. (2008) Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium Prochlorococcus marinus PCC 9511. J Biol Rhythms 23: 187-199.

239. O'Brien JM, Wolkin RH, Moench TT, Morgan JB, Zeikus JG (1984) Association of hydrogen metabolism with unitrophic or mixotrophic growth of Methanosarcina barkeri on carbon monoxide. J Bacteriol 158: 373-375.

240. Zeikus JG, Kerby R, Krzycki JA (1985) Single-carbon chemistry of acetogenic and methanogenic bacteria. Science 227: 1167-1173.

241. Rother M, Metcalf WW (2004) Anaerobic growth of Methanosarcina acetivorans C2A on carbon monoxide: an unusual way of life for a methanogenic archaeon. Proc Natl Acad Sci U S A 101: 16929-16934.

242. Huber R, Kurr M, Vacharaksa HW, .Stetter KO (1989) A novel group of abyssal methanogenic archaebacteria (Methanopyrus) growing at 110 [deg]C. 342: 833-834.

243. Breitung J, Borner G, Scholz S, Linder D, Stetter KO, et al. (1992) Salt dependence, kinetic properties and catalytic mechanism of N-formylmethanofuran:tetrahydromethanopterin formyltransferase from the extreme thermophile Methanopyrus kandleri. Eur J Biochem 210: 971-981.

244. Slesarev AI, Lake JA, Stetter KO, Gellert M, Kozyavkin SA (1994) Purification and characterization of DNA topoisomerase V. An enzyme from the hyperthermophilic prokaryote Methanopyrus kandleri that resembles eukaryotic topoisomerase I. J Biol Chem 269: 3295-3303.

245. Pierson BK, Castenholz RW (1974) A phototrophic gliding filamentous bacterium of hot springs, Chloroflexus aurantiacus, gen. and sp. nov. Arch Microbiol 100: 5-24.

246. Sirevag R, Castenholz R (1979) Aspects of carbon metabolism in Chloroflexus. Archives of Microbiology 120: 151-153.

247. Falb M, Pfeiffer F, Palm P, Rodewald K, Hickmann V, et al. (2005) Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis. Genome Res 15: 1336-1343.

248. Gonzalez O, Oberwinkler T, Mansueto L, Pfeiffer F, Mendoza E, et al. (2010) Characterization of growth and metabolism of the haloalkaliphile Natronomonas pharaonis. PLoS Comput Biol 6: e1000799.

249. Vreeland RH, Hochstein LH (1993) Biology of halophilic bacteria. CRC Press, Boca Raton, FL.

250. Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, et al. (2004) Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. Genome Res 14: 2221-2234.

251. Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, et al. (2000) Genome sequence of Halobacterium species NRC-1. Proc Natl Acad Sci U S A 97: 12176-12181.

252. Dassarma S, Kennedy SP, Berquist B, Victor Ng W, Baliga NS, et al. (2001) Genomic perspective on the photobiology of Halobacterium species NRC-1, a phototrophic, phototactic, and UV-tolerant haloarchaeon. Photosynth Res 70: 3-17.

253. Ginzburg M, Sachs L, Ginzburg BZ (1970) Ion metabolism in a Halobacterium. I. Influence of age of culture on intracellular concentrations. J Gen Physiol 55: 187-207.

254. Oppermann H, Levinson AD, Varmus HE (1981) The structure and protein kinase activity of proteins encoded by nonconditional mutants and back mutants in the sec gene of avian sarcoma virus. Virology 108: 47-70.

255. Haveman SA, DiDonato RJ, Jr., Villanueva L, Shelobolina ES, Postier BL, et al. (2008) Genome-wide gene expression patterns and growth requirements suggest that Pelobacter carbinolicus reduces Fe(III) indirectly via sulfide production. Appl Environ Microbiol 74: 4277-4284.

256. Knoblauch C, Sahm K, Jorgensen BB (1999) Psychrophilic sulfate-reducing bacteria isolated from permanently cold arctic marine sediments: description of Desulfofrigus oceanense gen. nov., sp. nov., Desulfofrigus fragile sp. nov., Desulfofaba gelida gen. nov., sp. nov., Desulfotalea psychrophila gen. nov., sp. nov. and Desulfotalea arctica sp. nov. Int J Syst Bacteriol 49 Pt 4: 1631-1643.

257. Rabus R, Ruepp A, Frickey T, Rattei T, Fartmann B, et al. (2004) The genome of Desulfotalea psychrophila, a sulfate-reducing bacterium from permanently cold Arctic sediments. Environ Microbiol 6: 887-902.

258. Frostl JM, Overmann J (1998) Physiology and tactic response of the phototrophic consortium "Chlorochromatium aggregatum". Arch Microbiol 169: 129-135.

259. Overmann J (2010) The phototrophic consortium "Chlorochromatium aggregatum" - a model for bacterial heterologous multicellularity. Adv Exp Med Biol 675: 15-29.

260. Kanzler BE, Pfannes KR, Vogl K, Overmann J (2005) Molecular characterization of the nonphotosynthetic partner bacterium in the consortium "Chlorochromatium aggregatum". Appl Environ Microbiol 71: 7434-7441.

261. Ferreira AC, Nobre MF, Rainey FA, Silva MT, Wait R, et al. (1997) Deinococcus geothermalis sp. nov. and Deinococcus murrayi sp. nov., two extremely radiation-resistant and slightly thermophilic species from hot springs. Int J Syst Bacteriol 47: 939-947.

262. Brim H, Venkateswaran A, Kostandarithes HM, Fredrickson JK, Daly MJ (2003) Engineering Deinococcus geothermalis for bioremediation of high-temperature radioactive waste environments. Appl Environ Microbiol 69: 4575-4582.

263. Makarova KS, Omelchenko MV, Gaidamakova EK, Matrosova VY, Vasilenko A, et al. (2007) Deinococcus geothermalis: the pool of extreme radiation resistance genes shrinks. PLoS One 2: e955.

264. Higgins CF (1992) ABC transporters: from microorganisms to man. Annu Rev Cell Biol 8: 67-113.

265. Kim W, Whitman WB (1999) Isolation of acetate auxotrophs of the methane-producing archaeon Methanococcus maripaludis by random insertional mutagenesis. Genetics 152: 1429-1437.

266. Smart JP, Cliff MJ, Kelly DJ (2009) A role for tungsten in the biology of Campylobacter jejuni: tungstate stimulates formate dehydrogenase activity and is transported via an ultra-high affinity ABC system distinct from the molybdate transporter. Mol Microbiol 74: 742-757.

267. Gerber S, Comellas-Bigler M, Goetz BA, Locher KP (2008) Structural basis of trans-inhibition in a molybdate/tungstate ABC transporter. Science 321: 246-250.

268. Bevers LE, Hagedoorn PL, Krijger GC, Hagen WR (2006) Tungsten transport protein A (WtpA) in Pyrococcus furiosus: the first member of a new class of tungstate and molybdate transporters. J Bacteriol 188: 6498-6505.

269. Sevcenco AM, Bevers LE, Pinkse MW, Krijger GC, Wolterbeek HT, et al. (2010) Molybdenum incorporation in tungsten aldehyde oxidoreductase enzymes from Pyrococcus furiosus. J Bacteriol 192: 4143-4152.

270. Oelgeschlager E, Rother M (2008) Carbon monoxide-dependent energy metabolism in anaerobic bacteria and archaea. Arch Microbiol 190: 257-269.

271. Bonam D, Lehman L, Roberts GP, Ludden PW (1989) Regulation of carbon monoxide dehydrogenase and hydrogenase in Rhodospirillum rubrum: effects of CO and oxygen on synthesis and activity. J Bacteriol 171: 3102-3107.

272. Kerby RL, Ludden PW, Roberts GP (1995) Carbon monoxide-dependent growth of Rhodospirillum rubrum. J Bacteriol 177: 2241-2244.

273. Balch WE, Fox GE, Magrum LJ, Woese CR, Wolfe RS (1979) Methanogens: reevaluation of a unique biological group. Microbiol Rev 43: 260-296.

274. Hippe H, Caspari D, Fiebig K, Gottschalk G (1979) Utilization of trimethylamine and other N-methyl compounds for growth and methane formation by Methanosarcina barkeri. Proc Natl Acad Sci U S A 76: 494-498.

275. Kiene RP, Oremland RS, Catena A, Miller LG, Capone DG (1986) Metabolism of reduced methylated sulfur compounds in anaerobic sediments and by a pure culture of an estuarine methanogen. Appl Environ Microbiol 52: 1037-1045.

276. Hao B, Gong W, Ferguson TK, James CM, Krzycki JA, et al. (2002) A new UAG-encoded residue in the structure of a methanogen methyltransferase. Science 296: 1462-1466.

277. Srinivasan G, James CM, Krzycki JA (2002) Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. Science 296: 1459-1462.

278. Krzycki JA (2004) Function of genetically encoded pyrrolysine in corrinoid-dependent methylamine methyltransferases. Curr Opin Chem Biol 8: 484-491.

279. Hecker M, Volker U (1998) Non-specific, general and multiple stress resistance of growth-restricted Bacillus subtilis cells by the expression of the sigmaB regulon. Mol Microbiol 29: 1129-1136.

280. Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol 24: 1263-1269.

281. Wang Q, Shao Y, Huong VT, Park WJ, Park JM, et al. (2008) Fine-scale population structure of Accumulibacter phosphatis in enhanced biological phosphorus removal sludge. J Microbiol Biotechnol 18: 1290-1297.

282. Kovacs AT, Rakhely G, Balogh J, Maroti G, Fulop A, et al. (2005) Anaerobic regulation of hydrogenase transcription in different bacteria. Biochem Soc Trans 33: 36-38.

283. Lenz O, Ludwig M, Schubert T, Burstel I, Ganskow S, et al. (2010) H2 conversion in the presence of O2 as performed by the membrane-bound [NiFe]-hydrogenase of Ralstonia eutropha. Chemphyschem 11: 1107-1119.

284. Patankar AV, Gonzalez JE (2009) Orphan LuxR regulators of quorum sensing. FEMS Microbiol Rev 33: 739-756.

285. Barnes MR, Duetz WA, Williams PA (1997) A 3-(3-hydroxyphenyl)propionic acid catabolic pathway in Rhodococcus globerulus PWD1: cloning and characterization of the hpp operon. J Bacteriol 179: 6145-6153.

286. Dagley S, Chapman PJ, Gibson DT (1965) The metabolism of beta-phenylpropionic acid by an Achromobacter. Biochem J 97: 643-650.

287. Diaz E, Ferrandez A, Garcia JL (1998) Characterization of the hca cluster encoding the dioxygenolytic pathway for initial catabolism of 3-phenylpropionic acid in Escherichia coli K-12. J Bacteriol 180: 2915-2923.

288. Coates JD, Chakraborty R, Lack JG, O'Connor SM, Cole KA, et al. (2001) Anaerobic benzene oxidation coupled to nitrate reduction in pure culture by two strains of Dechloromonas. Nature 411: 1039-1043.

289. Rabus R, Kube M, Heider J, Beck A, Heitmann K, et al. (2005) The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. Arch Microbiol 183: 27-36.

290. Vial L, Groleau MC, Dekimpe V, Deziel E (2007) Burkholderia diversity and versatility: an inventory of the extracellular products. J Microbiol Biotechnol 17: 1407-1429.

291. Gilad J (2007) Burkholderia mallei and Burkholderia pseudomallei: the causative micro-organisms of glanders and melioidosis. Recent Pat Antiinfect Drug Discov 2: 233-241.

292. Genin S, Boucher C (2004) Lessons learned from the genome analysis of ralstonia solanacearum. Annu Rev Phytopathol 42: 107-134.

293. Mergeay M, Monchy S, Vallaeys T, Auquier V, Benotmane A, et al. (2003) Ralstonia metallidurans, a bacterium specifically adapted to toxic metals: towards a catalogue of metal-responsive genes. FEMS Microbiol Rev 27: 385-410.

294. Pohlmann A, Fricke WF, Reinecke F, Kusian B, Liesegang H, et al. (2006) Genome sequence of the bioplastic-producing "Knallgas" bacterium Ralstonia eutropha H16. Nat Biotechnol 24: 1257-1262.

295. Finneran KT, Johnsen CV, Lovley DR (2003) Rhodoferax ferrireducens sp. nov., a psychrotolerant, facultatively anaerobic bacterium that oxidizes acetate with the reduction of Fe(III). Int J Syst Evol Microbiol 53: 669-673.

296. Risso C, Sun J, Zhuang K, Mahadevan R, DeBoy R, et al. (2009) Genome-scale comparison and constraint-based metabolic reconstruction of the facultative anaerobic Fe(III)-reducer Rhodoferax ferrireducens. BMC Genomics 10: 447.

297. Crossman L, Dow JM (2004) Biofilm formation and dispersal in Xanthomonas campestris. Microbes Infect 6: 623-629.

298. Creczynski-Pasa TB, Antonio RV (2004) Energetic metabolism of Chromobacterium violaceum. Genet Mol Res 3: 162-166.

299. Duran N, Menck CF (2001) Chromobacterium violaceum: a review of pharmacological and industiral perspectives. Crit Rev Microbiol 27: 201-222.

300. Carepo MS, Azevedo JS, Porto JI, Bentes-Sousa AR, Batista Jda S, et al. (2004) Identification of Chromobacterium violaceum genes with potential biotechnological application in environmental detoxification. Genet Mol Res 3: 181-194.

301. Vijayan AP, Anand MR, Remesh P (2009) Chromobacterium violaceum sepsis in an infant. Indian Pediatr 46: 721-722.

302. Ajithdoss DK, Porter BF, Calise DV, Libal MC, Edwards JF (2009) Septicemia in a neonatal calf associated with Chromobacterium violaceum. Vet Pathol 46: 71-74.

303. Watson SW, Waterbury JB (1971) Characteristics of two marine nitrite oxidizing bacteria, Nitrospina gracilis nov. gen. nov. sp. and Nitrococcus mobilis nov. gen. nov. sp. . Archives of Microbiology 77: 203-230.

304. Murray RG, Watson SW (1965) Structure of Nitrosocystis Oceanus and Comparison with Nitrosomonas and Nitrobacter. J Bacteriol 89: 1594-1609.

305. Klotz MG, Arp DJ, Chain PS, El-Sheikh AF, Hauser LJ, et al. (2006) Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium Nitrosococcus oceani ATCC 19707. Appl Environ Microbiol 72: 6299-6315.

306. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, et al. (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329: 52-56.

307. Potter MC (1911) Electrical Effects Accompanying the Decomposition of Organic Compounds. Proc R Soc Lond: 260-276.

308. Kim S, Singh P, Park J, Park S, Friedman A, et al. (2011) Genetic and molecular characterization of a blue light photoreceptor MGWC-1 in Magnaporth oryzae. Fungal Genet Biol 48: 400-407.

309. Vladimirov N, Sourjik V (2009) Chemotaxis: how bacteria use memory. Biol Chem 390: 1097-1104.

310. Henrichsen J (1972) Bacterial surface translocation: a survey and a classification. Bacteriol Rev 36: 478-503.

311. Spormann AM (1999) Gliding motility in bacteria: insights from studies of Myxococcus xanthus. Microbiol Mol Biol Rev 63: 621-641.

312. Bagramyan K, Trchounian A (2003) Structural and functional features of formate hydrogen lyase, an enzyme of mixed-acid fermentation from Escherichia coli. Biochemistry (Mosc) 68: 1159-1170.

313. Ragsdale SW (1991) Enzymology of the acetyl-CoA pathway of CO2 fixation. Crit Rev Biochem Mol Biol 26: 261-300.

314. Regnier P, Hajnsdorf E (2009) Poly(A)-assisted RNA decay and modulators of RNA stability. Prog Mol Biol Transl Sci 85: 137-185.

315. Bock A, Forchhammer K, Heider J, Leinfelder W, Sawers G, et al. (1991) Selenocysteine: the 21st amino acid. Mol Microbiol 5: 515-520.

316. Jacob C, Giles GI, Giles NM, Sies H (2003) Sulfur and selenium: the role of oxidation state in protein structure and function. Angew Chem Int Ed Engl 42: 4742-4758.

317. Arbogast S, Ferreiro A (2010) Selenoproteins and protection against oxidative stress: selenoprotein N as a novel player at the crossroads of redox signaling and calcium homeostasis. Antioxid Redox Signal 12: 893-904.

318. Pollock MR, Knox R, Gell PGH (1942) Bacterial Reduction of Tetrathionate. Nature: 94-94.

319. Oltmann L, van der Beek E, Stouthamer A (1975) Reduction of inorganic sulphur compounds by facultatively aerobic bacteria. Plant and Soil 43: 153-169.

320. Papavassiliou J, Samaraki-Lyberopoulou V, Piperakis G (1969) Production of tetrathionate reductase by Salmonella. Can J Microbiol 15: 238-240.

321. Tuttle JH (1980) Thiosulfate Oxidation and Tetrathionate Reduction by Intact Cells of Marine Pseudomonad Strain 16B. Appl Environ Microbiol 39: 1159-1166.

322. Kimura T, Nishioka H (1997) Intracellular generation of superoxide by copper sulphate in Escherichia coli. Mutat Res 389: 237-242.

323. Beswick PH, Hall GH, Hook AJ, Little K, McBrien DC, et al. (1976) Copper toxicity: evidence for the conversion of cupric to cuprous copper in vivo under anaerobic conditions. Chem Biol Interact 14: 347-356.

324. Outten FW, Huffman DL, Hale JA, O'Halloran TV (2001) The independent cue and cus systems confer copper tolerance during aerobic and anaerobic growth in Escherichia coli. J Biol Chem 276: 30670-30677.

325. Tottey S, Harvie DR, Robinson NJ (2005) Understanding how cells allocate metals using metal sensors and metallochaperones. Acc Chem Res 38: 775-783.

326. De la Cerda B, Castielli O, Duran RV, Navarro JA, Hervas M, et al. (2007) A proteomic approach to iron and copper homeostasis in cyanobacteria. Brief Funct Genomic Proteomic 6: 322-329.

327. Galindo CL, Gutierrez C, Jr., Chopra AK (2006) Potential involvement of galectin-3 and SNAP23 in Aeromonas hydrophila cytotoxic enterotoxin-induced host cell apoptosis. Microb Pathog 40: 56-68.

328. Gray SJ (1984) Aeromonas hydrophila in livestock: incidence, biochemical characteristics and antibiotic susceptibility. J Hyg (Lond) 92: 365-375.

329. Seshadri R, Joseph SW, Chopra AK, Sha J, Shaw J, et al. (2006) Genome sequence of Aeromonas hydrophila ATCC 7966T: jack of all trades. J Bacteriol 188: 8272-8282.

330. Pham CA, Jung SJ, Phung NT, Lee J, Chang IS, et al. (2003) A novel electrochemically active and Fe(III)-reducing bacterium phylogenetically related to Aeromonas hydrophila, isolated from a microbial fuel cell. FEMS Microbiol Lett 223: 129-134.

331. Smith EF, Townsend CO (1907) A Plant-Tumor of Bacterial Origin. Science 25: 671-673.

332. Tomlinson AD, Fuqua C (2009) Mechanisms and regulation of polar surface attachment in Agrobacterium tumefaciens. Curr Opin Microbiol 12: 708-714.

333. Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, et al. (2001) Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58. Science 294: 2323-2328.

334. Topping JF, Wei W, Clarke MC, Muskett P, Lindsey K (1995) Agrobacterium-mediated transformation of Arabidopsis thaliana. Application in T-DNA tagging. Methods Mol Biol 49: 63-76.

335. Newell CA (2000) Plant transformation technology. Developments and applications. Mol Biotechnol 16: 53-65.

336. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, et al. (2001) The genome of the natural genetic engineer Agrobacterium tumefaciens C58. Science 294: 2317-2323.

337. Muller RH, Jorks S, Kleinsteuber S, Babel W (1999) Comamonas acidovorans strain MC1: a new isolate capable of degrading the chiral herbicides dichlorprop and mecoprop and the herbicides 2,4-D and MCPA. Microbiol Res 154: 241-246.

338. Muller RH, Babel W (2004) Delftia acidovorans MC1 resists high herbicide concentrations--a study of nutristat growth on (RS)-2-(2,4-Dichlorophenoxy)propionate and 2,4-dichlorophenoxyacetate. Biosci Biotechnol Biochem 68: 622-630.

339. Engelhardt B, Martin-Simonet MT, Rott LS, Butcher EC, Michie SA (1998) Adhesion molecule phenotype of T lymphocytes in inflamed CNS. J Neuroimmunol 84: 92-104.

340. Baldermann C, Lupas A, Lubieniecki J, Engelhardt H (1998) The regulated outer membrane protein Omp21 from Comamonas acidovorans is identified as a member of a new family of eight-stranded beta-sheet proteins by its sequence and properties. J Bacteriol 180: 3741-3749.

341. Mathes A, Engelhardt H (1998) Nonlinear and asymmetric open channel characteristics of an ion-selective porin in planar membranes. Biophys J 75: 1255-1262.

342. Lovley DR, Coates JD (2000) Novel forms of anaerobic respiration of environmental relevance. Curr Opin Microbiol 3: 252-256.

343. Bond DR, Holmes DE, Tender LM, Lovley DR (2002) Electrode-reducing microorganisms that harvest energy from marine sediments. Science 295: 483-485.

344. Holmes DE, Bond DR, O'Neil RA, Reimers CE, Tender LR, et al. (2004) Microbial communities associated with electrodes harvesting electricity from a variety of aquatic sediments. Microb Ecol 48: 178-190.

345. Ishii S, Watanabe K, Yabuki S, Logan BE, Sekiguchi Y (2008) Comparison of electrode reduction activities of Geobacter sulfurreducens and an enriched consortium in an air-cathode microbial fuel cell. Appl Environ Microbiol 74: 7348-7355.

346. Liu Y, Harnisch F, Fricke K, Sietmann R, Schroder U (2008) Improvement of the anodic bioelectrocatalytic activity of mixed culture biofilms by a simple consecutive electrochemical selection procedure. Biosens Bioelectron 24: 1012-1017.

347. Van Eldere J (2003) Multicentre surveillance of Pseudomonas aeruginosa susceptibility patterns in nosocomial infections. J Antimicrob Chemother 51: 347-352.

348. Hassett DJ, Cuppoletti J, Trapnell B, Lymar SV, Rowe JJ, et al. (2002) Anaerobic metabolism and quorum sensing by Pseudomonas aeruginosa biofilms in chronically infected cystic fibrosis airways: rethinking antibiotic treatment strategies and drug targets. Adv Drug Deliv Rev 54: 1425-1443.

349. Rabaey K, Boon N, Siciliano SD, Verhaege M, Verstraete W (2004) Biofuel cells select for microbial consortia that self-mediate electron transfer. Appl Environ Microbiol 70: 5373-5382.

350. Walcott RR, Gitaitis RD, Castro AC (2003) Role of Blossoms in Watermelon Seed Infestation by Acidovorax avenae subsp. citrulli. Phytopathology 93: 528-534.

351. Lessner DJ, Parales RE, Narayan S, Gibson DT (2003) Expression of the nitroarene dioxygenase genes in Comamonas sp. strain JS765 and Acidovorax sp. strain JS42 is induced by multiple aromatic compounds. J Bacteriol 185: 3895-3904.

352. Byrne-Bailey KG, Weber KA, Chair AH, Bose S, Knox T, et al. (2010) Completed genome sequence of the anaerobic iron-oxidizing bacterium Acidovorax ebreus strain TPSY. J Bacteriol 192: 1475-1476.

353. Rabaey K, Read ST, Clauwaert P, Freguia S, Bond PL, et al. (2008) Cathodic oxygen reduction catalyzed by bacteria in microbial fuel cells. ISME J 2: 519-527.

354. Erable B, Duteanu NM, Ghangrekar MM, Dumas C, Scott K (2010) Application of electro-active biofilms. Biofouling 26: 57-71.

355. Elbein AD (1974) The metabolism of alpha,alpha-trehalose. Adv Carbohydr Chem Biochem 30: 227-256.

356. Elbein AD, Pan YT, Pastuszak I, Carroll D (2003) New insights on trehalose: a multifunctional molecule. Glycobiology 13: 17R-27R.

357. Edwards KJ, Bond PL, Gihring TM, Banfield JF (2000) An archaeal iron-oxidizing extreme acidophile important in acid mine drainage. Science 287: 1796-1799.

358. Gihring TM, Bond PL, Peters SC, Banfield JF (2003) Arsenic resistance in the archaeon "Ferroplasma acidarmanus": new insights into the structure and evolution of the ars genes. Extremophiles 7: 123-130.

359. Baker-Austin C, Dopson M, Wexler M, Sawers RG, Bond PL (2005) Molecular insight into extreme copper resistance in the extremophilic archaeon 'Ferroplasma acidarmanus' Fer1. Microbiology 151: 2637-2646.

360. Dopson M, Baker-Austin C, Bond PL (2005) Analysis of differential protein expression during growth states of Ferroplasma strains and insights into electron transport for iron oxidation. Microbiology 151: 4127-4137.

361. Baker-Austin C, Potrykus J, Wexler M, Bond PL, Dopson M (2010) Biofilm development in the extremely acidophilic archaeon 'Ferroplasma acidarmanus' Fer1. Extremophiles 14: 485-491.

362. Meldrum FC, Mann S, Heywood BR, Frankel RB, Bazylinski DA (1993) Electron Microscopy Study of Magnetosomes in a Cultured Coccoid Magnetotactic Bacterium. Proc R Soc Lond B 251: 231-236

363. Schleper C, Puehler G, Holz I, Gambacorta A, Janekovic D, et al. (1995) Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. J Bacteriol 177: 7050-7059.

364. Futterer O, Angelov A, Liesegang H, Gottschalk G, Schleper C, et al. (2004) Genome sequence of Picrophilus torridus and its implications for life around pH 0. Proc Natl Acad Sci U S A 101: 9091-9096.

365. van de V, Driessen AJ, Zillig W, Konings WN (1998) Bioenergetics and cytoplasmic membrane stability of the extremely acidophilic, thermophilic archaeon Picrophilus oshimae. Extremophiles 2: 67-74.

366. Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, et al. (2000) The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum. Nature 407: 508-513.

367. Kawashima T, Amano N, Koike H, Makino S, Higuchi S, et al. (2000) Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium. Proc Natl Acad Sci U S A 97: 14257-14262.

368. DeLong EF (2000) Extreme genomes. Genome Biol 1: REVIEWS1029.

369. Lowy FD (1998) Staphylococcus aureus infections. N Engl J Med 339: 520-532.

370. van Belkum A, Melles DC, Nouwen J, van Leeuwen WB, van Wamel W, et al. (2009) Co-evolutionary aspects of human colonisation and infection by Staphylococcus aureus. Infect Genet Evol 9: 32-47.

371. Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. Annu Rev Biochem 70: 369-413.

372. Mizuuchi K, O'Dea MH, Gellert M (1978) DNA gyrase: subunit structure and ATPase activity of the purified enzyme. Proc Natl Acad Sci U S A 75: 5960-5963.

373. Gellert M, Mizuuchi K, O'Dea MH, Nash HA (1976) DNA gyrase: an enzyme that introduces superhelical turns into DNA. Proc Natl Acad Sci U S A 73: 3872-3876.

374. Cozzarelli NR (1980) DNA gyrase and the supercoiling of DNA. Science 207: 953-960.

375. Wang JC (1996) DNA topoisomerases. Annu Rev Biochem 65: 635-692.

376. Maxwell A (1997) DNA gyrase as a drug target. Trends Microbiol 5: 102-109.

377. Bernard P, Kezdy KE, Van Melderen L, Steyaert J, Wyns L, et al. (1993) The F plasmid CcdB protein induces efficient ATP-dependent DNA cleavage by gyrase. J Mol Biol 234: 534-541.

378. Dwyer DJ, Kohanski MA, Hayete B, Collins JJ (2007) Gyrase inhibitors induce an oxidative damage cellular death pathway in Escherichia coli. Mol Syst Biol 3: 91.

379. Hanawalt PC (1966) The U.V. sensitivity of bacteria: its relation to the DNA replication cycle. Photochem Photobiol 5: 1-12.

380. Jaffe A, Ogura T, Hiraga S (1985) Effects of the ccd function of the F plasmid on bacterial growth. J Bacteriol 163: 841-849.

381. Kreuzer KN, Cozzarelli NR (1979) Escherichia coli mutants thermosensitive for deoxyribonucleic acid gyrase subunit A: effects on deoxyribonucleic acid replication, transcription, and bacteriophage growth. J Bacteriol 140: 424-435.

382. Willmott CJ, Critchlow SE, Eperon IC, Maxwell A (1994) The complex of DNA gyrase and quinolone drugs with DNA forms a barrier to transcription by RNA polymerase. J Mol Biol 242: 351-363.

383. Couturier M, Bahassi el M, Van Melderen L (1998) Bacterial death by DNA gyrase poisoning. Trends Microbiol 6: 269-275.

384. Bernard P, Couturier M (1992) Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes. J Mol Biol 226: 735-745.

385. Miki T, Park JA, Nagao K, Murayama N, Horiuchi T (1992) Control of segregation of chromosomal DNA by sex factor F in Escherichia coli. Mutants of DNA gyrase subunit A suppress letD (ccdB) product growth inhibition. J Mol Biol 225: 39-52.

386. Dao-Thi MH, Van Melderen L, De Genst E, Afif H, Buts L, et al. (2005) Molecular basis of gyrase poisoning by the addiction toxin CcdB. J Mol Biol 348: 1091-1102.

387. Hayes F (2003) Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest. Science 301: 1496-1499.

388. Rawlings DE (1999) Proteic toxin-antitoxin, bacterial plasmid addiction systems and their evolution with special reference to the pas system of pTF-FC2. FEMS Microbiol Lett 176: 269-277.

389. Van Melderen L (2010) Toxin-antitoxin systems: why so many, what for? Curr Opin Microbiol 13: 781-785.

390. Weinberg ED (1984) Iron withholding: a defense against infection and neoplasia. Physiol Rev 64: 65-102.

391. Boyd PW, Jickells T, Law CS, Blain S, Boyle EA, et al. (2007) Mesoscale iron enrichment experiments 1993-2005: synthesis and future directions. Science 315: 612-617.

392. B. NJ, T. P, A. LS (1980) High Affinity Iron Transport in Microorganisms. AMERICAN CHEMICAL SOCIETY 140: 263-278.

393. Klebba PE (2003) Three paradoxes of ferric enterobactin uptake. Front Biosci 8: s1422-1436.

394. Matsushiro A (1963) Specialized transduction of tryptophan markers in Escherichia coli K12 by bacteriophage phi-80. Virology 19: 475-482.

395. Gratia JP (1964) [Resistance to Colicin B in Escherichia Coli. Specificity Relations among Colicins B, I and V and Phage T-4. Genetic Study]. Ann Inst Pasteur (Paris) 107: SUPPL:132-151.

396. Bassford PJ, Jr., Bradbeer C, Kadner RJ, Schnaitman CA (1976) Transport of vitamin B12 in tonB mutants of Escherichia coli. J Bacteriol 128: 242-247.

397. Kaserer WA, Jiang X, Xiao Q, Scott DC, Bauler M, et al. (2008) Insight from TonB hybrid proteins into the mechanism of iron transport through the outer membrane. J Bacteriol 190: 4001-4016.

398. Brun A, Englund E (1986) Brain changes in dementia of Alzheimer's type relevant to new imaging diagnostic methods. Prog Neuropsychopharmacol Biol Psychiatry 10: 297-308.

399. Tsang PH, Li G, Brun YV, Freund LB, Tang JX (2006) Adhesion of single bacterial cells in the micronewton range. Proc Natl Acad Sci U S A 103: 5764-5768.

400. Ireland MM, Karty JA, Quardokus EM, Reilly JP, Brun YV (2002) Proteomic analysis of the Caulobacter crescentus stalk indicates competence for nutrient uptake. Mol Microbiol 45: 1029-1041.

401. Wagner JK, Setayeshgar S, Sharon LA, Reilly JP, Brun YV (2006) A nutrient uptake role for bacterial cell envelope extensions. Proc Natl Acad Sci U S A 103: 11772-11777.

402. Landt SG, Lesley JA, Britos L, Shapiro L (2010) CrfA, a small noncoding RNA regulator of adaptation to carbon starvation in Caulobacter crescentus. J Bacteriol 192: 4763-4775.

403. Fischer B, Rummel G, Aldridge P, Jenal U (2002) The FtsH protease is involved in development, stress response and heat shock control in Caulobacter crescentus. Mol Microbiol 44: 461-478.

404. Yeh YC, Comolli LR, Downing KH, Shapiro L, McAdams HH (2010) The caulobacter Tol-Pal complex is essential for outer membrane integrity and the positioning of a polar localization factor. J Bacteriol 192: 4847-4858.

405. Crossman LC, Gould VC, Dow JM, Vernikos GS, Okazaki A, et al. (2008) The complete genome, comparative and functional analysis of Stenotrophomonas maltophilia reveals an organism heavily shielded by drug resistance determinants. Genome Biol 9: R74.

406. Sader HS, Jones RN (2005) Antimicrobial susceptibility of uncommonly isolated non-enteric Gram-negative bacilli. Int J Antimicrob Agents 25: 95-109.

407. Denton M, Kerr KG (1998) Microbiological and clinical aspects of infection associated with Stenotrophomonas maltophilia. Clin Microbiol Rev 11: 57-80.

408. Winther L, Andersen RM, Baptiste KE, Aalbaek B, Guardabassi L (2010) Association of Stenotrophomonas maltophilia infection with lower airway disease in the horse: a retrospective case series. Vet J 186: 358-363.

409. Lidstrom M (2006) Aerobic Methylotrophic Prokaryotes. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E, editors. The Prokaryotes: Springer New York. pp. 618-634.

410. Galbally IE, Kirstine W (2002) The Production of Methanol by Flowering Plants and the Global Cycle of Methanol. Journal of Atmospheric Chemistry 43: 195-229.

411. Yabuuchi E, Yamamoto H, Terakubo S, Okamura N, Naka T, et al. (2001) Proposal of Sphingomonas wittichii sp. nov. for strain RW1T, known as a dibenzo-p-dioxin metabolizer. Int J Syst Evol Microbiol 51: 281-292.

412. Seah SY, Ke J, Denis G, Horsman GP, Fortin PD, et al. (2007) Characterization of a C-C bond hydrolase from Sphingomonas wittichii RW1 with novel specificities towards polychlorinated biphenyl metabolites. J Bacteriol 189: 4038-4045.

413. Miller TR, Delcher AL, Salzberg SL, Saunders E, Detter JC, et al. (2010) Genome sequence of the dioxin-mineralizing bacterium Sphingomonas wittichii RW1. J Bacteriol 192: 6101-6102.

414. Timoney JF (2004) The pathogenic equine streptococci. Vet Res 35: 397-409.

415. Weese JS, Jarlot C, Morley PS (2009) Survival of Streptococcus equi on surfaces in an outdoor environment. Can Vet J 50: 968-970.

416. Mushtaq N, Ezzati M, Hall L, Dickson I, Kirwan M, et al. (2011) Adhesion of Streptococcus pneumoniae to human airway epithelial cells exposed to urban particulate matter. J Allergy Clin Immunol 127: 1236-1242 e1232.

417. Xu J, Gordon JI (2003) Honor thy symbionts. Proc Natl Acad Sci U S A 100: 10452-10459.

418. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638.

419. Turnbaugh PJ, Gordon JI (2009) The core gut microbiome, energy balance and obesity. J Physiol 587: 4153-4158.

420. Hooper LV, Midtvedt T, Gordon JI (2002) How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annu Rev Nutr 22: 283-307.

421. Backhed F, Ding H, Wang T, Hooper LV, Koh GY, et al. (2004) The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A 101: 15718-15723.

422. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, et al. (2005) Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A 102: 11070-11075.

423. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022-1023.

424. Jernberg C, Lofmark S, Edlund C, Jansson JK (2010) Long-term impacts of antibiotic exposure on the human intestinal microbiota. Microbiology 156: 3216-3223.

425. Mazmanian SK, Liu CH, Tzianabos AO, Kasper DL (2005) An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. Cell 122: 107-118.

426. Mazmanian SK, Round JL, Kasper DL (2008) A microbial symbiosis factor prevents intestinal inflammatory disease. Nature 453: 620-625.

427. Savage DC (1977) Microbial ecology of the gastrointestinal tract. Annu Rev Microbiol 31: 107-133.

428. Wexler HM (2007) Bacteroides: the good, the bad, and the nitty-gritty. Clin Microbiol Rev 20: 593-621.

429. Reid G (2004) When microbe meets human. Clin Infect Dis 39: 827-830.

430. Troy EB, Kasper DL (2010) Beneficial effects of Bacteroides fragilis polysaccharides on the immune system. Front Biosci 15: 25-34.

431. Mallozzi M, Viswanathan VK, Vedantam G (2010) Spore-forming Bacilli and Clostridia in human disease. Future Microbiol 5: 1109-1123.

432. Wells CL, Wilkins TD (1996) Clostridia: Sporeforming Anaerobic Bacilli. In: Baron S, editor. Medical Microbiology. 4th ed. Galveston (TX).

433. Bartlett JG (2002) Clinical practice. Antibiotic-associated diarrhea. N Engl J Med 346: 334-339.

434. Hopkins MJ, Macfarlane GT (2002) Changes in predominant bacterial populations in human faeces with age and with Clostridium difficile infection. J Med Microbiol 51: 448-454.

435. Hopkins MJ, Macfarlane GT (2003) Nondigestible oligosaccharides enhance bacterial colonization resistance against Clostridium difficile in vitro. Appl Environ Microbiol 69: 1920-1927.

436. Villemur R, Lanthier M, Beaudet R, Lepine F (2006) The Desulfitobacterium genus. FEMS Microbiol Rev 30: 706-733.

437. Zijnge V, van Leeuwen MB, Degener JE, Abbas F, Thurnheer T, et al. (2010) Oral biofilm architecture on natural teeth. PLoS One 5: e9321.

438. Ehlers KT, Rusan M, Fuursted K, Ovesen T (2009) Fusobacterium necrophorum: most prevalent pathogen in peritonsillar abscess in Denmark. Clin Infect Dis: 1467-1472.

439. Rosado P, Gallego L, Junquera L, de Vicente JC (2009) Lemierre's syndrome: a serious complication of an odontogenic infection. Med Oral Patol Oral Cir Bucal 14: e398-401.

440. Lorber B (2005) Treatment of brain abscess due to Listeria monocytogenes. Clin Infect Dis 41: 419.

441. Jones DT, Woods DR (1986) Acetone-butanol fermentation revisited. Microbiol Rev 50: 484-524.

442. Dürre P (1998) New insights and novel developments in clostridial acetone/butanol/isopropanol fermentation. Applied Microbiology and Biotechnology 49: 639-648.

443. Fischer RJ, Helms J, Durre P (1993) Cloning, sequencing, and molecular analysis of the sol operon of Clostridium acetobutylicum, a chromosomal locus involved in solventogenesis. J Bacteriol 175: 6959-6969.

444. Smith CJ, Tribble GD, Bayley DP (1998) Genetic elements of Bacteroides species: a moving story. Plasmid 40: 12-29.

445. Shoemaker NB, Vlamakis H, Hayes K, Salyers AA (2001) Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon. Appl Environ Microbiol 67: 561-568.

446. Salyers AA, Shoemaker NB, Stevens AM, Li LY (1995) Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. Microbiol Rev 59: 579-590.

447. Salyers AA, Shoemaker NB, Li LY (1995) In the driver's seat: the Bacteroides conjugative transposons and the elements they mobilize. J Bacteriol 177: 5727-5731.

448. Beinert H, Kiley PJ (1999) Fe-S proteins in sensing and regulatory functions. Curr Opin Chem Biol 3: 152-157.

449. Beinert H, Holm RH, Munck E (1997) Iron-sulfur clusters: nature's modular, multipurpose structures. Science 277: 653-659.

450. Zheng L, Cash VL, Flint DH, Dean DR (1998) Assembly of iron-sulfur clusters. Identification of an iscSUA-hscBA-fdx gene cluster from Azotobacter vinelandii. J Biol Chem 273: 13264-13272.

451. Takahashi Y, Tokumoto U (2002) A third bacterial system for the assembly of iron-sulfur clusters with homologs in archaea and plastids. J Biol Chem 277: 28380-28383.

452. Johnson DC, Dean DR, Smith AD, Johnson MK (2005) Structure, function, and formation of biological iron-sulfur clusters. Annu Rev Biochem 74: 247-281.

453. Ayala-Castro C, Saini A, Outten FW (2008) Fe-S cluster assembly pathways in bacteria. Microbiol Mol Biol Rev 72: 110-125, table of contents.

454. Zheng L, White RH, Cash VL, Jack RF, Dean DR (1993) Cysteine desulfurase activity indicates a role for NIFS in metallocluster biosynthesis. Proc Natl Acad Sci U S A 90: 2754-2758.

455. Sa-Nogueira I, Nogueira TV, Soares S, de Lencastre H (1997) The Bacillus subtilis L-arabinose (ara) operon: nucleotide sequence, genetic organization and expression. Microbiology 143 (Pt 3): 957-969.

456. Schleif R (2000) Regulation of the L-arabinose operon of Escherichia coli. Trends Genet 16: 559-565.

457. Schonert S, Seitz S, Krafft H, Feuerbaum EA, Andernach I, et al. (2006) Maltose and maltodextrin utilization by Bacillus subtilis. J Bacteriol 188: 3911-3922.

458. Boos W, Shuman H (1998) Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation. Microbiol Mol Biol Rev 62: 204-229.

459. Hanson RS, Peterson JA, Yousten AA (1970) Unique biochemical events in bacterial sporulation. Annu Rev Microbiol 24: 53-90.

460. Driks A (1999) Bacillus subtilis spore coat. Microbiol Mol Biol Rev 63: 1-20.

461. Guarner F, Malagelada JR (2003) Gut flora in health and disease. Lancet 361: 512-519.

462. Kitaoka M, Tian J, Nishimoto M (2005) Novel putative galactose operon involving lacto-N-biose phosphorylase in Bifidobacterium longum. Appl Environ Microbiol 71: 3158-3162.

463. Hebbeln P, Rodionov DA, Alfandega A, Eitinger T (2007) Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module. Proc Natl Acad Sci U S A 104: 2909-2914.

464. Rodionov DA, Hebbeln P, Gelfand MS, Eitinger T (2006) Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters. J Bacteriol 188: 317-327.

465. Zhulin IB (2009) It is computation time for bacteriology! J Bacteriol 191: 20-22.

466. Rodionov DA, Hebbeln P, Eudes A, ter Beek J, Rodionova IA, et al. (2009) A novel class of modular transporters for vitamins in prokaryotes. J Bacteriol 191: 42-51.

467. Ji XB, Hollocher TC (1988) Reduction of nitrite to nitric oxide by enteric bacteria. Biochem Biophys Res Commun 157: 106-108.

468. Nathan C (1992) Nitric oxide as a secretory product of mammalian cells. FASEB J 6: 3051-3064.

469. Fang FC (1997) Perspectives series: host/pathogen interactions. Mechanisms of nitric oxide-related antimicrobial activity. J Clin Invest 99: 2818-2825.

470. Poole LB (2005) Bacterial defenses against oxidants: mechanistic features of cysteine-based peroxidases and their flavoprotein reductases. Arch Biochem Biophys 433: 240-254.

471. Brandes N, Rinck A, Leichert LI, Jakob U (2007) Nitrosative stress treatment of E. coli targets distinct set of thiol-containing proteins. Mol Microbiol 66: 901-914.

472. Ternan NG, Mc Grath JW, Mc Mullan G, Quinn JP (1998) Review: Organophosphonates: occurrence, synthesis and biodegradation by microorganisms. World Journal of Microbiology and Biotechnology 14: 635-647.

473. White AK, Metcalf WW (2004) Two C-P lyase operons in Pseudomonas stutzeri and their roles in the oxidation of phosphonates, phosphite, and hypophosphite. J Bacteriol 186: 4730-4739.

474. Chen CM, Ye QZ, Zhu ZM, Wanner BL, Walsh CT (1990) Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the phn (psiD) genes involved in alkylphosphonate uptake and C-P lyase activity in Escherichia coli B. J Biol Chem 265: 4461-4471.

475. Pittard J (1964) Effect of phage-controlled restriction on genetic linkage in bacterial crosses. J Bacteriol 87: 1256-1257.

476. Nelson KE, Fleischmann RD, DeBoy RT, Paulsen IT, Fouts DE, et al. (2003) Complete genome sequence of the oral pathogenic Bacterium porphyromonas gingivalis strain W83. J Bacteriol 185: 5591-5601.

477. Sakamoto M, Kitahara M, Benno Y (2007) Parabacteroides johnsonii sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 57: 293-296.

478. Boente RF, Ferreira LQ, Falcao LS, Miranda KR, Guimaraes PL, et al. (2010) Detection of resistance genes and susceptibility patterns in Bacteroides and Parabacteroides strains. Anaerobe 16: 190-194.

479. Wu XY, Shi KL, Xu XW, Wu M, Oren A, et al. (2010) Alkaliphilus halophilus sp. nov., a strictly anaerobic and halophilic bacterium isolated from a saline lake, and emended description of the genus Alkaliphilus. Int J Syst Evol Microbiol 60: 2898-2902.

480. Takai K, Moser DP, Onstott TC, Spoelstra N, Pfiffner SM, et al. (2001) Alkaliphilus transvaalensis gen. nov., sp. nov., an extremely alkaliphilic bacterium isolated from a deep South African gold mine. Int J Syst Evol Microbiol 51: 1245-1256.

481. Cao X, Liu X, Dong X (2003) Alkaliphilus crotonatoxidans sp. nov., a strictly anaerobic, crotonate-dismutating bacterium isolated from a methanogenic environment. Int J Syst Evol Microbiol 53: 971-975.

482. Phillips JE (1961) The commensal role of Actinobacillus lignieresi. J Pathol Bacteriol 82: 205-208.

483. Guettler MV, Rumler D, Jain MK (1999) Actinobacillus succinogenes sp. nov., a novel succinic-acid-producing strain from the bovine rumen. Int J Syst Bacteriol 49 Pt 1: 207-216.

484. Van der Werf MJ, Guettler MV, Jain MK, Zeikus JG (1997) Environmental and physiological factors affecting the succinate product ratio during carbohydrate fermentation by Actinobacillus sp. 130Z. Arch Microbiol 167: 332-342.

485. Janssen PH, Kirs M (2008) Structure of the archaeal community of the rumen. Appl Environ Microbiol 74: 3619-3625.

486. Irbis C, Ushida K (2004) Detection of methanogens and proteobacteria from a single cell of rumen ciliate protozoa. J Gen Appl Microbiol 50: 203-212.

487. Miller TL, Wolin MJ, Zhao HX, Bryant MP (1986) Characteristics of methanogens isolated from bovine rumen. Appl Environ Microbiol 51: 201-202.

488. Whitford MF, Teather RM, Forster RJ (2001) Phylogenetic analysis of methanogens from the bovine rumen. BMC Microbiol 1: 5.

489. Leahy SC, Kelly WJ, Altermann E, Ronimus RS, Yeoman CJ, et al. (2010) The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions. PLoS One 5: e8926.

490. Warnick TA, Methe BA, Leschine SB (2002) Clostridium phytofermentans sp. nov., a cellulolytic mesophile from forest soil. Int J Syst Evol Microbiol 52: 1155-1160.

491. Koehler TM (2009) Bacillus anthracis physiology and genetics. Mol Aspects Med 30: 386-396.

492. Green DH, Wakeley PR, Page A, Barnes A, Baccigalupi L, et al. (1999) Characterization of two Bacillus probiotics. Appl Environ Microbiol 65: 4288-4291.

493. Belaich JP, Tardif C, Belaich A, Gaudin C (1997) The cellulolytic system of Clostridium cellulolyticum. J Biotechnol 57: 3-14.

494. Doi RH, Goldstein M, Hashida S, Park JS, Takagi M (1994) The Clostridium cellulovorans cellulosome. Crit Rev Microbiol 20: 87-93.

495. Chesson A, Stewart CS, Dalgarno K, King TP (1986) Degradation of isolated grass mesophyll, epidermis and fibre cell walls in the rumen and by cellulolytic rumen bacteria in axenic culture. journal of applied microbiology 60: 327–336.

496. Cai S, Li J, Hu FZ, Zhang K, Luo Y, et al. (2010) Cellulosilyticum ruminicola, a newly described rumen bacterium that possesses redundant fibrolytic-protein-encoding genes and degrades lignocellulose with multiple carbohydrate- borne fibrolytic enzymes. Appl Environ Microbiol 76: 3818-3824.

497. Kooistra J, Venema G (1991) Cloning, sequencing, and expression of Bacillus subtilis genes involved in ATP-dependent nuclease synthesis. J Bacteriol 173: 3644-3655.

498. Goldmark PJ, Linn S (1972) Purification and properties of the recBC DNase of Escherichia coli K-12. J Biol Chem 247: 1849-1860.

499. Oishi M (1969) An ATP-dependent deoxyribonuclease from Escherichia coli with a possible role in genetic recombination. Proc Natl Acad Sci U S A 64: 1292-1299.

500. Palas KM, Kushner SR (1990) Biochemical and physical characterization of exonuclease V from Escherichia coli. Comparison of the catalytic activities of the RecBC and RecBCD enzymes. J Biol Chem 265: 3447-3454.

501. Kooistra J, Vosman B, Venema G (1988) Cloning and characterization of a Bacillus subtilis transcription unit involved in ATP-dependent DNase synthesis. J Bacteriol 170: 4791-4797.

502. Kupsch J, Alonso JC, Trautner TA (1989) Analysis of structural and biological parameters affecting plasmid deletion formation in Bacillus subtilis. Mol Gen Genet 218: 402-408.

503. Meima R, Haijema BJ, Venema G, Bron S (1995) Overproduction of the ATP-dependent nuclease AddAB improves the structural stability of a model plasmid system in Bacillus subtilis. Mol Gen Genet 248: 391-398.

504. Peijnenburg AA, Bron S, Venema G (1987) Structural plasmid instability in recombination- and repair-deficient strains of Bacillus subtilis. Plasmid 17: 167-170.

505. Meima R, Haijema BJ, Dijkstra H, Haan GJ, Venema G, et al. (1997) Role of enzymes of homologous recombination in illegitimate plasmid recombination in Bacillus subtilis. J Bacteriol 179: 1219-1229.

506. Amundsen SK, Fero J, Salama NR, Smith GR (2009) Dual nuclease and helicase activities of Helicobacter pylori AddAB are required for DNA repair, recombination, and mouse infectivity. J Biol Chem 284: 16759-16766.

507. Gibson GR, Beatty ER, Wang X, Cummings JH (1995) Selective stimulation of bifidobacteria in the human colon by oligofructose and inulin. Gastroenterology 108: 975-982.

508. Gibson GR, Wang X (1994) Enrichment of bifidobacteria from human gut contents by oligofructose using continuous culture. FEMS Microbiol Lett 118: 121-127.

509. Bailey JS, Blankenship LC, Cox NA (1991) Effect of fructooligosaccharide on Salmonella colonization of the chicken intestine. Poult Sci 70: 2433-2438.

510. Buddington KK, Donahoo JB, Buddington RK (2002) Dietary oligofructose and inulin protect mice from enteric and systemic pathogens and tumor inducers. J Nutr 132: 472-477.

511. Zhang X, Vrijenhoek JE, Bonten MJ, Willems RJ, van Schaik W (2011) A genetic element present on megaplasmids allows Enterococcus faecium to use raffinose as carbon source. Environ Microbiol 13: 518-528.

512. Prade RA (1996) Xylanases: from biology to biotechnology. Biotechnol Genet Eng Rev 13: 101-131.

513. Moracci M, Cobucci Ponzano B, Trincone A, Fusco S, De Rosa M, et al. (2000) Identification and molecular characterization of the first alpha -xylosidase from an archaeon. J Biol Chem 275: 22082-22089.

514. Sunna A, Antranikian G (1997) Xylanolytic enzymes from fungi and bacteria. Crit Rev Biotechnol 17: 39-67.

515. Garcia Sanchez R, Karhumaa K, Fonseca C, Sanchez Nogue V, Almeida JR, et al. (2010) Improved xylose and arabinose utilization by an industrial recombinant Saccharomyces cerevisiae strain using evolutionary engineering. Biotechnol Biofuels 3: 13.

516. Navarre WW, Schneewind O (1999) Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. Microbiol Mol Biol Rev 63: 174-229.

517. Mandlik A, Swierczynski A, Das A, Ton-That H (2008) Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. Trends Microbiol 16: 33-40.

518. Scott JR, Zahner D (2006) Pili with strong attachments: Gram-positive bacteria do it differently. Mol Microbiol 62: 320-330.

519. Ton-That H, Liu G, Mazmanian SK, Faull KF, Schneewind O (1999) Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of Staphylococcus aureus at the LPXTG motif. Proc Natl Acad Sci U S A 96: 12424-12429.

520. Suree N, Yi SW, Thieu W, Marohn M, Damoiseaux R, et al. (2009) Discovery and structure-activity relationship analysis of Staphylococcus aureus sortase A inhibitors. Bioorg Med Chem 17: 7174-7185.

521. Mao H, Hart SA, Schink A, Pollok BA (2004) Sortase-mediated protein ligation: a new method for protein engineering. J Am Chem Soc 126: 2670-2671.

522. Jin DJ, Cabrera JE (2006) Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in Escherichia coli. J Struct Biol 156: 284-291.

523. Dabrowska G, Prusiniska J, Goc A (2006) [The stringent response--bacterial mechanism of an adaptive stress response]. Postepy Biochem 52: 87-93.

524. Potrykus K, Cashel M (2008) (p)ppGpp: still magical? Annu Rev Microbiol 62: 35-51.

525. Kerscher L, Nowitzki S, Oesterhelt D (1982) Thermoacidophilic archaebacteria contain bacterial-type ferredoxins acting as electron acceptors of 2-oxoacid:ferredoxin oxidoreductases. Eur J Biochem 128: 223-230.

526. Schloss PD, Handelsman J (2003) Biotechnological prospects from metagenomics. Curr Opin Biotechnol 14: 303-310.

Chapter 2. Rothamsted Park Grass soil study of undisturbed microbial communities: an evenness vision

1. **Introduction (French part)**
2. **TerraGenome: a consortium for the sequencing of a soil metagenome**
3. **Assessing the soil metagenome for studies of microbial diversity Structure, fluctuation and magnitude of a grassland soil metagenome**

The main objectives of my PhD were i) to improve methods to access the genetic diversity (called metagenome) present in the Rothamsted Park Grass soil (reference of the Terragenome consortium; see section 1) and ii) to characterize microbial communities present in this environment using a direct sequencing approach.

This chapter represents what was done to reach these two objectives.

The strategy I developed to stimulate the accessibility of soil metagenomes is presented in the section 3. This study was done using a fingerprint (RISA) and a microarray technology, and emphasized important DNA extraction biases and new possibilities to increase the detection of taxa by fractionating soil, cells and DNA. In my point of view, one of the main conclusions of this study was that the diversity is highly underestimated when using only one approach to access DNA from soil, what is generally done. Thus soil biodiversity estimations have to be raised.

We decided then to apply partially this approach using next sequencing technologies (section 4). Our goal was not to sequence and assemble this soil metagenome, too diverse for actual tools, but more to characterize the functional potential of predominant microorganisms. Thus 13 sequencing efforts were done (one million reads each) using various DNA extraction methods, two seasons and two depths to maximize the standard deviation of each annotated function or taxa. After defining a global picture of this environment, it was possible to compare the life style of these communities to other, sequenced from other environment (e.g., oceans).

Introduction:

Les microorganismes sont apparus il y a plus de 3.5 milliards d'années (Allwood et al., 2006), ~1.5 milliards d'années après la formation de notre planète. La flexibilité génétique dans une vaste période de temps géologique a permis aux microorganismes de s'adapter à tous les écosystèmes virtuellement concevable de la Terre (e.g. Huber et al., 2007; Pointing et al., 2009; Larose et al., 2010). Parmi les écosystèmes contemporains, le sol, qui est le produit d'une vie microbienne et macrobienne, exhibe la plus grande densité et diversité phylogénétique par unité de volume (Van Elsas et al., 2006; Roesch et al., 2007), avec approximativement un milliard de cellules par gramme, et comprenant une diversité estimée allant de milliers à millions d'especes (Knietch et al., 2003).

Les communautés microbiennes du sol sont indispensable a la santé de la planète; elles sont le moteur des grands cycles géochimiques (Falkowski et al., 2001) et aident au developpement des plantes (Ortíz-Castro et al., 2009). Actuellement, il y a toujours un considérable manque de compréhension des mécanismes d'interaction et des métabolismes qui existent parmi les membres des communautés microbiennes et de leur écosystème. La connaissance actuelle de la diversité taxinomique et fonctionnelle, le potentiel métabolique des communautés, ainsi que les conséquences en termes d'adaptation évolutive sont largement basés sur l'information partielle extraite des études de microorganismes cultivés. Une dépendance d'études d'organismes cultivés a fondamentalement limité notre compréhension de la diversité des interactions dans ce système. En effet, les organismes cultivés de la terre représentent seulement une fraction de la vie microscopique présente dans le sol (e.g., ceux capables de se développer dans des conditions contrôlées en laboratoire ; Schloss and Handelsman, 2003).

L'application des méthodologies de metagenomique à des échantillons de sol a été entravé par des challenges techniques considérables, comme l'extraction non biaisée et représentative d'un échantillon de matériel génétique provenant d'organismes possédant des membranes cellulaires très différentes et donc un ADN plus ou moins accessible (Delmont et al., 2011; Demaneche et al, 2008; Ginolhac et al, 2004; Handelsman et al, 1998; Rajendhran and Gunasekaran, 2008). Ce problème est exacerbé par la distribution des communautés microbiennes du sol (Ranjard and Richaume, 2001). A l'inverse des systèmes marins qui permettent de claires observations temporelles et géographiques (Rusch et al., 2007; Gilbert et al. 2009, 2010), seulement une fraction de la diversité microbienne du sol, assumée comme étant immense (e.g. Tringe et al., 2005; Roesch et al., 2007; Morales et al., 2009), a pour l'instant été explorée malgré une abondance de données acquises lors de centaines d'expériences de haute qualité scientifique.

Malgré de nombreux efforts, les données en provenance du sol sont rares en comparaison de celles collectées à partir d'autre écosystèmes. Le seule metagenome de sol contemporain publié (Tringe et al., 2005) ne contient que 100 millions de paires de bases d'ADN, ce qui représente potentiellement un simple millionième d'un pourcent du materiel genetique

pouvant être extrait à partir d'un gramme de sol (basé sur la supposition d'une moyenne de taille de génome de 4 millions de paires de bases et d'un milliard de cellules par gramme de sol).

Le manque relatif de séquences générées à partir du sol et accessibles représente un paradoxe intéressant: l'environnement le plus diverse de la planète a reçu le moins d'attention a partir d'analyses métagénomiques. Afin de redresser la balance, nous avons réalisé l'investigation la plus en profondeur de metagenomique du sol en utilisant la technologie de pyroséquençage.

Le sol est essentiel pour la production de nourriture d'une population planétaire augmentant constamment ; il est donc essentiel de découvrir les mécanismes par lesquelles les microorganismes influencent la production des plantes. Elucider le rôle que jouent les microorganismes dans le sol devrait permettre la manipulation de cet écosystème pour le bénéfice de la vie. Construite à partir des précédentes investigations (Vogel et al., 2009; Delmont et al., 2011), cette étude décrit un effort sans précédent de caractérisation de la diversité microbienne et du potentiel fonctionnel d'un unique écosystème de sol présent dans le site dénommé « Park Grass » de la station expérimentale de Rothamsted ; cette station représente la plus ancienne expérience agriculturale du monde, se déroulant sans discontinue depuis 1856. Dans le but d'explorer cet unique environnement, pratiquement 5 milliards de paires de bases de séquences métagénomiques ont été produites à partir de trois profondeurs et de trois périodes recouvrant deux années. Afin de répondre a l'inquiétude vis-à-vis de l'influence des techniques d'extraction d'ADN sur la diversité microbienne du sol (Delmont et al., 2011), nous avons accompli 11 techniques d'extraction différentes, a la fois afin d'augmenter la diversité détectée et d'explorer l'effet de ces méthodes sur la diversité microbienne observée.

Utilisant MG RAST afin d'annoter les séquences, les différents jeux de données ont été comparés les uns aux autres, ainsi qu'a des données correspondant à d'autres environnements afin de placer ces données dans un contexte plus global.

EDITORIAL

TerraGenome: a consortium for the sequencing of a soil metagenome

Vogel and colleagues invite the microbiology community to participate in an ambitious and extraordinary sequencing project to uncover the soil metagenome.

Timothy M. Vogel and Pascal Simonet are at the Environmental Microbial Genomics Group, Laboratoire AMPERE, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France.
Janet K. Jansson is at the Lawrence Berkeley National Laboratory, Division of Earth Sciences, Berkeley, California 94720, USA.
Penny R. Hirsch is at Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK.
James M. Tiedje is at the Center for Microbial Ecology, Michigan State University, East Lansing, Michigan 48824, USA.
Jan Dirk van Elsas is at the Department of Microbial Ecology, Centre for Ecological and Evolutionary Studies, University of Groningen, PO BOX 149750 AA Haren, The Netherlands.
Mark J. Bailey is at the Centre for Ecology & Hydrology, CEH-Wallingford, Maclean Building, Crowmarsh Gifford, Wallingford, Oxon OX10 8BB, UK.
Renaud Nalin is at the LibraGen, 3 rue des Satellites, 31400 Toulouse, France.
Laurent Philippot is at INRA, Université de Bourgogne, UMR Microbiologie du Sol et de l'Environnement, CMSE, BP 86510, 21065 Dijon Cedex, France.
Correspondence to T.M.V.
e-mail: Timothy.vogel@ec-lyon.fr

The microorganisms in the 'living soil' are fundamental to all higher life on our planet and are responsible for terrestrial processes that determine our quality of life, including soil fertility, carbon cycling and nutrient cycling. Given the importance of soil functions to most aspects of our lives, surprisingly little is understood about the vulnerability of soil to perturbations or its functional resilience; for example, to changes in land use or climate. We do not fully understand soil biogeochemistry or spatial complexity, or how soil processes, such as carbon cycling, contribute to climate change. Although microorganisms are responsible for key functions in soils, only a small percentage (less than 0.5%) have been grown in the laboratory and genome sequences are only available for a select few.

Soil is the most biodiverse environment on the Earth: it is estimated to contain approximately 1,000 Gbp of microbial genome sequences per gram of soil! Compared with the Human Genome project (in which 3 Gbp were sequenced)[1] and sequencing projects that target microbial habitats, such as the Sargasso Sea (for which 6 Gbp were sequenced)[2], metagenomic sequencing of soil remains rudimentary and constitutes a new and ambitious challenge. We propose that soil should be our next global metagenomic sequencing initiative.

Sequencing the soil metagenome will bring considerable economic and environmental value. The soil microbial community is a gold mine for genes and pathways that encode novel biocatalysts for biosynthetic or biodegradation processes, including degradation of pollutants, synthesis of biofuels and production of novel drugs. Sequencing of the soil metagenome will also provide insights into the ecology of microorganisms that are beneficial to, or threaten, crop production, and that ensure the quality and provision of ecosystem services.

Recent developments in high-throughput sequencing methods have put the goal of deciphering the soil metagenome within our reach. Large-scale metagenomic sequencing efforts will be necessary to resolve the intricacies of the soil microbiome and to provide sufficient data to understand soil microbial community diversity and function. The success of soil metagenomics depends on a combination of intelligent sample selection, efficient DNA extraction methods, cloning, screening strategies and sequencing approaches, together with open system data management and sharing. Owing to the magnitude of this task, we propose that a coordinated international effort should be established to combine the skills of the global scientific community to focus on sequencing and annotating the soil metagenome. To catalyse this process, we are seeking agreement and cooperation from the scientific community in reaching a primary objective: the complete sequencing of a 'reference' soil metagenome. The soil system chosen for investigation, Park Grass, is an internationally recognized agroecology field experiment that has been running for more than 150 years at the UK agricultural sciences institute, Rothamsted Research. This ambitious reference sequencing effort cannot be undertaken by a single laboratory or even by a single country. We therefore invite the international community to participate in this project, and hope to eventually expand the project to other soil sites. The information gleaned from this project will serve as a starting block or platform for other soil metagenomic sequencing efforts and will generate new hypotheses to test. This initiative will also spur complementary efforts in other 'omic approaches, such as transcriptomics, proteomics and metabolomics of soil to add more layers of information about gene expression, activity and function. In addition, we will need to develop and apply new approaches to cultivate the previously uncultivated and rare members of the soil community to assign functions to the vast number of unknown or hypothetical genes that will undoubtedly be found. The TerraGenome international sequencing consortium, which is dedicated to soil metagenomics, has just been launched to coordinate these efforts.

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
2. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PloS Biol.* **5**, e77 (2007).

FURTHER INFORMATION
TerraGenome international sequencing consortium: http://www.terragenome.org/

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Feb. 2011, p. 1315–1324 Vol. 77, No. 4
0099-2240/11/$12.00 doi:10.1128/AEM.01526-10

Accessing the Soil Metagenome for Studies of Microbial Diversity▽†

Tom O. Delmont,[1] Patrick Robe,[2] Sébastien Cecillon,[1] Ian M. Clark,[3] Florentin Constancias,[1] Pascal Simonet,[1] Penny R. Hirsch,[3] and Timothy M. Vogel[1]*

Environmental Microbial Genomics Group, Laboratoire Ampère, Ecole Centrale de Lyon, Université de Lyon, 36 Avenue Guy de Collongue, 69134 Ecully, France[1]; LibraGen, 3 Rue des Satellites, 31400 Toulouse, France[2]; and Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, United Kingdom[3]

Received 27 June 2010/Accepted 13 December 2010

Soil microbial communities contain the highest level of prokaryotic diversity of any environment, and metagenomic approaches involving the extraction of DNA from soil can improve our access to these communities. Most analyses of soil biodiversity and function assume that the DNA extracted represents the microbial community in the soil, but subsequent interpretations are limited by the DNA recovered from the soil. Unfortunately, extraction methods do not provide a uniform and unbiased subsample of metagenomic DNA, and as a consequence, accurate species distributions cannot be determined. Moreover, any bias will propagate errors in estimations of overall microbial diversity and may exclude some microbial classes from study and exploitation. To improve metagenomic approaches, investigate DNA extraction biases, and provide tools for assessing the relative abundances of different groups, we explored the biodiversity of the accessible community DNA by fractioning the metagenomic DNA as a function of (i) vertical soil sampling, (ii) density gradients (cell separation), (iii) cell lysis stringency, and (iv) DNA fragment size distribution. Each fraction had a unique genetic diversity, with different predominant and rare species (based on ribosomal intergenic spacer analysis [RISA] fingerprinting and phylochips). All fractions contributed to the number of bacterial groups uncovered in the metagenome, thus increasing the DNA pool for further applications. Indeed, we were able to access a more genetically diverse proportion of the metagenome (a gain of more than 80% compared to the best single extraction method), limit the predominance of a few genomes, and increase the species richness per sequencing effort. This work stresses the difference between extracted DNA pools and the currently inaccessible complete soil metagenome.

The soil microbial community is relatively diverse (9, 31), with arguably the highest level of prokaryotic diversity of any environment (32, 41). One gram of soil has been reported to contain up to 10 billion microorganisms and thousands of different species (20). This soil species pool represents a goldmine for genes involved in pharmaceutical and industrial applications (42) and in the biodegradation of human-made pollutants (4, 13). Currently, less than 1% of this diversity is considered to be cultivable by traditional techniques (34), a problem that can be circumvented by metagenomic approaches. Metagenomic approaches have been applied to study a range of soil environments (8, 10, 15, 17, 28), and comparisons with cultivation techniques should include biases in the methods used to extract DNA from soil. Different DNA extraction methods are widely used, although they each have biases that restrict the diversity of the so-called metagenomic DNA (6, 12, 18, 22, 24, 25). Therefore, the total microbial diversity of soil might still be underestimated, independent of the method used to calculate the species (or operational taxonomic unit [OTU]) diversity in a soil. Indeed, the relative dominance of certain groups in DNA extracted from soil will mask less abundant species, thus confounding estimates of soil microbial community structure.

Recently developed technologies provide relatively quick and deep sequencing of metagenomic DNA samples at a moderate cost (19, 35), although metagenomic DNA sequencing, however completely sequenced, depends on the DNA extracted. Deciphering soil function based on soil metagenome sequencing (such as that proposed previously by the Terragenome International Consortium [43]) requires extraction of the DNA from all members of the soil microbial community. The difficulty is that every protocol facilitates the extraction of part of the microbially diverse population to the detriment of the rest. Biodiversity estimates from a variety of methods (Fig. 1) already range from 10^4 species (32, 38) to 10^7 species (14) per gram of soil. Therefore, a measure of the dependence of biodiversity estimates on metagenomic access would aid in an understanding of whether sequencing depth or DNA extraction diversity is driving diversity estimations.

Our approach was to combine different methods to recover different spectra of community diversity in order to increase access to the biodiverse soil community. We applied four classes of DNA (or microbial) separation techniques that significantly resolve DNA diversity. These techniques are based on (i) vertical soil sampling, (ii) cell separation in a density gradient, (iii) cell lysis stringency, and (iv) DNA fragment size distribution (Fig. 2). Although the respective methods used are not without some overlap, we have shown that they can be adjusted to increase the relative diversity of the final DNA pool. In other words, by varying the conditions of the four

* Corresponding author. Mailing address: Environmental Microbial Genomics, Laboratoire Ampere, Ecole Centrale de Lyon, Université de Lyon, 36 Avenue Guy de Collongue, 69134 Ecully, France. Phone: 33 4 72 18 65 14. Fax: 33 4 78 43 37 17. E-mail: tvogel@ec-lyon.fr.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 23 December 2010.

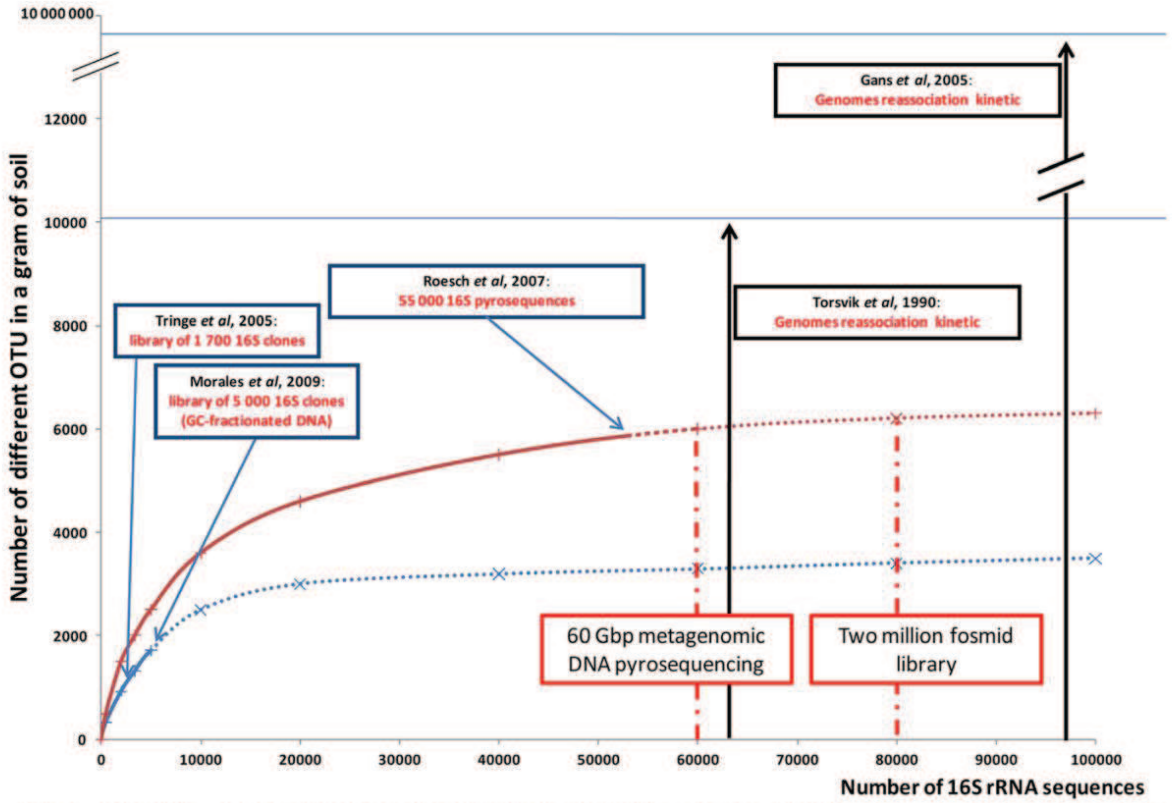


FIG. 1. Theoretical contribution of the Terragenome Initiative to soil diversity exploration, which starts with 60 "454" titanium plates and the construction of a 2-million-fosmid (40-kb inserts) clone library in the context of soil microbial diversity estimation studies (Metasoil Project). (Based on data from references 14, 26, 32, 37, and 39.)

methods and applying a phylogenetic technique to track relative diversity and the less represented species, the final DNA pool can be optimized for increased nucleic acid diversity. This strategy was compared to other more common approaches (including the individual application of one of the methods used here) in order to illustrate the advantages of this approach. Although applying these four variables might improve the already distorted view of the relative abundance of species, the aim here is to enhance species and gene discovery by maximizing the identification of the genetic diversity of a DNA pool before high-throughput sequencing efforts or the construction of libraries is performed.

MATERIALS AND METHODS

Soil samples. Samples were collected from two sites: Ecully, France, and an untreated control plot (plot "3d") of Park Grass (lat 51.481481°N, long 0.222231°E), Rothamsted, England (see http://www.rothamsted.ac.uk/for further information), in October 2008 and March 2009, respectively. The Park Grass soil is an internationally recognized resource and is targeted as a reference soil for soil metagenomic studies (43). It is classified as chromic luvisol according to FAO guidelines (11) and is a silty clay loam overlying clay with flints with a pH of 5.2 (measured in H_2O). Park Grass covers 249 m^2 (13.28 by 18.75 m), and the sampling strategy consisted of taking randomized soil samples in four areas of the plot (horizontal sampling) and at seven depths (vertical sampling, each 3 cm between 0 and 21 cm). The Ecully soil (silty topsoil) was sampled in a grassland area (lat 45.470759°N, long 4.460152°E) at the same seven depths. Samples were placed into plastic bags and transported on ice. Soil was homogenized manually by thorough physical mixing. All tools and materials used were washed and sterilized.

DNA extraction methods. DNA extraction from soil is a key step in the metagenomic approach (3, 12, 21). Two different methods are routinely used. In the first method, direct extraction, cells are lysed within the soil sample (27, 40, 44). We used two direct DNA extraction protocols that involve bead beating: a method described previously Griffiths et al. (16) that uses the FastPrep lysing matrix (MP bead beating; Bio101 Biomedical) and the MoBio UltraClean soil DNA isolation kit. For both protocols, DNA was extracted from 0.5 g of soil. For the alternative method, cells were first removed from the soil (60 g) and then lysed (2). This method is commonly called indirect extraction and has been reported to separate prokaryotic from eukaryotic cells via a Nycodenz density gradient (1, 7, 23). During the centrifugation, the Nycodenz gradient is stabilized at a density of 1.3 g/ml and should isolate prokaryotes to form a cellular fraction called the cell ring (Fig. 2). We fractionated the gradient into six parts, each 5 ml (four fractions above the cell ring, the cell ring, and one below the cell ring [total of 30 ml]), by varying the centrifugation speed (1,000 × *g*, 2,000 × *g*, 5,000 × *g*, and 9,000 × *g* for 40 min). After centrifugation at each speed, the Nycodenz gradient was subsampled from the top down by pipetting out 5-ml samples. The cell ring was within the fifth subsample.

After cell separation in the gradient, we used different cell lysis protocols, which have various degrees of stringency: the MP bead-beating protocol, the Epicentre Gram-positive kit, the Nucleospin tissue kit, and five agarose plug protocols called protocols A, B, C, D, and E.

Agarose plugs. The extraction of soil bacteria was performed on fresh soil samples as previously described by Bertrand et al. (3), using the Nycodenz

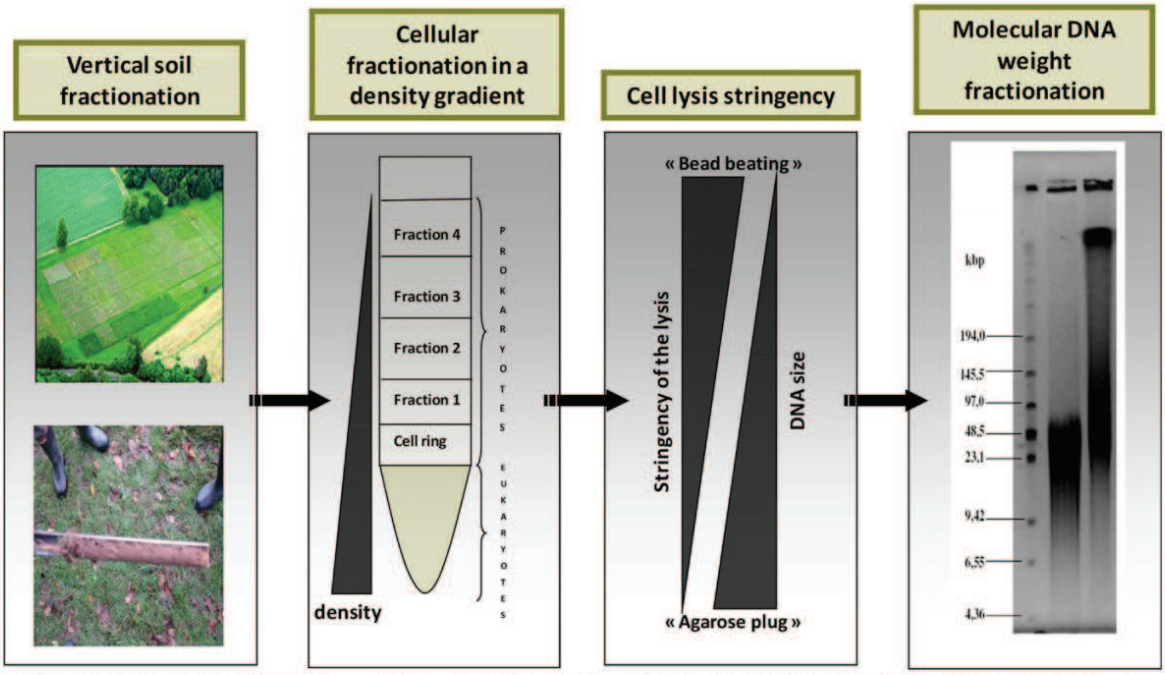


FIG. 2. Schematic of the different classes of DNA separation methods, starting with physical distance in the field and then density differences in Nycodenz gels, resistance to cell lysis, and finally DNA size separation by pulsed-field gel electrophoresis.

gradient separation method. The collected bacterial cell fraction was washed with ultrapure water and then centrifuged for 10 min at 12,000 × *g*. The cell pellet was then resuspended in a 50 mM Tris (pH 8)–100 mM EDTA buffer, mixed with an equal volume of molten 1.6% Incert agarose, and then transferred into disposable plug molds (Bio-Rad). The lysis of the soil bacteria was then performed with agarose. After the different lysis methods were used, agarose plugs were equilibrated in a 10 mM Tris (pH 8.0)–1 mM EDTA storage buffer.

(i) Protocol A. For protocol A, agarose plugs were first transferred into 3 ml of G^- lysis buffer (1% lauroyl sarcosine, 500 mM EDTA-Na_2 [pH 9.5]) with 0.5 mg/ml of lysozyme and incubated at 37°C for 12 h. The agarose plugs were then incubated in 3 ml of G^- lysis buffer with 500 μg/ml of proteinase K at 56°C for 12 h.

(ii) Protocol B. For protocol B, agarose plugs were first transferred into 45 ml of LA lysis buffer (50 mM Tris [pH 8.0], 100 mM EDTA, 5 mg of lysozyme/ml, 0.5 mg of achromopeptidase/ml) and incubated at 37°C for 6 h. The agarose plugs were then incubated in 45 ml of SP lysis buffer (50 mM Tris [pH 8.0], 100 mM EDTA, 1% lauryl sarcosyl, 2 mg of proteinase K/ml) at 55°C for 24 h. An additional incubation for 24 h was performed with fresh SP buffer.

(iii) Protocol C. For protocol C, agarose plugs were first transferred into 3 ml of G^+ lysis buffer (6 mM Tris-HCl, 100 mM EDTA-Na_2, 1 M NaCl, 0.5% Brij 58, 0.2% sodium deoxycholate, 0.5% lauroyl sarcosine [pH 7.5]) with 0.5 mg/ml of lysozyme and incubated at 37°C for 12 h. The agarose plugs were then incubated in 3 ml of G^- lysis buffer with 500 μg/ml of proteinase K at 56°C for 12 h.

(iv) Protocol D. For protocol D, agarose plugs were incubated in 45 ml of SP lysis buffer (50 mM Tris [pH 8.0], 100 mM EDTA, 1% lauryl sarcosyl, 2 mg of proteinase K/ml) at 55°C for 24 h. An additional incubation for 24 h was performed with fresh SP buffer.

(v) Protocol E. For protocol E, agarose plugs were transferred into 45 ml of LA lysis buffer (50 mM Tris [pH 8.0], 100 mM EDTA, 5 mg of lysozyme/ml, 0.5 mg of achromopeptidase/ml) and incubated at 37°C for 6 h.

These plug protocols for differential DNA recovery have also been used for fosmid library construction that requires high-molecular-weight DNA to create clone libraries with different sequence diversities.

DNA size separation. Pulsed-field gel electrophoresis (PFGE) was used to separate the metagenomic DNA as a function of the fragment size distribution (1% low-melting-point agarose and 0.5× Tris-borate-EDTA [TBE], with a program of 2 s, 20 s, and 15 h). The DNA was then extracted from the gel by using agarase I (New England BioLabs Inc.). For the Rothamsted soil samples, a portion of DNA was physically sheared to generate a range of fragments that were smaller than those in the undisrupted portion, as demonstrated by the differential migration of the smears (Fig. 2).

Ribosomal intergenic spacer analysis (RISA). The intergenic spacer (IGS) region between the small (16S) and the large (23S) subunits of ribosomal sequences were amplified by PCR using primers 5′-TGCGGCTGGATCCCCTC CTT-3′ (forward) and 5′-CCGGGTTTCCCCATTCGG-3′ (reverse) (29). For the PCR mix, 2 μl of DNA (10 μM) was mixed with 1.25 μl of reverse and forward primers (10 μM) and 20.5 μl of distilled water (DH_2O). PCR cycles consisted of 95°C for 10 min and then 30 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 1 min, followed by 72°C for 15 min, with a Biometra thermocycler. One microliter of the PCR mix was then loaded into an Agilent DNA 7500 Lab on a Chip, and electropherograms were analyzed and data were normalized by using an Agilent 2100 Bioanalyzer. An example of different replicates is shown in Fig. S1 in the supplemental material in order to demonstrate the reproducibility of this fingerprint approach.

Phylochip analyses. The microarray format used in these experiments was that from Agilent Sureprint Technologies. The format used consisted of 8 blocks of 15,000 spots each on a standard glass slide format, 1 in. by 3 in. (25 mm by 75 mm). Each spot was formed by the *in situ* synthesis of 20-mer oligonucleotide probes. Each oligonucleotide probe occurred at least in triplicate within each block. All blocks were identical. This format provides for the hybridization of eight samples at the same time and on the same slide. The use of multiple slides was necessary for the hybridization of over eight samples. Probes were designed to target the *rrs* gene and to cover a wide part of the *Bacteria* and *Archaea* phylogenic tree. Probes were designed with the ARB software package and PhylArray (24a). We have chosen to design 20-mer probes with a melting temperature range of 65°C ± 5°C and with a weighted mismatch of less than 1.5. Our design includes oligonucleotide probes at different taxonomic levels. This microarray covers over 400 genera and 400 OTUs ("species" or "hits").

The *rrs* genes were amplified by PCR from total DNA by using universal primer pA (TAATACGACTCACTATAGAGAGTTTGATCCTGGCTCAG) and pH-T7 (AAGGAGGTGATCCAGCCGCA) (5) (universal for most members of the *Bacteria* and some of the *Archaea*) under standard conditions. The

amplification of DNA was performed with a 48-μl PCR mixture using 5 U of Ex *Taq* titanium polymerase. PCR was conducted at 94°C for 4 min and then with 35 cycles of 94°C for 45 s, 55°C for 45 s, and 68°C for 90 s, followed by 68°C for 5 min. Amplified PCR products were electrophoresed on a 1% agarose gel, and the desired 1.5-kb bands were removed and purified by using GFX PCR DNA and a gel band purification kit (Amersham Biosciences). Purified PCR products were then transcribed onto RNA using T7 RNA polymerase (Invitrogen) with the incorporation of labeled Cy3-UTP. Cy3 is a fluorescent dye, emitting light at 532 nm. RNA purification was performed by using the Qiagen RNeasy minikit according to the manufacturer's instructions. RNA fragmentation was achieved by the addition of 1.14 μl of Tris-Cl (1 mM) and 4.57 μl of $ZnSO_4$ (100 mM) to 40 μl of labeled RNA sample and incubation for 30 min at 60°C. Chemically fragmented labeled RNA was then hybridized to the phylochips.

Microarray scanning and data processing. An Innoscan (Carbonne, France) 700 scanner was used for scanning microarray slides according to the manufacturer's instructions. Raw hybridization fluorescence signals for each spot were determined based on the signal-to-noise ratio (SNR), which was calculated by using the following formula: SNR = (signal intensity − background)/standard deviation of the background. Hybridization fluorescence signals for all probes, including negative controls, were transformed by calculating the $\log_2$ of the signal. Since at least three replicates exist for all oligonucleotide probes, outliers were eliminated when any individual spot was greater than 2 standard deviations from the average of all replicates. Analysis of variance (ANOVA) was used to evaluate positive probes from the results for all microarray data from one experiment. Since the probes have different phylogenetic depths, the genera described here were those for which all relevant probes were positive. While all of the thousands of probes could not be independently verified, many of the probes were validated by the application of DNA from a single bacterium (33).

RESULTS

Two different soils were employed for the elaboration of the DNA-recovering strategy: Ecully, France, and Park Grass (plot 3d), Rothamsted, England. Different approaches were tested to separate and increase the metagenomic DNA extracted at one time and one place, and the diversity of the different samples was estimated with ribosomal intergenic spacer analysis (RISA) fingerprinting (electropherogram profiles are shown in Fig. 3). After preliminary tests, four methods appeared to separate metagenomic DNA into the most diverse fractions for the two soils: vertical soil sampling, density gradient (cell separation), cell lysis stringency, and DNA fragment size distribution (Fig. 2). The different methods could be applied sequentially to maximize differential DNA extraction (Fig. 2). Clearly, all the extracted DNA pools have distinct species diversities and distributions. However, the cell lysis stringency appeared to have the most influence on the diversity of the extracted DNA pool, as discussed below.

RISA profiles. RISA fingerprints representing electropherograms demonstrate the presence and absence of different populations within the DNA extracted from the microbial community (Fig. 3). The different DNA extraction methods applied to the two soils are compared in four categories (vertical samples [Fig. 3A and A′], Nycodenz separation [Fig. 3B and B′], the cell lysis procedure [Fig. 3C and C′], and DNA size differences [Fig. 3D and D′]). The fingerprints of the microbial community extracted by the different methods are all different (in contrast to the similar profiles seen with replicate samples) (see Fig. S1 in the supplemental material), although the differences are more pronounced for those methods that include different lysis procedures (Fig. 3C and C′). Soil sample depth (Fig. 3A and A′) and DNA size (Fig. 3D and D′) showed the fewest differences, although extreme size classes were noticeably different (i.e., 40 kb in C and 250 kb in C′). The use of different lysis procedures had the greatest impact on RISA diversity, together covering the entire spectrum of possible RISA peaks (Fig. 3C and C′), in contrast to DNA fractionated according to size, which has several areas without peaks (Fig. 3D and D′). In order to evaluate the differences between the different RISA profiles, the Rothamsted profiles were quantified and the different samples were compared with a principal component analysis (PCA). The PCA separated the groups principally as a function of the cell lysis procedure (Fig. 4). Within these large groups, other parameters are regrouped, such as fractions from different depths in the soil core, the Nycodenz gradient fractions (where "top" refers to different samples from above the cell ring), and the PFGE smear (where Bw1 is about 40 kb, Bw2 is about 100 kb, and Bw3 is about 250 kb). In addition, the MP bead-beating extraction method was applied to both the soil (direct cell extraction) and the Nycodenz cell ring. While the bead beating produced somewhat similar RISA profiles, the direct and indirect ("cell ring") samples were differentiated by the PCA (Fig. 4). Some replicates are provided in order to evaluate the relative importance of the RISA profiles. For example, the MoBio kit method was performed twice on the deepest soil sample (18 to 21 cm deep), and the bead beating was performed three times on the second depth fraction (4 to 6 cm). All of the replicates grouped relatively closely together (Fig. 4).

Taxonomic comparisons. For the Rothamsted soil, the difference between the metagenomic DNAs extracted by these different methods was further explored with the phylogenetic microarray in order to determine which genera were selectively extracted by one approach or the other. Comparisons of the microarray responses were therefore made between different extraction protocols.

In addition, the same DNA extraction protocol (MP bead beating) was used to evaluate the microbial diversity differences as a function of depth (vertical soil sampling). Phylochip analysis using 16S rRNA gene (*rrs*) hybridization showed significant diversity variations, with the frequency of *Bacillus* spp. increasing and that of *Mesorhizobium* species decreasing with depth. Some genera were detected in only one fraction. For example, *Sandarakinotalea* was detected only at the 3- to 6-cm depth; *Alkalibacillus* and *Ammoniphilus* were detected at the lowest depth (see Fig. S2 in the supplemental material). After centrifugation at a relatively low speed (2,000 × *g*), the density gradient was subsampled in six fractions (four fractions above the cell ring, one at the cell ring, and one below the cell ring of 5 ml each). One DNA extraction protocol (Epicentre Gram-positive kit) was used for phylochip comparisons. The frequency of detection of the genera *Glycomyces* and *Legionella* increased with depth in the Nycodenz gradient. Moreover, the populations of some genera were relatively isolated in one fraction and undetected or at very low levels in all others (e.g., *Marinobacter*, *Pseudoxanthomonas*, *Fervidobacterium*, and *Treponema*), emphasizing the value of varying the centrifugation speed to access different metagenomic DNAs (see Fig. S3 in the supplemental material).

After the soil and cell separation, different cellular lysis protocols were used to separate the metagenomic DNA as a function of the cell wall resistance to lysis. Seven different protocols were applied. In addition, two direct extraction protocols (DNA extracted directly from the soil), the MP bead-beating protocol and the MoBio Ultraclean soil DNA kit, were

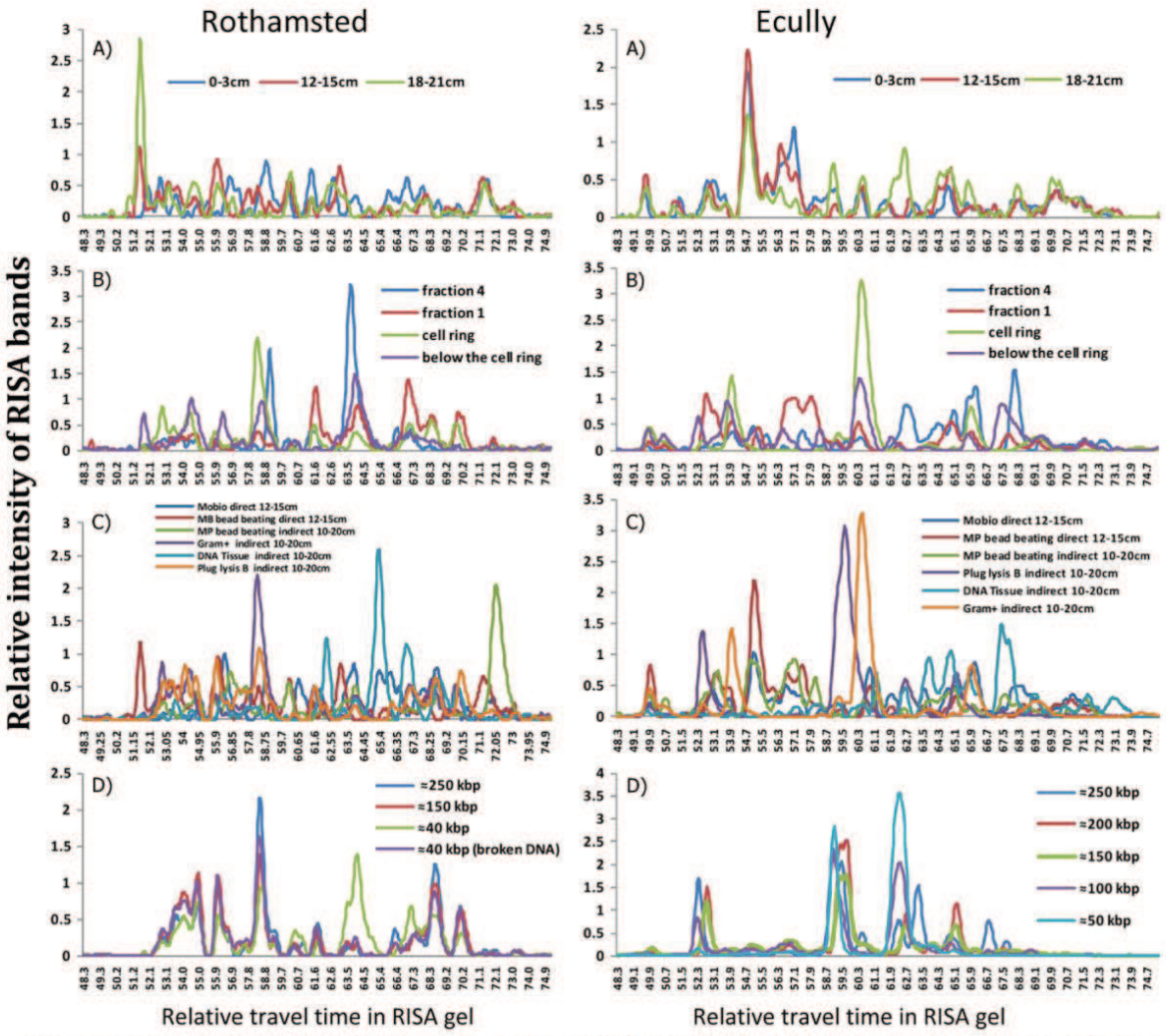


FIG. 3. Multiple examples of ribosomal intergenic spacer analysis (RISA) electropherograms of DNA from the Ecully and Park Grass, Rothamsted, soils illustrating the differences in the diversities of microbial community DNA as a function of the applied separation technique. Graphs represent relative RISA band intensities as a function of travel time in the gel. (A) Physical separation with MP bead-beating direct DNA extraction of soil samples from different depths (0 to 3, 12 to 15, and 18 to 21 cm deep). (B) Cellular fractionation in a density gradient (Gram-positive lysis after centrifugation at 5,000 × *g*). (C) Cell lysis with MP bead-beating and MoBio kit direct DNA extractions (depth, 12 to 15 cm) and Epicentre Gram-positive, bead-beating, and DNA tissue indirect DNA extractions (depth, 0 to 10 cm). (D) Metagenomic DNA fractionation by PFGE after extraction by plug protocol B (depth, 10 to 20 cm).

applied to the soils. The seven other protocols were indirect extraction protocols (cells extracted before lysis), the same MP bead-beating protocol, the Epicentre Gram-positive kit, the Nucleospin tissue kit, and five different agarose plug protocols, by varying the lysis stringency. Each lysis method facilitated the DNA extraction of a part of the microbially diverse population to the detriment of the rest. For example, MP bead-beating direct DNA extraction (fraction of 0 to 3 cm) facilitated the extraction of the genera *Brevundimonas* and *Mesorhizobium* but not the genera *Sphingobium* (detected only with plug lysis protocol E) or *Pseudomonas*. On the other hand, indirect bead-beating DNA extraction accessed more members of the *Pseudomonas* genus but not *Mesorhizobium* or *Gloeobacter* (see Fig. S4 in the supplemental material).

Finally, after an in-plug lysis (protocol B), DNA was separated as a function of its size distribution by pulsed-field gel electrophoresis. This separated DNA based on its molecular weight. The low-molecular-weight (30- to 50-kb) fraction was extracted and analyzed directly, and the DNA was fragmented so that the 250-kb fraction was fragmented down to the same size (30 to 50 kb) and then analyzed by phylochip analysis. Some genera were clearly unevenly represented in these two

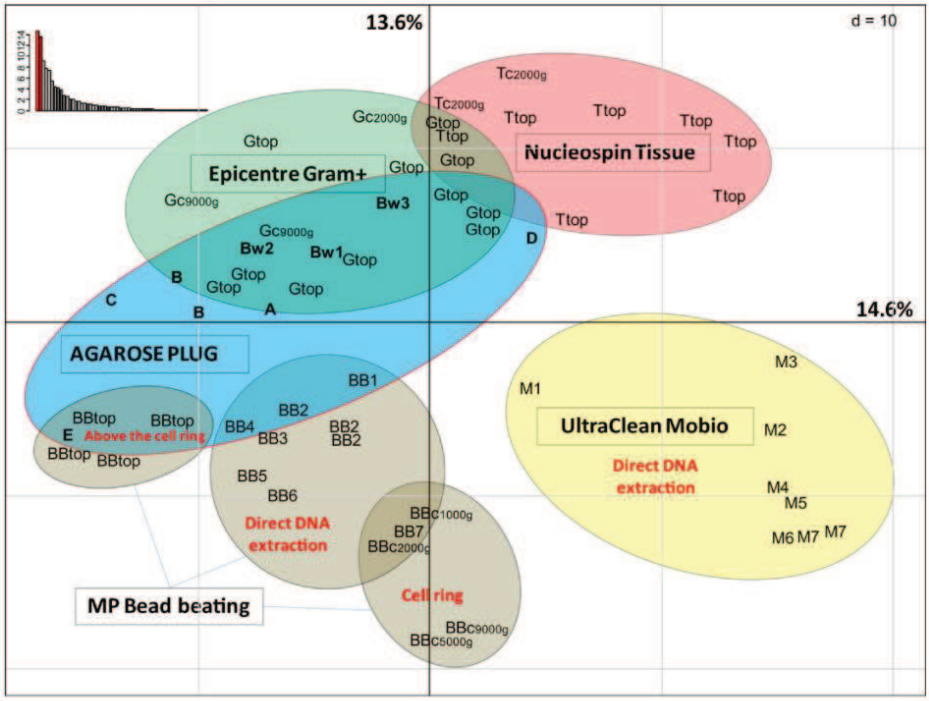


FIG. 4. Principal component analysis (showing the first and second components) of the matrix data for the RISA analysis from each DNA separation method. The percentages of variance of all axes are shown in the upper left corner. BB, bead beating; A, B, C, D, and E, agarose plug protocols; Bw1, Bw2, and Bw3, low-, medium-, and high-molecular-weight DNA extracted with plug protocol B; M, MoBio Ultraclean kit; G, Epicentre G$^+$ kit; c, cell ring from the Nycodenz density gradient separation; top, DNA recovered from the different fractions above the cell ring. The numbers 1 to 7 refer to the depth intervals (3 cm deep each) of the soil samples from the soil core, with 1 being 0 to 3 cm and 2 being 4 to 6 cm, etc.; 1,000g refers to DNA recovered from the cell ring in the Nycodenz gradient when the centrifuge was operated at 1,000 $\times$ *g* rather than the usual 9,000 $\times$ *g*.

DNA samples (see Fig. S5 in the supplemental material). Notably, the genera *Sulfurimonas*, *Xylella*, and *Leuconostoc* were undetected in the low-molecular-weight (30- to 50-kb) fraction but were easily detected in the high-molecular-weight fraction. On the other hand, the genera *Marinobacter* and *Rhodopirellula* were detected only in the low-molecular-weight fraction (30 to 50 kb). These results demonstrate the variation in genetic diversity in the soil metagenomic DNA smear and might explain some of the bias found in the fosmid clone libraries, as the DNA selected is generally between 25 and 40 kb.

The relative phylogenetic distributions (based on probe hybridization intensities) of soil DNA pools extracted as a function of all four parameters (soil depth, Nycodenz gradient depth, cell lysis stringency, and DNA size) were also compared. The presence or absence of different genera and their relative fluorescence intensities from the different DNA pools were plotted against those for the MP bead-beating direct lysis of the top soil fraction (Fig. 5, black line). Thus, this pool of DNA defines the order (descending) of the genera (not listed here) along the *x* axis from most abundant to least abundant (Fig. 5). This DNA pool had 218 identified genera, which is why the genera after the 218th genus were not detected in the MP bead-beating top soil fraction of "0 to 3 cm" but were detected in other DNA extracts. All other DNA pools were thus compared to this pool, and where there are peaks above the black line, the pool in question has more of a given genus, and where there are valleys, the given pool has less of a given genus than those determined by MP bead beating. For example, *Mesorhizobium* (Fig. 5, far left) is the most predominant genus in the reference DNA pool (MP bead-beating direct lysis of the top soil fraction of 0 to 3 cm), more so than in any other extraction method's DNA pool. Other examples include *Pseudomonas* in the DNA pool from the MP bead beating applied to the Nycodenz cell ring after centrifugation at 9,000 $\times$ *g* and *Bacillus* in the DNA pool from the MP bead-beating direct lysis on the bottom soil sample (Fig. 5). Note that when the reference pool (MP bead-beating direct lysis of the top soil sample) does not detect certain genera at all (Fig. 5, right), several different extraction DNA pools have relatively high levels of these genera (e.g., *Marinobacter* with the Gram-positive extraction of the Nycodenz cell ring at 1,900 $\times$ *g* and *Sphingobium* with cell lysis procedure E). Many genera were not detected by using a single DNA extraction protocol but were revealed by applying other protocols. While some protocols, like direct and indi-

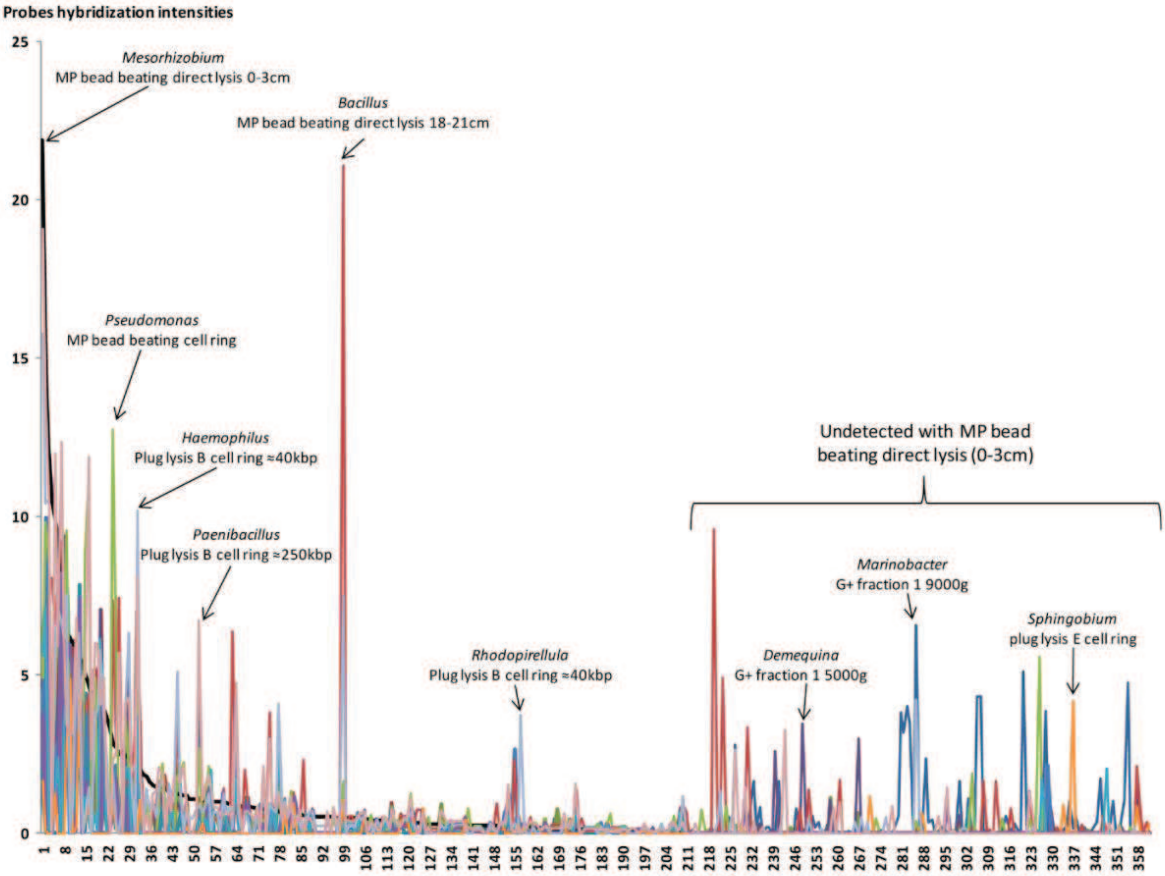


FIG. 5. Phylogenetic distribution (genus level) of 14 DNA pools for 360 different genera. The genus order is based on the decreasing percentage of those detected in the DNA pool extracted with MP bead-beating direct lysis of the surface (0- to 3-cm) soil sample (black line) from genera 1 to 218. The order from genera 218 to 360 (where the genera were not detected in the reference DNA pool) is alphabetical.

rect MP bead beating, access more genera than some of the more specific extraction protocols, the relative proportions are not the same. In any case, no single protocol accesses the entire microbial community metagenome. When the phylogenetic probes on the microarray are quantified by extraction techniques, the numbers of phyla, classes, genera, and potential species ("hits") vary considerably (e.g., from 50 to 214 genera) between protocols (Table 1). However, while

TABLE 1. Potential microbial biodiversity detected from the Rothamsted soil as a function of the extraction technique[a]

Method	No. of phyla	No. of classes	No. of orders	No. of families	No. of genera	Total no. of hits
MP bead-beating direct DNA extraction, 0–3 cm	19	29	60	107	210	1,892
MP bead-beating direct DNA extraction, 18–21 cm	18	28	58	107	214	2,024
G^+ fraction 4, 2,000 × *g*, 0–10 cm	12	18	29	39	51	333
G^+ cell ring, 9,000 × *g*, 0–10 cm	17	24	49	75	121	1,201
G^+ cell ring, 2,000 × *g*, 0–10 cm	15	22	47	72	121	974
Plug lysis protocol B cell ring, 0–10 cm, ~40 kbp	18	26	58	102	203	2,130
Plug lysis protocol B cell ring, 0–10 cm, ~250 kbp	17	26	56	95	182	1,887
Plug lysis protocol D cell ring, 10–23 cm	16	23	47	63	109	869
Plug lysis protocol E cell ring, 10–23 cm	12	16	27	32	50	270
23 different DNA extraction approaches	23	36	71	148	385	3,940

[a] The value of 2,000 × *g* refers to DNA recovered from the cell ring in the Nycodenz gradient when the centrifuge was operated at 2,000 × *g* rather than at the usual 5,000 × *g*.

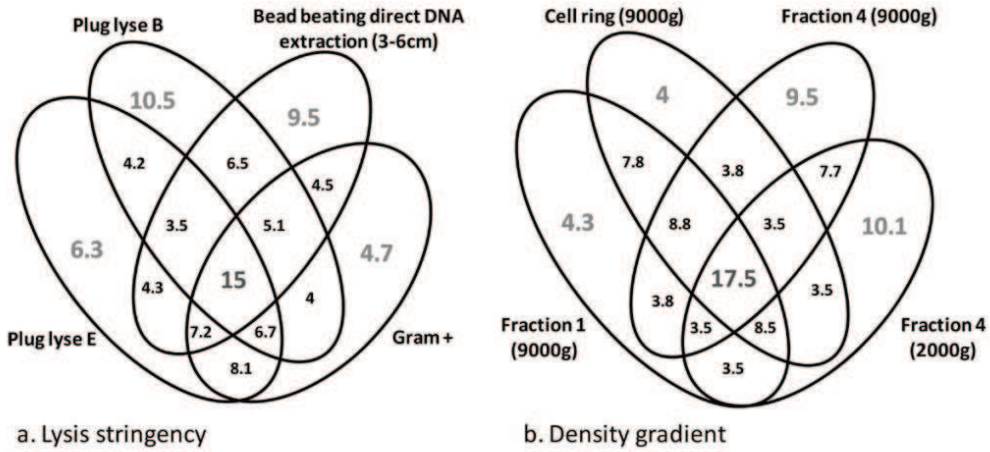


FIG. 6. Venn diagram showing percentages of probe hybridization coverage (out of over 3,000 total) between DNA extraction protocols as a function of the lysis stringency (a) and location in a Nycodenz density gradient at different centrifugation speeds (b).

some protocols detected relatively low numbers of genera (e.g., lysis procedure E), these protocols add to the overall recovery of diversity. For example, if all 23 different protocols were used, then 385 different genera would be detected (Table 1).

DISCUSSION

The exploration of the biodiversity in soils requires metagenomic approaches that extract DNA from all the *Bacteria* and *Archaea* present as comprehensively as is possible. The scale of the spatial variation of the microbial diversity in a soil must influence any attempts to recover the genomes of all members of the microbial community. RISA profiles showed that Park Grass diversity varies both horizontally and vertically; however, the vertical variation appeared to be greater. To increase the level of biodiversity recovered from soil, we applied a range of approaches to access the metagenomic DNA pool. These approaches were dependent on soil depth, cell separation in density gradients, cell lysis stringency, and DNA molecular weight. The often-applied strategy of sampling different locations at the site was not the most significant factor in increasing the level of diversity of DNA extracted from the soils tested here (Fig. 3 and 4). Rather, the most critical strategies were those applied to the soil samples in the laboratory to extract and fractionate cells and DNA. This implies that the sample size (roughly 100 g) was sufficient to capture the majority of the microbial community metagenome. Nevertheless, all of the different approaches, including vertical soil sampling, altered the accessible biodiversity. The relevant issue was the relative improvement achieved with every additional DNA extraction protocol.

While all cell lysis protocols have numerous biases that limit the diversity of the metagenomic DNA extracted (12), we used these biases to our advantage in order to access different soil microbial communities with different proportions of species represented. This approach separated the metagenomic DNA as a function of cell wall resistance to lysis. RISA analyses showed important differences between lysis methods. The PCA corresponding to RISA profiles of some Rothamsted soil DNA samples emphasized the importance of this step (Fig. 4). The lysis protocol was the major driving force in grouping microbially diverse communities and thus was a crucial step for DNA extraction differences. These different lysis methods had significant effects on the metagenomic DNA extracted from a soil, with different microbial populations being represented in each sample (Fig. 5). Furthermore, we made an effort to access different diverse populations with the agarose plug protocols (five different lysis protocols) so that this strategy could be coupled with fosmid clone library production.

No one protocol can provide an accurate determination of species distribution, and therefore, different DNA extraction protocols, more or less stringent, could be employed, and the DNA pools could then be mixed together to maximize the number of different species represented and to decrease the proportion of the dominant species with a consequent increase in the final level of metagenomic diversity. The true relative abundance of different species is not currently determinable, and both microarray approaches and attempts to validate "16S" clone libraries by quantitative PCR are unfortunately dealing with the same DNA extraction pool (e.g., see reference 26) and, thus, the same extraction bias. Nevertheless, improved knowledge of the species present in the soil will aid in our understanding of soil function independent of their relative abundances. Since the majority of microorganisms are probably underrepresented in soil (30, 36), they are not easily accessible for study. Our approach was to maximize the representation of different species in DNA extracted from the same soil using four different techniques in order to improve our understanding of soil biodiversity.

To visualize the impact of our strategy on accessing different levels of biodiversity in soil, sample DNA was analyzed with a phylochip containing the 20-mer complementary strands of the 16S rRNA gene (*rrs*). The different strategies clearly extracted different relative numbers of genera (Fig. 5 and Table 1), with some not detecting the presence of certain genera (Fig. 5).

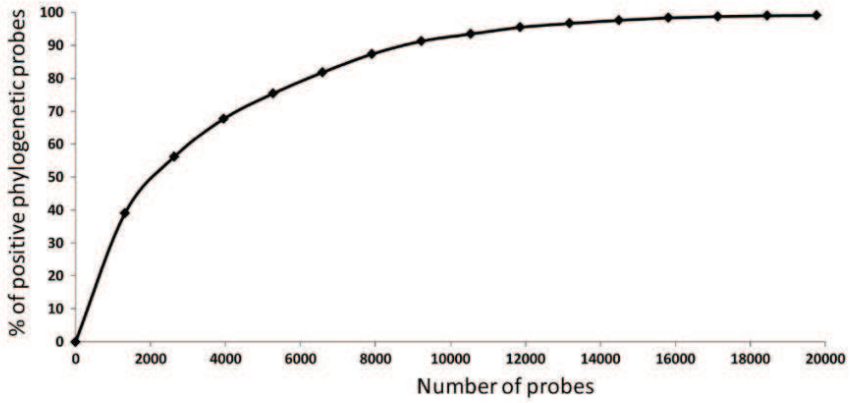


FIG. 7. Rarefaction curve based on phylogenetic microarray analyses of 15 different (based on extraction methods) DNA pools from the Rothamsted soil samples. The percentage of positive probes is plotted against the number of probes tested over multiple microarrays used to test different DNA pools.

These rather large differences confirm the requirement for multiple approaches when high levels of microbial diversity are sought. Clearly, there is some overlap between different DNA extraction strategies (Fig. 3 and 4). In the case of different lysis stringencies or cellular fractions in a density gradient, at least 15% of the biodiversity (as measured by positive phylogenetic microarray probes) was detected in all DNA extraction method variations. On the other hand, over 20% of the biodiversity was detected only in individual pools of extracted DNA (Fig. 6). The different approaches tested appear to access variable quantities of phyla, classes, families, genera, and species (corresponding to different "hits" in the NCBI database) (Table 1). Some of the methods accessed a maximum amount of diversity (e.g., MP direct DNA extraction and plug protocol B indirect DNA extraction), while others provide in-depth information on diversity (e.g., fraction 4 of the density gradient with the Gram-positive Epicentre kit or plug protocol E indirect DNA extraction), which can help metagenomic DNA assemblages and provide access to generally unrepresented genetic resources. Combining the outputs from the different methods provides a greater level of biodiversity than any individual approach, increasing the number of hypothetical species by 83.5% in comparison to the best individual DNA extraction method tested.

None of the different extraction protocols described here are suitable for high-throughput sequencing, although PCR approaches can be easily applied to prokaryote community studies. The yield is particularly low when DNA is extracted from the cell density gradient fractions above the cell ring and when the DNA is extracted from agarose gels. In theory, it is possible to use whole-genome amplification to increase yields, but the inherent bias in this method would considerably limit the utility of sequencing these fractionated parts of a soil metagenome. There is some anecdotal evidence that the lower the DNA yield, the more the DNA sample represents unique phyla. The challenge is to accumulate sufficient DNA with low-yield approaches to enable high-throughput sequencing. Sequencing may not be appropriate for comparisons across many samples but is likely to be crucial when species richness and diversity within a small number of soil samples need to be defined in detail.

We have defined a strategy for increasing the level of detection of metagenomic DNA diversity in two soils by employing multiple DNA extraction methods. By comparing these multiple methods, we showed that the spatial distance between soil samples did not have a major impact on the genetic diversity that was determined, in contrast to both depth and the different DNA extraction and purification methods. The mixed metagenomic DNA containing products from different soil depths and with different extraction factors (density gradient, cell lysis stringency, and DNA molecular weight) will maximize the representation of different species, although it may distort their relative abundance at the nucleic acid level. However, the "true" distribution is unknown, and no existing method provides this information. To the contrary, most methods provide limited views of the true soil biodiversity, and it is only by adopting a range of extraction and lysis methods that rare species are captured, thus increasing the number of species detected (Table 1). The increase in the phylochip probe diversity from these different DNA fractions follows standard rarefaction curves (Fig. 7). These results imply that the level of soil diversity is greater than estimations based on one DNA extraction method (e.g., see references 14, 32, and 38). Therefore, considerable efforts and technologies are needed to access not only DNA pools but also an entire metagenome for unbiased microbial ecology studies.

ACKNOWLEDGMENTS

We thank the French National Research Agency (ANR GMGE Metasoil). We also thank Libragen for its help and collaboration.

T.O.D. was funded by the Rhône-Alpes region. Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom.

REFERENCES

1. **Bakken, L. R.** 1985. Separation and purification of bacteria from soil. Appl. Environ. Microbiol. **49:**1482–1487.
2. **Berry, A. E., C. Chiocchini, T. Selby, M. Sosio, and E. M. Wellington.** 2003. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. FEMS Microbiol. Lett. **223:**15–20.

3. **Bertrand, H., et al.** 2005. High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. J. Microbiol. Methods **62:**1–11.
4. **Boubakri, H., M. Beuf, P. Simonet, and T. M. Vogel.** 2006. Development of metagenomic DNA shuffling for the construction of a xenobiotic gene. Gene **375:**87–94.
5. **Bruce, K. D., et al.** 1992. Amplification of DNA from native populations of soil bacteria by using the polymerase chain reaction. Appl. Environ. Microbiol. **58:**3413–3416.
6. **Carrig, C., O. Rice, S. Kavanagh, G. Collins, and V. O'Flaherty.** 2007. DNA extraction method affects microbial community profiles from soils and sediment. Appl. Microbiol. Biotechnol. **77:**955–964.
7. **Courtois, S., et al.** 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. Environ. Microbiol. **3:**431–439.
8. **Courtois, S., et al.** 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. Appl. Environ. Microbiol. **69:**49–55.
9. **Curtis, T. P., W. T. Sloan, and J. W. Scannell.** 2002. Estimating prokaryotic diversity and its limits. Proc. Natl. Acad. Sci. U. S. A. **99:**10494–10499.
10. **Demaneche, S., et al.** 2008. Antibiotic-resistant soil bacteria in transgenic plant fields. Proc. Natl. Acad. Sci. U. S. A. **105:**3957–3962.
11. **FAO.** 2006. Guidelines for soil description. FAO, Rome, Italy. ftp://ftp.fao.org/agl/agll/docs/guidel_soil_descr.pdf.
12. **Frostegård, A., et al.** 1999. Quantification of bias related to the extraction of DNA directly from soil. Appl. Environ. Microbiol. **65:**5409–5420.
13. **Galvao, T. C., W. W. Mohn, and V. de Lorenzo.** 2005. Exploring the microbial biodegradation and biotransformation gene pool. Trends Biotechnol. **23:**497–506.
14. **Gans, J., M. Wolinsky, and J. Dunbar.** 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science **309:**1387–1390.
15. **Ginolhac, A., et al.** 2004. Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. Appl. Environ. Microbiol. **70:**5522–5527.
16. **Griffiths, R. I., A. S. Whitely, A. G. O'Donnell, and M. J. Bailey.** 2000. Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. Appl. Environ. Microbiol. **66:**5488–5491.
17. **Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman.** 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. **5:**R245–R249.
18. **Head, I. M., J. R. Saunders, and R. W. Pickup.** 1998. Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. Microb. Ecol. **35:**1–21.
19. **Kahvejian, A., J. Quackenbush, and J. F. Thompson.** 2008. What would you do if you could sequence everything? Nat. Biotechnol. **26:**1125–1133.
20. **Knietch, A., T. Waschkowitz, S. Bowien, A. Henne, and R. Daniel.** 2003. Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on Escherichia coli. J. Microbiol. Biotechnol. **5:**46–56.
21. **Lakay, F. M., A. Botha, and B. A. Prior.** 2007. Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils. J. Appl. Microbiol. **102:**265–273.
22. **LaMontagne, M. G., F. C. Michel, P. A. Holden, and C. A. Reddy.** 2002. Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. J. Microbiol. Methods **49:**255–264.
23. **Lefevre, F., et al.** 2008. Drugs from hidden bugs: their discovery via untapped resources. Res. Microbiol. **159:**153–161.
24. **Martin-Laurent, F., et al.** 2001. DNA extraction from soils: old bias for new microbial diversity analysis methods. Appl. Environ. Microbiol. **67:**2354–2359.

24a. **Milton, C., et al.** 2007. PhylArray: phylogenetic probe design algorithm for microarray. Bioinformatics **23:**2550–2557.

25. **Morales, S. E., T. F. Cosart, J. V. Johnson, and W. E. Holben.** 2008. Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. Appl. Environ. Microbiol. **75:**668–675.
26. **Morales, S. E., and W. E. Holben.** 2009. Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by in silico analysis. Appl. Environ. Microbiol. **75:**2677–2683.
27. **Ogram, A., G. S. Sayler, and T. Barbay.** 1987. The extraction and purification of microbial DNA from sediments. J. Microbiol. Methods **7:**57–66.
28. **Rajendhran, J., and P. Gunasekaran.** 2008. Strategies for accessing soil metagenome for desired applications. Biotechnol. Adv. **26:**576–590.
29. **Ranjard, L., E. Brothier, and S. Nazaret.** 2000. Sequencing bands of ribosomal intergenic spacer analysis fingerprints for characterization and microscale distribution of soil bacterium populations responding to mercury spiking. Appl. Environ. Microbiol. **66:**5334–5339.
30. **Rappé, M. S., and S. J. Giovannoni.** 2003. The uncultured microbial majority. Annu. Rev. Microbiol. **57:**369–394.
31. **Robe, P., R. Nalin, C. Capellano, T. M. Vogel, and P. Simonet.** 2003. Extraction of DNA from soil. Eur. J. Soil Biol. **39:**183–190.
32. **Roesch, L. L., et al.** 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. **1:**283–290.
33. **Sanguin, H., et al.** 2006. Potential of a 16S rRNA-based taxonomic microarray for analyzing the rhizosphere effects of maize on Agrobacterium spp. and bacterial communities. Appl. Environ. Microbiol. **72:**4302–4312.
34. **Schloss, P. D., and J. Handelsman.** 2003. Biotechnological prospects from metagenomics. Curr. Opin. Biotechnol. **14:**303–310.
35. **Shendure, J., and J. Hanlee.** 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26:**1135–1145.
36. **Sogin, M. L., et al.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. U. S. A. **103:**12115–12120.
37. **Torsvik, V., J. Goksoyr, and F. L. Daae.** 1990. High diversity in DNA of soil bacteria. Appl. Environ. Microbiol. **56:**782–787.
38. **Torsvik, V., L. Ovreas, and T. F. Thingstad.** 2002. Prokaryotic diversity—magnitude, dynamics, and controlling factors. Science **296:**1064–1066.
39. **Tringe, S. G., et al.** 2005. Comparative metagenomics of microbial communities. Science **308:**554–557.
40. **Van Elsas, J. D., V. Mantynen, and A. C. Wolters.** 1997. Soil DNA extraction and assessment of the fate of Mycobacterium cholorophenolicum strain PC-1 in different soils by 16S ribosomal gene sequence based most probable number PCR and immunofluorescence. Biol. Fertil. Soils **24:**188–195.
41. **Van Elsas, J. D., J. K. Jansson, and J. T. Trevors.** 2006. Modern soil microbiology II. CRC Press, Boca Raton, FL.
42. **Van Elsas, J. D., et al.** 2008. The metagenomics of disease-suppressive soils—experiences from the Métacontrol project. Trends Biotechnol. **26:**591–601.
43. **Vogel, T. M., et al.** 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat. Rev. Microbiol. **7:**252.
44. **Zhou, J., M. A. Bruns, and J. M. Tiedje.** 1996. DNA recovery from soils of diverse composition. Appl. Environ. Microbiol. **62:**316–322.

Supplement figures:

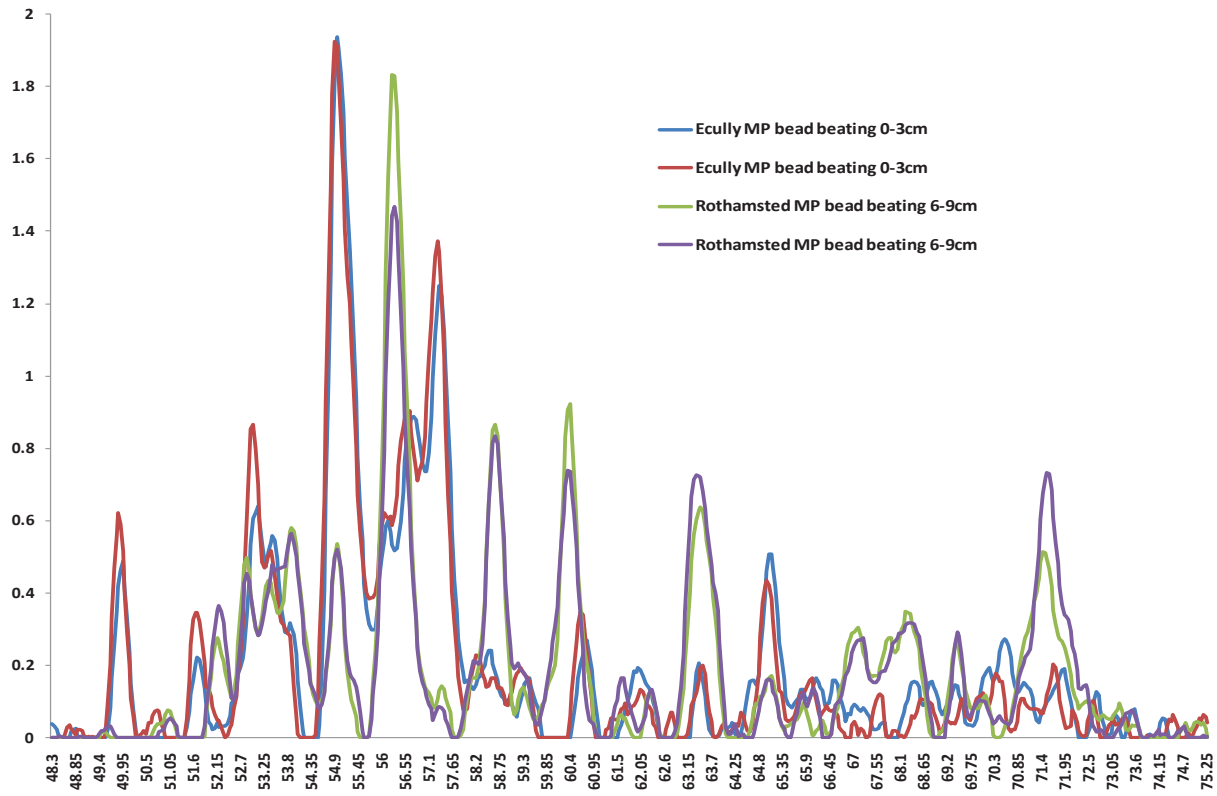


Figure S1: Example of replicates corresponding to different lyses, PCR and RISA analyses.

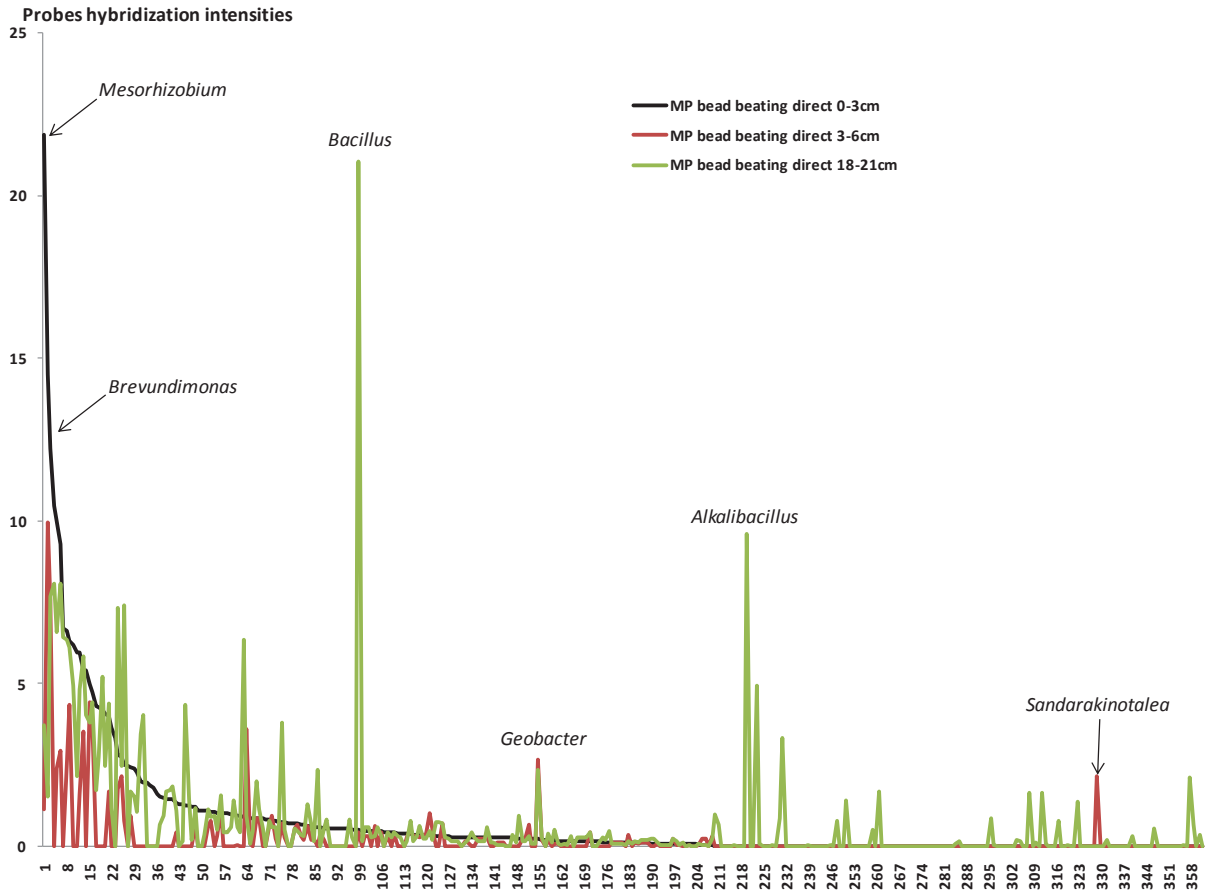


Figure S2: Phylogenetic distribution (genera level) of three DNA pools as a function of the decreasing distribution of the DNA pool extracted with the MP bead beating direct lysis (0-3 cm) protocol (relative proportion based on probes hybridization intensities).

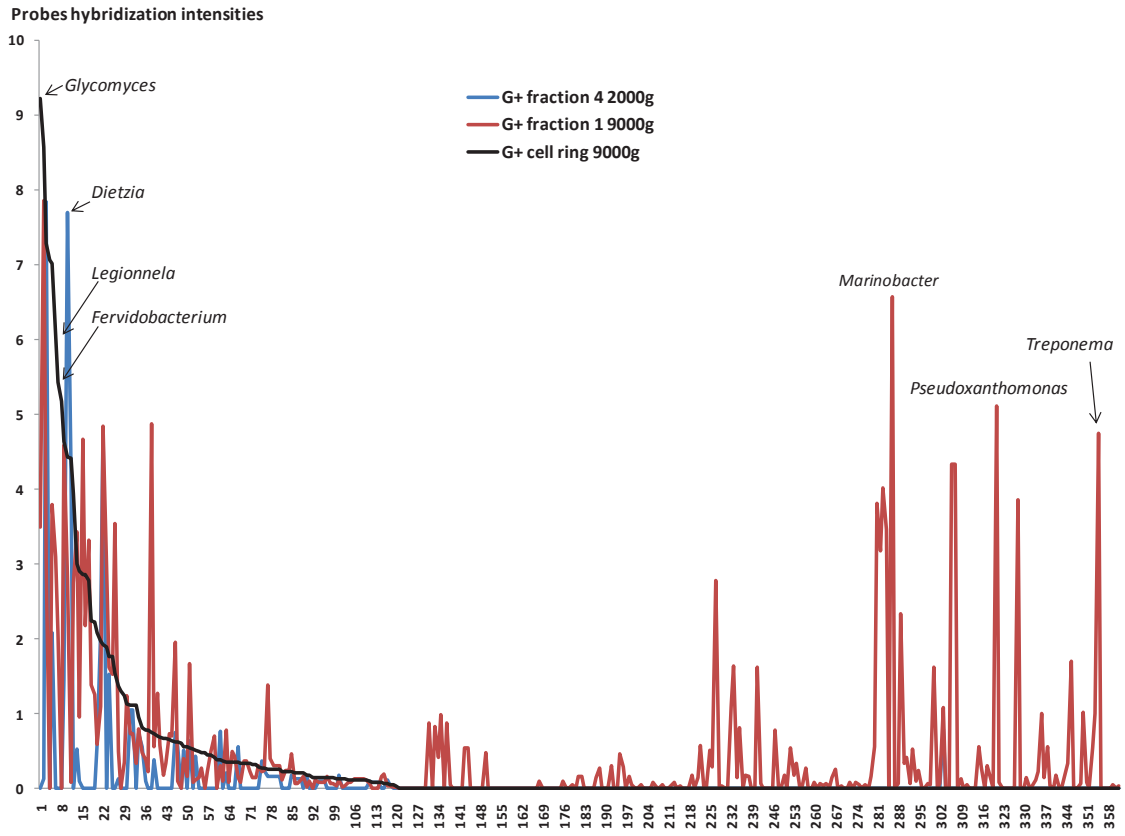


Figure S3: Phylogenetic distribution (genera level) of three DNA pools as a function of the decreasing distribution of the DNA pool corresponding to a cell ring lysed with the Gram positive protocol (relative proportion based on probes hybridization intensities).

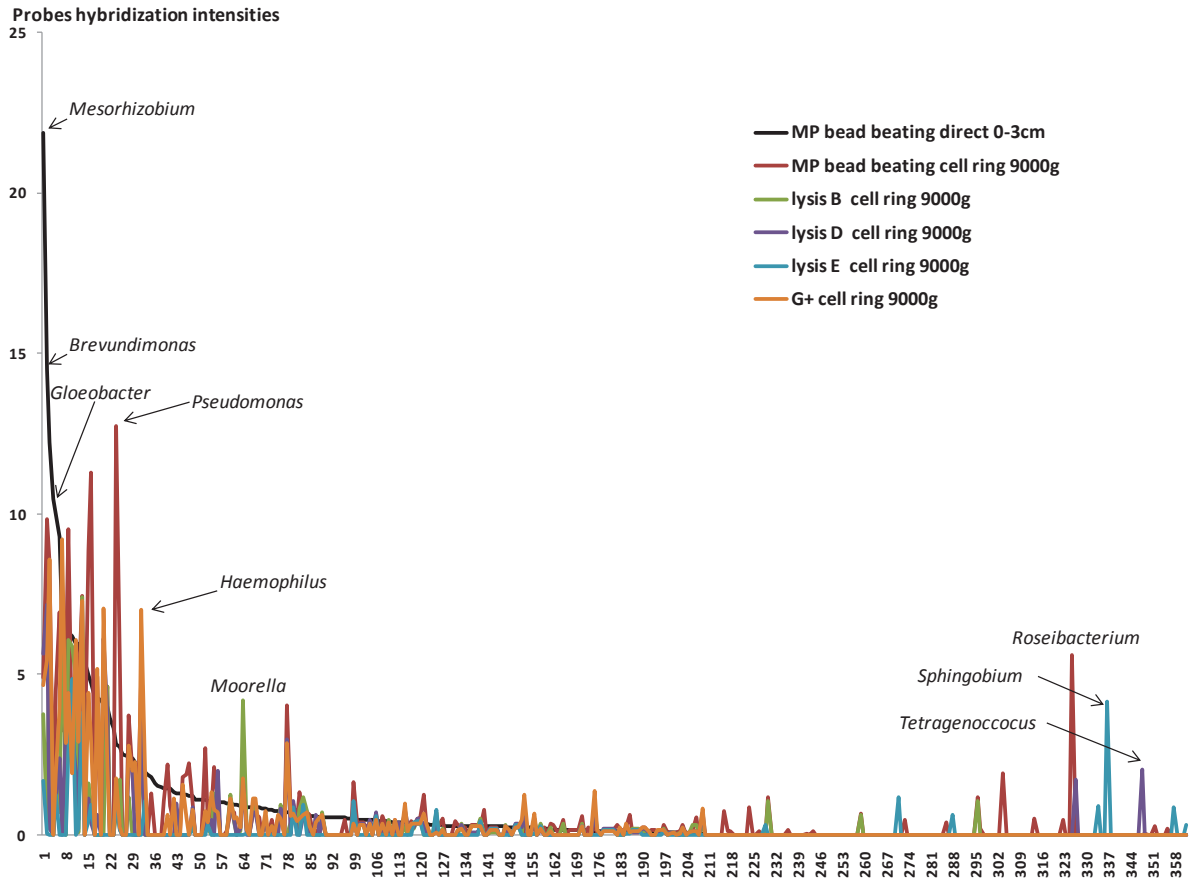


Figure S4: Phylogenetic distribution (genera level) of six DNA pools as a function of the decreasing distribution of the DNA pool corresponding to the MP bead beating direct lysis (0-3 cm) (relative proportion based on probes hybridization intensities).

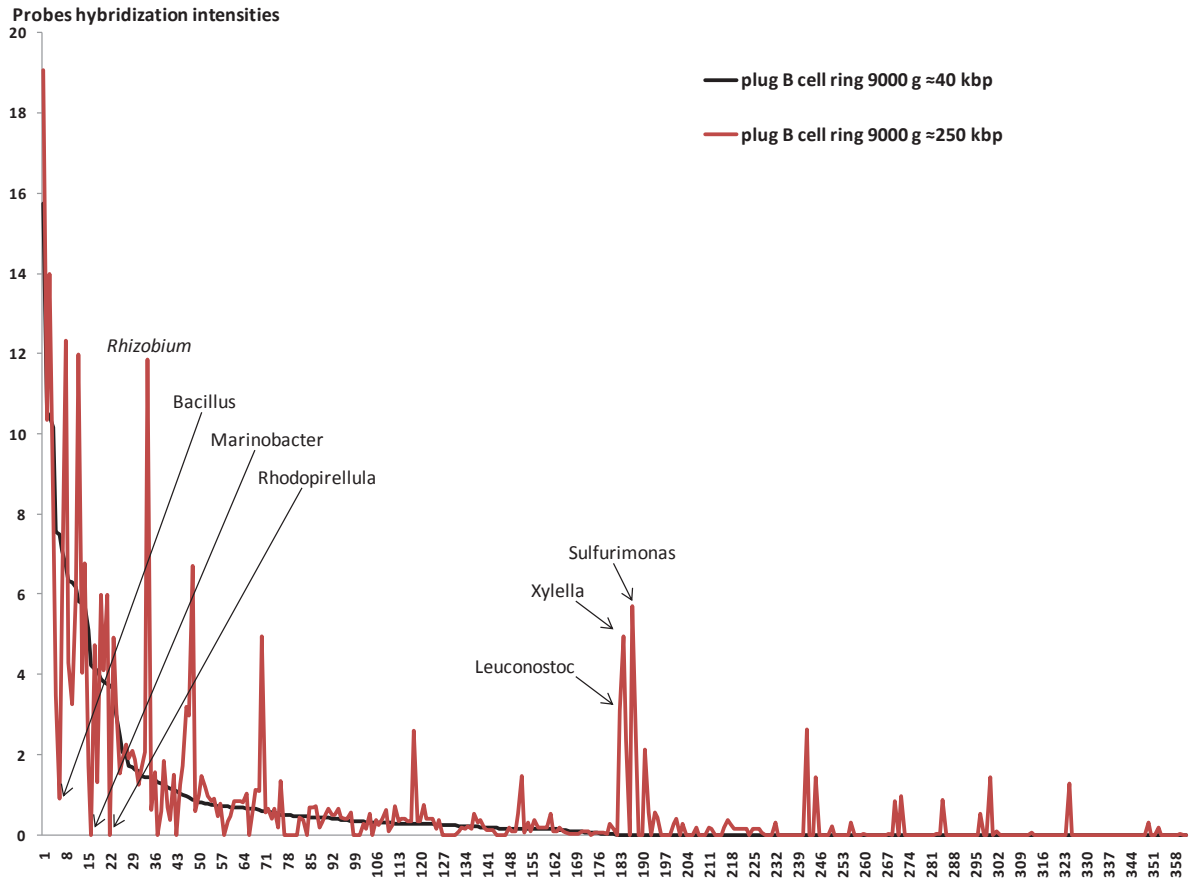


Figure S5: Phylogenetic distribution (genera level) of two DNA pools as a function of the decreasing distribution of the DNA pool corresponding to 40 kb fraction in a PFGE (plug B lysis (0-3 cm)of the cell ring (9000g)) (relative proportion based on probes hybridization intensities).

Structure, Fluctuation and Magnitude of a Natural Grassland Soil Metagenome

Tom O. Delmont[1], Emmanuel Prestat[1], Kevin P. Keegan[2], Michael Faubladier[3], Patrick Robe[4], Ian M. Clark[5], Eric Pelletier[6,7,8], Penny R. Hirsch[5], Folker Meyer[2], Jack A Gilbert[2,9],Denis Le Paslier[6,7,8], Pascal Simonet[1] and Timothy M. Vogel[1*]

Accepted on ISMEj

[1] Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France.

[2] Institute of Genomic and Systems Biology, Argonne National Laboratory, Lemont, IL, 60439, USA.

[3] Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS UMR5558, Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France

[4] LibraGen, 3 rue des Satellites, 31400 Toulouse, France.

[5] Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK.

[6] Commissariat à l'Energie Atomique, Genoscope, 91000 Evry, France.

[7]. Centre National de la Recherche Scientifique, UMR8030, 91000 Evry, France.

[8]. Université d'Evry Val d'Essonne 91000 Evry, France.

[9] Department of Ecology and Evolution, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, U.S.A.

Running title: Grassland Soil Metagenomics

Abstract:

The soil ecosystem is critical for human health, affecting aspects of the environment from key agricultural and edaphic parameters to critical influence on climate change. Soil has more unknown biodiversity than any other ecosystem. We have applied diverse DNA extraction methods coupled with high throughput pyrosequencing to explore $4.88x10^9$ base pairs of metagenomic sequence data from the longest continually studied soil environment (Park Grass experiment at Rothamsted Research in the UK). Results emphasize important DNA extraction biases and unexpectedly low seasonal and vertical soil metagenomic functional class variations. Clustering-based subsystems (CBSS) and carbohydrate metabolism had the largest quantity of annotated reads assigned although less than 50 % of reads were assigned at an E value cutoff of 10^{-5}. In addition, with the more detailed subsystems, cAMP signaling in bacteria (3.24 ± 0.27 % of the annotated reads) and the Ton and Tol transport systems (1.69 ± 0.11%) were relatively highly represented. The most highly represented genome from the database was that for a *Bradyrhizobium* species. The metagenomic variance created by integrating natural and methodological fluctuations represents a global picture of the Rothamsted soil metagenome that can be used for specific questions and future inter-environmental metagenomic comparisons. However, only 1% of annotated sequences correspond to already sequenced genomes at 96% similarity and E values of less than 10^{-5}, thus, considerable genomic reconstructions efforts still have to be performed.

Introduction

Microorganisms first appeared more than $3.5x10^9$ years ago (Allwood, 2006, 714-8), ~1.5 billon years after the formation of our planet. Genetic flexibility over a vast expanse of geological time has enabled microorganisms to adapt to virtually every conceivable ecosystem on earth (e.g. Huber et al., 2007; Pointing et al., 2009; Larose et al., 2010). Among contemporary ecosystems, soil, which is a product of microbial and macrobial life, exhibits the greatest density and phylogenetic diversity per unit volume (Van Elsas et al., 2006; Roesch et al., 2007), with approximately 10^9 cells per gram, comprising a diversity that is estimated to range from thousands to millions of taxa (Knietch et al., 2003).

Soil microbial communities are indispensable for the health of our planet; they drive major geochemical cycles (Falkowski et al., 2001) and help to support healthy plant growth (Ortíz-Castro et al., 2009). Yet, there is still a considerable lack of understanding of the mechanisms of interaction and metabolism that exist among members of the microbial community and their ecosystem. Existing knowledge, concerning the phylogenetic and functional diversity, community metabolic potential, and consequences of evolutionary adaptation, is based largely on partial information gained from studies performed on microorganisms that have been cultivated from soil on a small scale or 16S rRNA gene sequences.

A dependence on studies of cultivable organisms may limit our fundamental understanding of the diversity of interactions in this system. The organisms cultured from soil so far, represent a fraction of the soil biota, e.g. those amenable to growth in controlled laboratory conditions (Schloss and Handelsman, 2003; Davis et al., 2011). Attempts to apply metagenomic methodology to soil samples have been hampered by extreme technical challenges, such as extracting an unbiased and representational sample of genetic material from organisms with very different cell membranes and accessible DNA (Delmont et al., 2011b,c; Demaneche et al., 2008; Ginolhac et al., 2004; Handelsman et al., 1998; Rajendhran and Gunasekaran, 2008). This problem is exacerbated by the uneven spatial distribution of microbial communities in soil ({Grundmann, 2004, 119-127;Ranjard, 2001, 707-16}. Unlike marine systems, which are generally well mixed and amenable to temporal and biogeographic observations (Gilbert et al. 2009, 2010), soil systems surveys have, despite a wealth of valuable data acquired from hundreds of well designed experiments and surveys, uncovered only a fraction of the assumed immense microbial diversity of the soil metagenome (e.g., Tringe et al., 2005; Roesch et al., 2007; Morales et al., 2009). In spite of numerous efforts to

study parameters influencing its diversity using cultural independent approaches (e.g., soil pH or nitrogen fertilisation, Rousk et al., 2010; Ramirez et al., 2010), data from soil are scarcer than those collected from other commonly encountered ecosystems. The only contemporary published soil metagenome (Tringe et al., 2005) contains just 100 million base pairs of DNA, which is potentially a mere millionth of one percent of the genetic material that could be extracted from a gram of soil (based on an assumption of 4 million base pairs per average microbial genome and 10^9 cells per gram of soil). The relative lack of available soil related sequence data presents an interesting paradox, that the most diverse environment on earth has received the least attention from metagenomic analysis (Vogel et al., 2009) although the first soil metagenome dates from 2005. To redress this balance, we have performed an in-depth investigation of a temperate European ungrazed grassland soil metagenome using pyrosequencing technology.

Building on our previous investigations (Delmont et al., 2011b,c), this study describes an unprecedented effort to characterize the microbial diversity and functional potential of a single soil ecosystem that found in the Park Grass Experiment at Rothamsted Research; the location of the oldest agricultural experiments in the world, run continuously since 1856 (Silvertown et al., 2006). In an attempt to explore this unique environment, almost $5x10^9$ base pairs of metagenomic sequence data (Titanium pyrosequencing reads) were produced from soils collected from three depths and at three time points spanning two years. To address concerns regarding the influence of DNA extraction technique bias on microbial diversity (Delmont et al., 2011b), we performed 11 different extraction techniques to improve the diversity of the sequenced microbial genomes. The MG-RAST (Meyer et al., 2008) annotated content of the samples were compared to each other, and to samples of two previously reported, non-soil, environments, so as to place the samples in a more global context.

Material and methods

Soil samples: Samples were collected from the untreated control plot (3d) of Park Grass Experiment, Rothamsted Research, Hertfordshire, UK (Silvertown et al., 2006) in March 2009, July 2009 and July 2010. The overall sample handling is outlined in Figure 1. Soil samples from the top 21 centimeters were collected (Delmont et al., 2011b) by sterile manual corers (10 cm diameter) in plot 3D at random locations, but not where previous samples had been taken, and were placed in sterile plastic bags, sealed and placed on ice 24 hours until

processing. Previous investigations of this soil demonstrated very little horizontal change in diversity, but measureable changes with depth (Delmont et al., 2011b). Hence, the core samples were fractioned into either seven subsamples as a function of the depth (every three centimeters for the direct lysis (described below) and into two depths for the indirect lysis (Delmont et al., 2011c; described below). The aim of this step is to homogenize the quantity of extracted DNA (which decreases with depth) represented in the final pool for each fraction. The different subsamples were then homogenized separately manually by thorough mixing and stored at -20°C for the direct lysis and at 4°C during a maximum period of one week for the indirect lysis. To access rhizospheric microbial communities, a soil core (0-21cm) was sieved (0.2mm) and grass roots were extracted. Soil attached to roots was then recovered in a water column. The column helped the physical separation between roots and soil present at it surfaces. The few grams of recovered soil were then mixed prior to DNA extraction. The metadata for the site and samples are provided in Table S1.

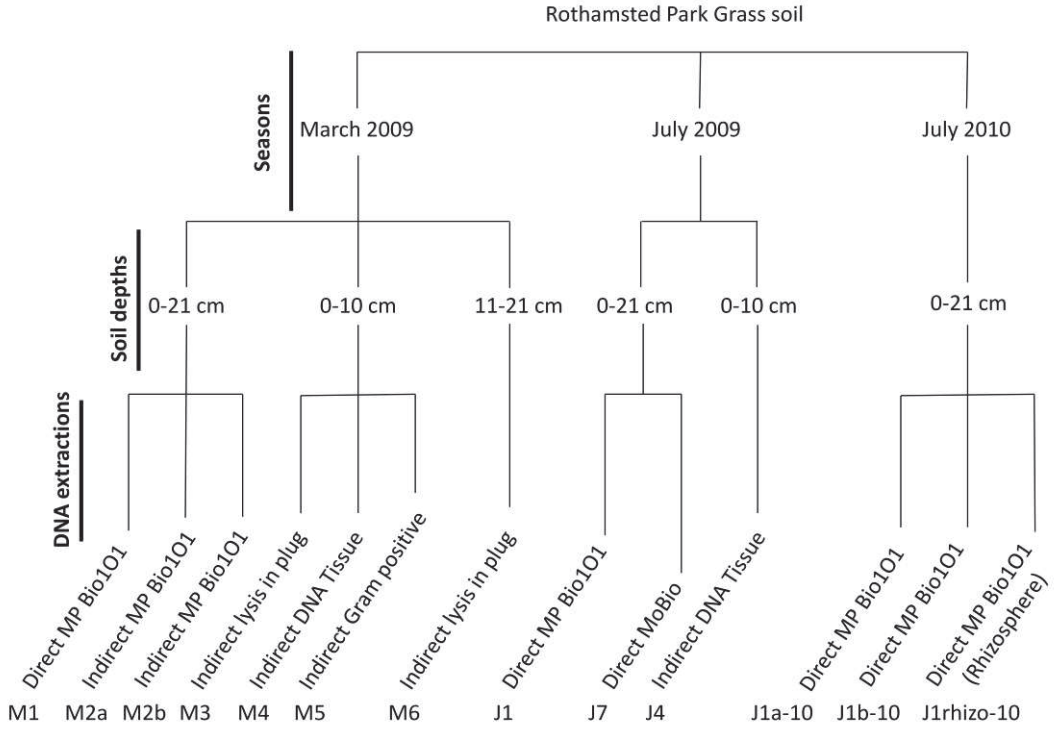


Figure 1. The sampling and DNA extraction schematic for the thirteen pyrosequencing runs. The two pairs, M2a/M2b and J1a-10/J1b-10, are respectively replicate runs from the same DNA extraction and distinct DNA samples extracted sequentially from the same soil sample.

DNA extraction method: Different extraction procedures were used to process the soil samples (Figure 1). We selected DNA extraction methods that use a wide range of approaches

to extract and lyse cells. Among the selected methods, some were already known to provide a high DNA yield (e.g. BIO1O1), or DNA quality or increased DNA length (in plug lysis), others provided a low yield but could still potentially represent a difficult to access microbial communities. The main goal of this experimental design was to create DNA pools with a large variance in order to uncover a wider range of community members within this soil metagenome at both functional and taxonomical levels.

Direct soil lysis: utilized one of two bead beating protocols, (Fast prep MP Bio1O1 Biomedical, Eschwege, Germany) (Griffiths et al., 2000) with 0.5 g of soil. This approach was named "direct MP Bio101" (M1, J1, and replicates J1a10 and J1b10). In addition, rhizospheric soil from July 2010 was extracted with the same protocol (J1rhizo10) and soil from July 2009 was extracted with another bead beating method, the MoBio PowerSoil® DNA Isolation Kit (Carlsbad, USA) (J7). Several different indirect DNA extraction methods were used by first extracting cells on a Nycodenz® gradient gel (density of 1.3) (Bertrand et al., 2005) and then applying one of the following four lyses with the extracted cells: 1) the same bead beating protocol, called "indirect MP Bio101" (replicates M2a and M2b from March 2009); 2) the Nucleospin® Tissue kit, named "indirect DNA Tissue" (M4 and J4, March and July 2009, respectively); 3) the Gram positive kit, named "indirect Gram positive" (M5); and finally 4) a lysis using agarose plugs called "indirect lysis in plug". (M3 and M6 from 0 to 10cm and 11 to 21 cm depths, March 2009, respectively – see figure 1). Plugs were first transferred in 3 ml of G^- lysis buffer (1% lauroyl sarcosine, 500 mM of EDTA Na2, pH 9.5) with 0.5 mg/ml of lysozyme and incubated at 37°C for 12 h. The agarose plugs were then incubated in 3 ml of G^- lysis buffer with 500 µg/ml of proteinase K at 56°C for 12 h, and finally equilibrated in a 10 mM Tris (pH 8.0), 1 mM EDTA storage buffer). This enzymatic lysis was performed in a stable environment (the agarose plug) and was performed without any physical perturbations (*e.g.*, tube mixing that break DNA). This method is generally used to provide high quality and long DNA sequences for the construction of fosmid libraries or for genome size. General information about the different DNA extraction yields used is presented in the table 1.

	Quantity of soil used	Principal type of lysis	average DNA length after extraction	DNA yield per kilogram of soil
MP BIO1O1 rhisosphere soil	0.5 g	Mechanical	10 kbp	40 mg
MP BIO1O1 soil	0.5 g	Mechanical	10 kbp	10 mg
MoBIO soil	0.5 g	Mechanical	10 kbp	2 mg
MP BIO1O1 on extracted cells (Nycodenz)	300-400 g	Mechanical	10 kbp	150 µg
In plug on extracted cells (Nycodenz)	300-400 g	chemio-enzymatic	> 500 kbp	120 µg
DNA Tissue on extracted cells (Nycodenz)	300-400 g	chemio-enzymatic	20-40 kbp	30 µg
Gram positive on extracted cells (Nycodenz)	300-400 g	chemio-enzymatic	20-40 kbp	5 µg

Table 1. Quality and quantity of DNA extracted from the Rothamsted Parkgrass soil with different DNA extraction approaches.

Pyrosequencing runs: A minimum of 10 µg of DNA were used for each Roche/454 pyrosequencing run on a 454 pyrosequencer (GS FLX Titanium Series Reagents ; Roche 454; Shirley, NY, USA). Processing of samples (prior to sequencing) did not involve prior amplification step. For the direct lysis, equal quantities of DNA extracted from the seven fractions from 0 to 21cm were pooled together. J1a10 and J1b10 correspond to distinct extractions from the same soil core. For the indirect approach corresponding to soil from 0 to 21cm, equal quantities of DNA extracted from the two fractions (0 to 10 and 11 to 21cm) were pooled together. For the indirect lysis using the bead beating protocol (0 to 21cm, March 2009), two pyrosequencing runs (M2a and M2b) were performed from the same DNA pool (> 20 micrograms). The sequence data are publically available (http://www.genomenviron.org/Projects/METASOIL.html).

Data analyses: Artificial duplicates were deleted using cd-hit-454 with default parameters (Niu et al. 2010). Sequences were then annotated on the MG RAST (v.02) online software (Meyer et al., 2008). Reads were distributed into different metabolic subsystems. Similarity search between pyrosequencing reads and the SEED database (Overbeek et al., 2005) have been processed with a maximum E value of 10^{-5}. All compared distributions were normalized as a function of the number of annotated sequences for each metagenome. Data corresponding to both functional and taxonomical distributions were then statistically analyzed within the STAMP software (Parks and Beiko, 2010). Fisher's exact tests were performed and annotated functions and taxa with p-values < 0.05 were considered to be significantly different between the different experiments

Tests on assembly productivity were performed using Newbler (Margulies et al., 2005). Newbler was run directly from the ".sff" files produced by the pyrosequencer using the following parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed step: 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap

identity : 90%; Alignment identity score: 2; Alignment difference score: -3. The minimum read length in the data set was 40. Each deeply sequenced dataset was assembled separately using the 454 GS de novo assembler software (Newbler v2.0.00.22,), and all contigs were used for subsequent analysis. In addition, MetaGeneMark (version 2.7d using the parameter file for metagenome gene prediction version 1) have been used to search genes from the 100 largest contigs, and the 1006 genes predicted were analyzed via MG-RAST.

Results

Thirteen pyrosequencing runs were performed with DNA extracted from the Rothamsted Research (Park Grass) site. Grassland soil samples were taken at different depths and three different time points over 1.5 years. DNA was extracted from the samples using 6 different DNA extraction protocols (see materials and methods and figure 1). Two samples were sequenced in duplicate (J1a10 and J1b10, and M2a and M2b) to explore the reproducibility of the metagenomic profile. A total of 12,575,129 reads were generated (length average of 385.9 ± 31.8 bp) and 34.5 (±3.3) % of them were annotated with the MG-RAST online server (E value < 10^{-5}) (Meyer et al., 2008). Based on the protein database used by MG-RAST, 88.64 (±1.44) % of these annotated sequences had closest homology to a protein found in Bacteria, 0.91 (±0.23) % to Eukarya, and 1.41 (±0.16) % to Archaea. Thus, almost 9% of annotated sequences were not classified at the domain level. All the annotated reads were compared to SEED-NR, FIGFams for functional assignments and then used in subsystem reconstructions. The closest matched gene was the source of information about the functional (metabolic) subsystem that the read was binned into and about the "taxa" represented by this read. Therefore, the taxa cited here correspond to the genomes in the database that best matched the given read as long as the E value was smaller than 10^{-5}. Major functions and taxa identified can be found in tables S2 through S4.

Functional comparison:

Functional differences between the 13 datasets generated from Rothamsted, two datasets from other soils, and one from an aquatic environment were derived by exploring the relative number of reads associated with the 835 functional subsystems detected at least in one metagenome (Figure 2).

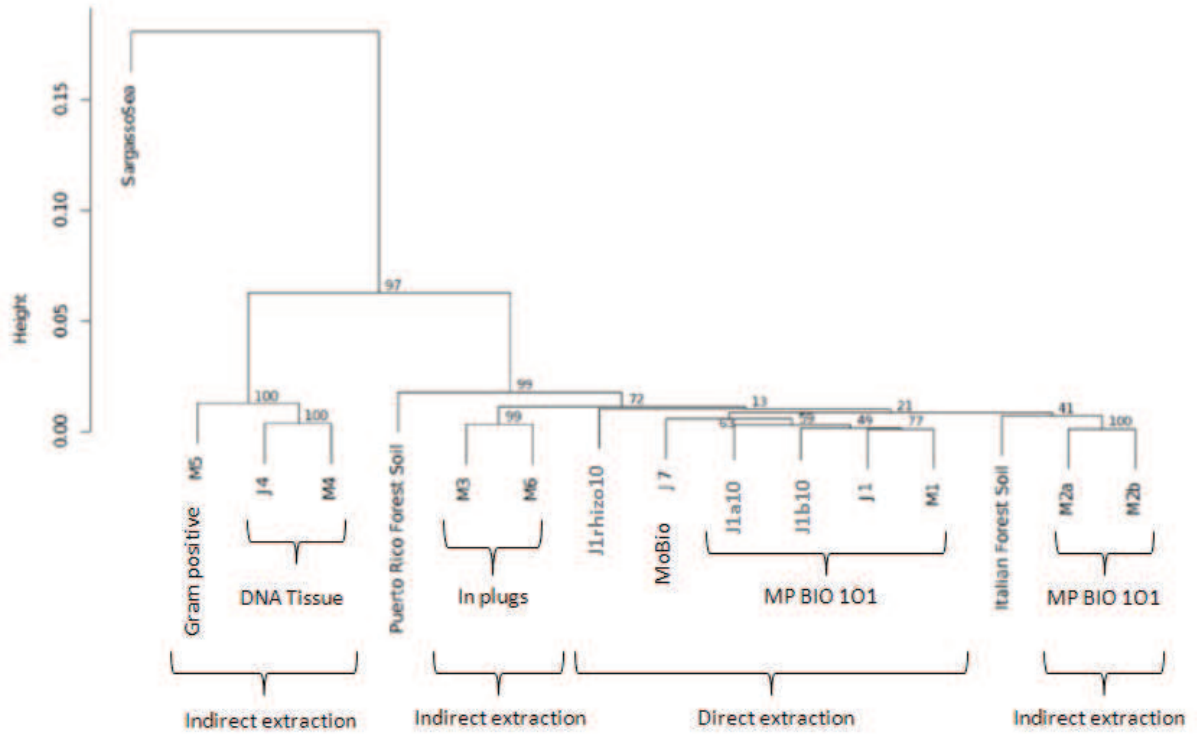


Figure 2. Cluster tree confronting the thirteen pyrosequencing runs, two other soil metagenomes and a metagenome corresponding to Sargasso Sea environment based on the number of reads assigned to each of the 835 metabolic subsystems detected by MG-RAST at least in one dataset. The tree was constructed using Euclidean distances, nPCA ordination method, and complete cluster method.

Bootstrap values are provided. The method of DNA extraction correlates with sample grouping. Samples, M1, J1, J1.a and J1.b were directly extracted using the MP Bio101 kit. Sample J7 which lies in the same general group was extracted directly with MoBio Powersoil kit. The sample from the application of direct MP Bio101 on the rhizosphere soil (J2) is closely associated with this group. The bootstrap values are not particularly high within this group. Three sample pairs on the other hand had significant bootstrap values (>90%) grouping them apart from the other samples: 1) the replicate samples from the application of MP Bio101 to the cells first removed from soil via the Nycodenz gradient (M2.a and M2.b); 2) the two depth samples extracted by indirect lysis in agarose plugs (M3 and M6); and 3) the two samples from different seasons extracted after Nycodenz by use of the DNA tissue kit (M4 and J4). In order to assess the statistical likelihood of the subsystem distribution differences between samples, STAMP software (Parks and Beiko, 2010) based on a bootstrap approach using Fisher's exact tests were applied to the MG-RAST (subsystem functional level 3) outputs. This approach determined what percentage of the 835 subsystems were significantly (at 95% CI) different between any pair-wise comparison. Replicate runs (M2a / M2b and

J1a10 / J1b10) had between 7.3 and 7.7 % dissimilar subsystems and seasonal variations had 8.6 and 11.7 % dissimilar subsystems for direct (M1/J1) and indirect (M4/J4) extractions respectively. When different lysis methods (e.g., M4, M5, and M6) were applied to the bacterial cells removed by Nycodenz gradient gel before DNA extraction, significant differences in subsystem distributions (16.9 – 39.8% dissimilar subsystems) were observed at the 95% CI. Using sequences corresponding to communities extracted from two distinct horizons (0 to 10cm: M3 and 11 to 20cm: M6), 27.01% of the detected functional subsystems possessed statistically different distributions. Two types of geographical comparisons were made. One was between Rothamsted soil and soil from Italy (Vallombrosa forest soil, defined as a Cambric Umbrisol) extracted with the same method (MP Bio101) in our laboratory (the related Italian soil metagenome is represented by approximately 100 000 sequences) and these two soils had 14.1% dissimilar subsystems. The second was Rothamsted sequences compared to those from Puerto Rico (located in the Luquillo experimental forest and defined as a tropical rain forest soil, Metagenome ID of 4446153.3 on MG RAST, one million reads), which were extracted and sequenced elsewhere. They had between 30.98% and 33.13% dissimilar subsystems. The most extreme comparison was between Rothamsted soil and the Sargasso Sea (72% dissimilar subsystems) as indicated also by the distance in figure 2.

Among the major (29) metabolic classes, clustering-based subsystems (CBSS) and carbohydrate metabolism had the largest quantity of annotated reads assigned (Figure 3). Virulence and amino acid and derivatives were next in prevalence (Figure 3). The cluster-based subsystems contain such functions as proteosomes, ribosomes and recombination-related clusters. The virulence subsystem contains diverse functions also, such as resistance to antibiotics and toxic compounds, and pathogenicity islands. Some subsystems were relatively minor such as photosynthesis, prophage, dormancy and sporulation (Figure 3). Although there was significant (at the 95%CI) differences in the distribution of reads in some (from about 7 to 40%) of the different metabolic subsystems from the different pyrosequencing runs of DNA extracted from Rothamsted soil, the standard deviation around the mean of all of the pyrosequencing runs varied between 2 and 50 %, with the higher variance for the metabolic classes with relative few assigned reads (*e.g.*, macromolecular synthesis; error bars in figure 3).

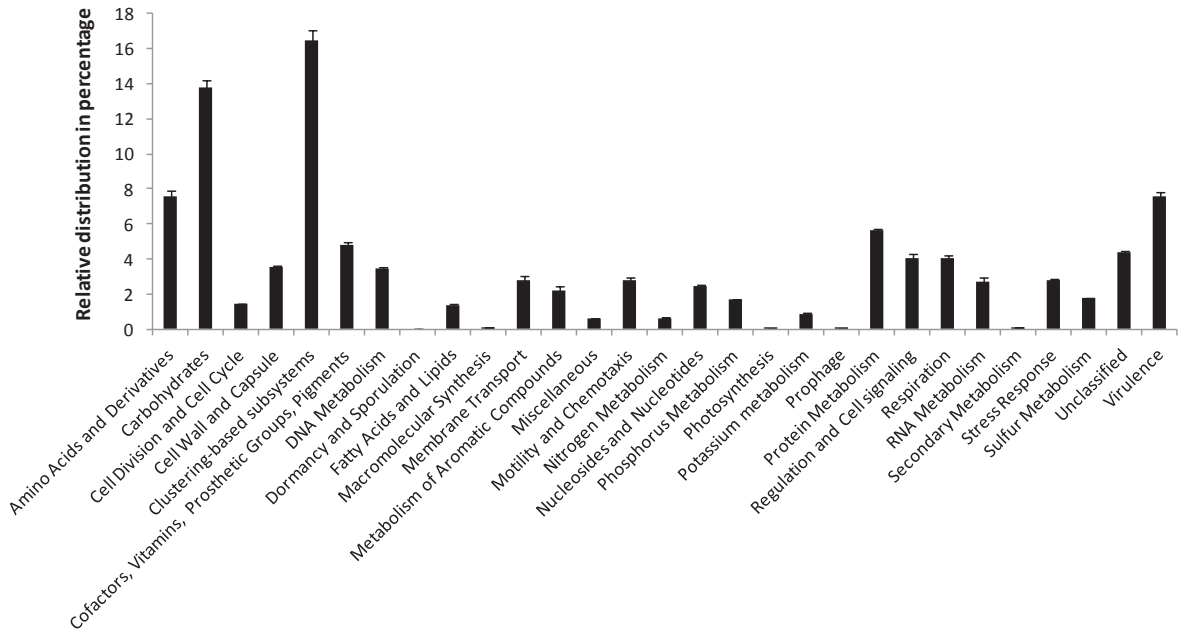


Figure 3. Relative distribution (in percentage of annotated reads) of the 29 major metabolic subsystems (using SEED subsystems in the MG-RAST program) detected in the Rothamsted soil metagenome. Standard deviations correspond to the variability among sequencing runs. The stars represent the relative distribution among the 100 largest contigs after assembly.

When comparing the assigned reads at a finer functional subsystem classification within MG-RAST, the most prevalent subsystem (out of the 835 different categories) in the soil was the cAMP signaling in bacteria with 3.24 ± 0.27 % of the annotated reads (Table S2). The next most prevalent subsystem was the Ton and Tol transport systems at 1.69 ± 0.11% of the annotated reads (Table S2). These prevalent systems varied less than 10% between DNA extraction pools except for the distribution of CO_2 uptake carboxysome related genes, which varied from 0.56% in M3 to 1.43% in M4, which represents an increase of 60.7% (average for the thirteen pyrosequencing runs was 0.70 ± 0.42 % of reads).

Taxonomic comparison:

The 56 ± 4.4% of annotated protein sequences showed closest homology to a total of 1214 unique taxa using the taxonomic annotation of functional SEED subsystems. The most dominant putative taxon was *Solibacter usitatus* (6.72 ± 0.29 % of annotated reads). Other taxa with relatively high number of assigned reads were *Blastopirellula marina (*4.96 ± 2.88%), *Bradyrhizobium japonicum (4.89* ± 0.635%) and *Acidobacteria bacterium (3.64* ± 0.94%) (Table S3). The legitimacy of the read assignment at an E value of 10^{-5} cut-off is provided in part by the E value distribution of the different reads assigned to the reference genome. In the case of *Bradyrhizobium japonicum*, the majority of assigned reads had E values lower than 10^{-30} and in the case of *Blastopirellula marina*, the E values were in general larger than 10^{-30}. Using the taxonomic classification of functional gene fragments, it is possible to use all annotated reads to determine community structure (Figure 4).

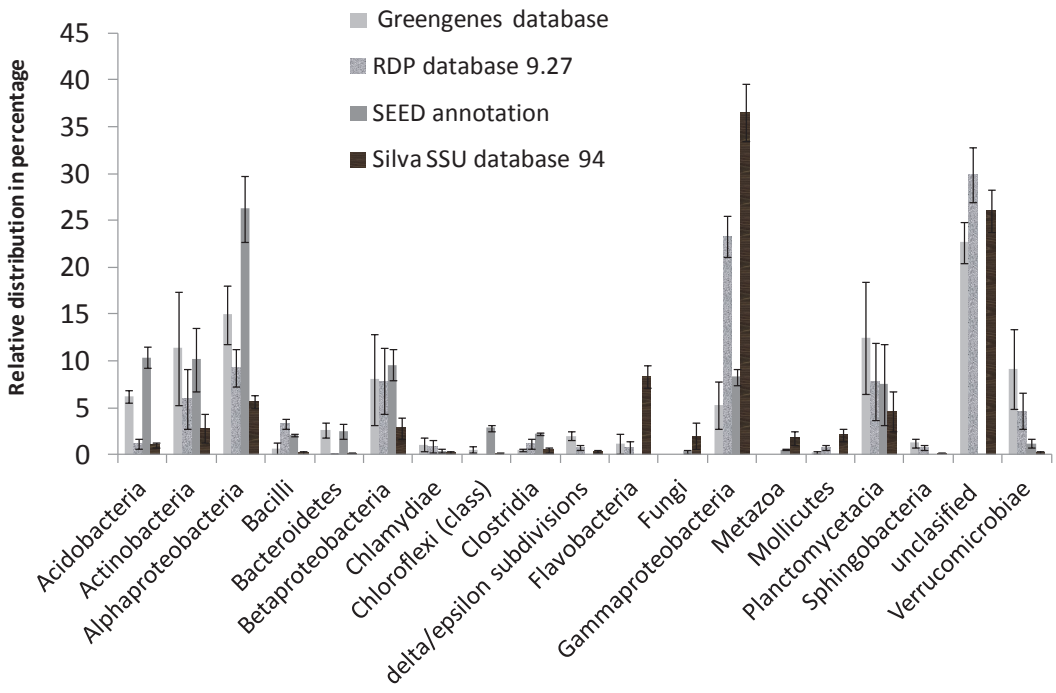


Figure 4. Relative distribution of microbial classes in the Rothamsted soil metagenome. Standard deviations correspond to the fluctuation of the relative distribution between different pyrosequencing runs. The total number of reads annotated by the different methods is not the same as the SEED annotation using all annotated reads and the others use only identified 16S rRNA genes (*rrs*). The version of Greengenes database used within MG-RAST was from 2008. The stars represent the relative distribution among the 100 largest contigs after assembly based on SEED annotation.

However, it is also possible to use 16S rRNA sequences to determine the community structure, although the number of reads is considerably less than for the SEED annotation. Three different databases accessible within the MG-RAST platform and used with the MG-RAST software were used to determine community structure with standard deviations calculated from the variance of the 13 different pyrosequencing runs (Figure 4). While there is a general agreement: alpha-, beta- and gammaproteobacteria and Actinobacteria dominate all four methods, there are some important differences in the relative number of reads in different classifications. For example, the Silva SSU database 94 has a much higher percentage of reads in Flavobacteria than the other systems (Figure 4). In order to use functional genes other than 16S rRNA for taxa or at least genera identification, a more accurate and limited analysis constrained the similarity at 96% or better (still with an E value of 10^{-5}). When this was performed using SEED, only 0.35 ±0.09% of the total reads (or about 1% of the annotated reads) were used to identify bacterial taxa (Table S4). The most abundant taxa identified from Rothamsted soil were members of the *Bradyrhizobium*, *Rhodopseudomonas* and *Nitrobacter* genera (Alphaproteobacteria); the *Solibacter* and *Acidobacteria* genera (Acidobacteria) and *Pseudomonas* (Gammaproteobacteria) and *Burkholderia* (Betaproteobacteria) genera. *Blastopirellula marina* was no longer associated with any of the reads.

Soil metagenome assembly:

Sequence data were assembled to provide a metric describing the depth of sequencing applied to the community metagenome; in part this was used to estimate the minimum quantity of sequencing required to completely sequence all the members of the soil microbial community. The extreme minimum could be considered as the quantity of sequences where no singleton is left unassembled, even if practically this minimum is insufficient to assemble all the genomes. Ten random read subsamples of increasing metagenome size (read quantity) were run through the Newbler assembler (Table S5). No attempt was made here to optimize the assembly process. The ten subsamples ranged in size from 1257242 to 12572342 reads (with 487554794 to 4874169257 total number of bases) and produced from 7478 to 266600 contigs (Table S5). The largest contig size increased from 6361 bp to 22645 bp with three times as many reads but then decreased and leveled off at about 15400 bp with increasing number of reads (Table S5). The fraction of reads that were not included in any contig ("singletons") fell

from roughly 0.93 to 0.76 when increasing the number of reads ten-fold (Figure 5A insert).

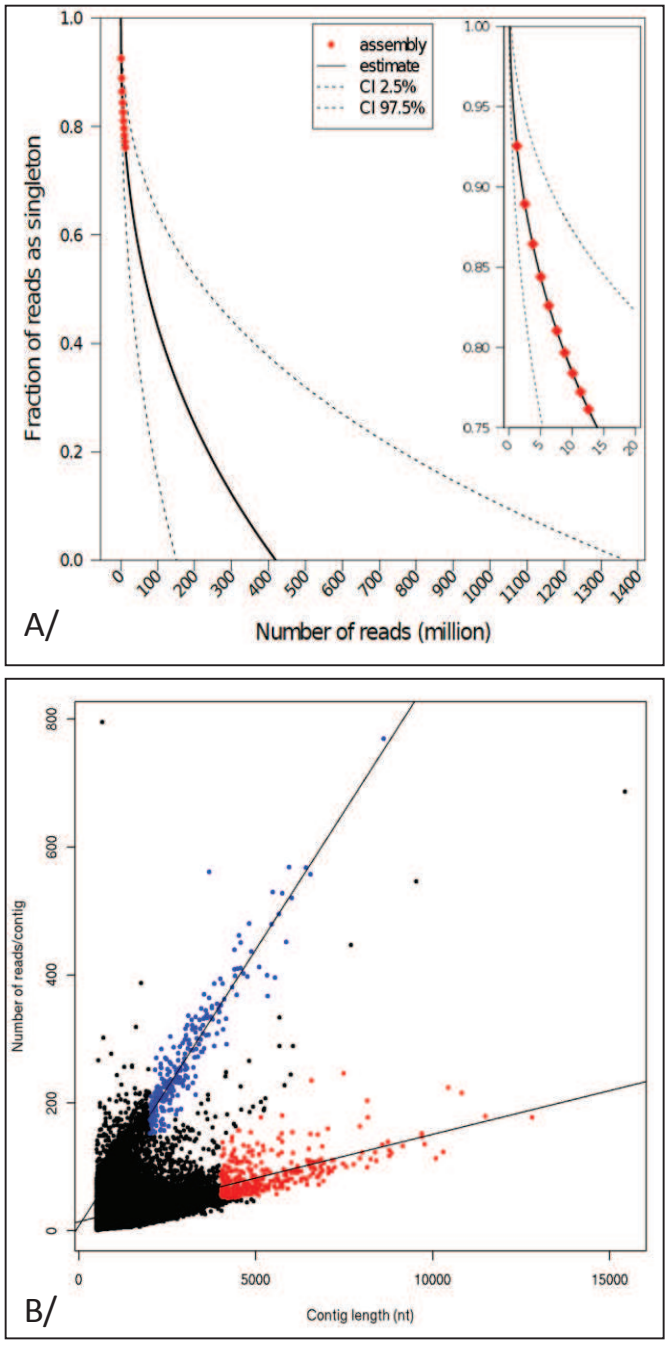


Figure 5. Panel A. Relation between number of 454 sequence reads used in the Newbler assembler and the percentage of reads not combined with any other reads (singletons). A best fit equation for this relationship is: pSingleton = a*[nbReads]^{b} + c with the following four parameters: Estimated value, Std. Error, t value, Pr(>|t|) - for a: -6.714x10^{-4} , 5.409x10^{-5} , -12.41, 5.06x10^{-6}; for b: 3.703x10^{-1}, 4.446x10^{-3} , 83.30, 9.46x10^{-12}; for c: 1.047, 2.372x10^{-3} , 441.56, < 2x10^{-16}. Panel B. Plot of the number of reads per contig as a function of the length of the contigs produced with all the reads from the 13 pyrosequencing runs using the 13 pools of DNA extracted from the Park Grass soil at Rothamsted Research.

This data was fitted and extrapolated to the point where no read would be orphaned. This extrapolation was at about 400 million 454 reads (average of 386 bp in length) with the 95% confidence interval stretching from less than 200 million reads to almost 1400 million reads (Figure 5A).

The maximum contig length did not continue to increase with increasing read number, but the number of reads per contig did develop two general trends (Figure 5B). These two trends are schematically represented by the two lines in figure 5B. The denser trend has a slope represented by a contig coverage of about 30x (when the assembler needs/uses 30X to build the contigs) and the smaller trend has a contig coverage of about 4.5x. Contigs from these two trends were selected, broken in coding sequences by MetaGeneMark and then annotated using MG RAST. Globally, the trend corresponding to coverage of 30x possessed more sequences related to Firmicutes (10.99%) and Verrucomicrobia (21.85%). In contrast, the trend corresponding to low coverage assembled contigs (4.5x coverage) possessed a majority of sequences related to Proteobacteria (66.06%). Independent of the two observed trends, the 100 largest contigs created from the entire sequence pool were also annotated by MG RAST and in general the relative proportion of different functional and phylogenic classes (stars in figures 3 and 4) were similar to that for the sequences directly with some exceptions. There were fewer virulence subsystem hits and significantly more fatty acids and protein metabolism hits.

Discussion

Soil is one of the most diverse environments on earth and the depth of the microbial diversity is still poorly understood. High throughput sequencing technologies, coupled with appropriate DNA extraction methods, provide a means to explore the soil ecosystem with an unprecedented level of detail (Vogel et al, 2009). In this study, pyrosequencing from 13 samples generated nearly $5x10^9$ base pairs of sequence data with average read size of 386 bp. Three key parameters were varied: soil depth, sample collection season, and DNA extraction method. Sequence samples were annotated with the MG-RAST online server, revealing broad functional (835 of 878 possible functional subsystems) and taxonomic (detection of 1214 putative taxa) diversity in the Rothamsted Park Grass soil metagenome.

The most abundant functional subsystems in the Rothamsted soil were seemed to be related to microbial cAMP signaling and Ton and Tol transport (Table S2). The same subsystems were prevalent in metagenomes in soil at Waseca farm, in Puerto Rico and Italy. These trends in soil functional content are robust enough to be observed on a global scale. cAMP is an important secondary messenger in Eukarya and Bacteria. cAMP is a universal cell energy/metabolism regulator as well as being involved with cell-cell signaling. Soil bacteria might have to deal with frequently fluctuating substrate levels so that they would need extra regulation rather than interacting with plants. Interestingly, since cAMP is also a subversion mechanism, some bacterial pathogens might also subvert plant cAMP production for their own benefit, through injection of adenylate cyclase and/or various toxins that alter adenylate cyclase levels (adenylate cyclase is essential to the production of cAMP) (Agarwal et al., 209; Akhter et al., 2008). Iron is an essential element for most organisms (Weinberg, 1984), but can be a limiting reagent for life (often in oceans, Boyd et al, 2007) due to its insolubility in aerobic environments at neutral pH. In response to this stress, some bacteria possess high-affinity transport systems (Crosa et al., 2004) and generate high-affinity siderophores that complex extracellular iron (Neilands et al, 1980) to optimize its acquisition. The presence of Ton related proteins in the soil is likely due to TonB, an energy dependent cell envelope protein that assists iron uptake through accommodation of ferric siderophores, too large to cross porins, through the outer membrane (Klebba et al, 2003).

MG-RAST annotation also revealed the presence of several highly abundant cluster-based subsystems (CBSS). These are groups of functionally coupled genes (genes found proximal to each other in the genomes of diverse taxa) whose functional attributes are not well understood. The relatively high abundance of these subsystems across all Park Grass samples, as well as the other sequenced soils, suggests that they play key roles in soil ecosystems across the globe, and should be explored in future research efforts to understand the composition of soil ecosystems. The CBSS-258594.1.peg.3339, CBSS-269799.3.peg.2220, CBSS-83332.1.peg.3803, CBSS-249196.1.peg.364 (Table S2) are thought to be a galactoglycan biosynthesis, a molybdenum oxidoreductase, a PKS-related, and a fatty acid metabolism subsystem, respectively.

The comparison of the runs corresponding to the same DNA sample (M2a/M2b) provided important information about the reproducibility of pyrosequence generation in highly biodiverse environments. The Fisher's exact test operated by the STAMP software did identify some functions (about 7%) and taxa that varied significantly (at the 95% CI) between

replicates. The lower p-value was on the order of 10^{-7} when comparing M2a and M2b at the functional level, so some comparisons between seasons and depths were possible. Based on these observations, functional comparisons having at most a minimum p-value of 10^{-8} (cut-off based on the observed technological reproducibility) were considered to have distributions that varied significantly. Unfortunately, the technological reproducibility is not the only limit for robust metagenomic comparisons. Even if a stringent p-value is used, the DNA extraction approach influenced the experimental conclusions. When comparing the seasonal effect by using two different extraction approaches (direct:M1/J1 and indirect M4/J4), some differences in relative predominance of different subsystems were found. Based on the comparison of M1 and J1, sequences related to the type 4 secretion and conjugative transfer and cellulosome subsystems are more represented in March (p-value of 10^{-8} in the two cases). When comparing M4 and J4, the cellulosome subsystem is still detected more in March (p-value<10^{-15}), but the type 4 secretion and conjugative transfer is not. In contrast, sequences related to bacterial cAMP signaling are more present in July (p-value of 10^{-12}), but only when comparing M4 and J4. Thus, only sequences related to cellulosome dominated one season's metagenome independent of the extraction method applied. Major environmental difference between the two studied seasons was temperature (from 6°C in March to 16.6°C in July). In addition, snow lay on the ground for weeks in February of the same year, thus limiting active grass growth. As a consequence, soluble root exudates were possibly in short supply during this relatively cold period and cellulosome from root residues would be the main source of carbon and energy supporting soil microbial communities.

On the other hand, depth had more effect with sequences related to genes involved in bacterial chemotaxis, Ton and Tol transport systems, flagellum mechanism, D-ribose and L-Arabinose utilization represented more in the surface sample (0 to 10 cm) and sequences related to selenocysteine metabolism and tRNA aminoacylation represented more at depth (11 to 20 cm). However these results were generated using only one DNA extraction method. In comparison to depth and seasonal variables, the extraction method was able to influence functional distributions (Figure 2), especially when using methods with striking differences in cell lysis (e.g., Gram positive kit versus in agarose plug lysis or DNA tissue). Thus, the stringency of lysis appears to be a crucial step for soil metagenomic analysis, confirming previous results with RISA and phylogenetic microarray analyses (Delmont et al, 2011b).

In addition, when studying the distribution of sequences based on their G+C%, clear variations were found among the different runs. Direct lysis versus indirect lysis had more

impact on the G+C% profile than any other variable. The indirect lysis provided more sequences possessing a higher G+C ratio (from 60 to 72%), while the direct lysis had a more even distribution with more sequences in the 50 to 58 G+C% range (Figure S1). Both metagenomic standard deviations and G+C% ratio profile fluctuations are limited by the experiments and variables used. However, this effort provides both significant soil metagenomic sequences and data useful to appreciate methodological differences in microbial community diversity accessibility.

Given the relatively low functional subsystem variations between different soils (figure 2), soil microbial community metagenomes from Rothamsted, Puerto Rico, Italy and the Waseca farm soil (Tringe et al., 2005) could be compared to metagenomes from oceans and human feces.

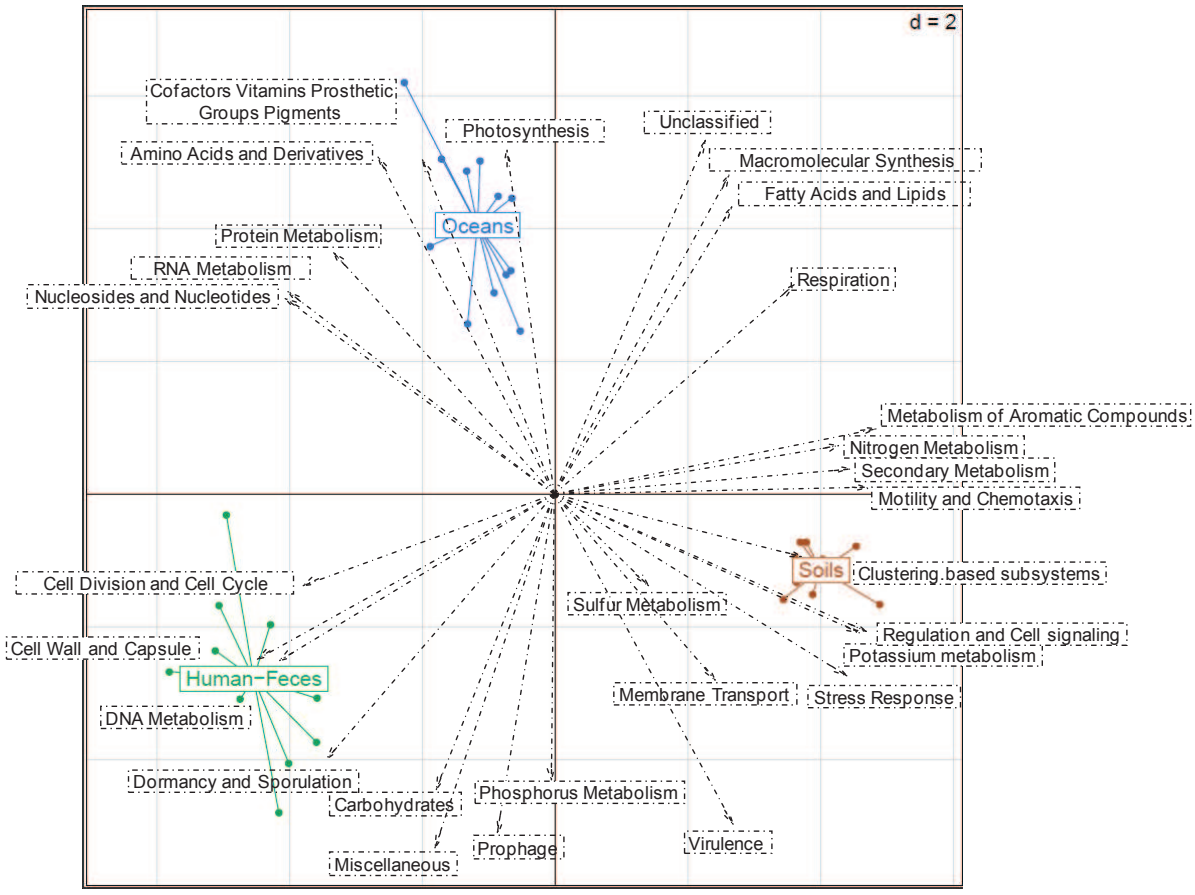


Figure 6. The principal component analysis of three ecosystems using the relative distribution of reads in the different metabolic subsystems for the metagenomic sequences available in the public database in addition to those produced here. The large metabolic classes as determined by MG-RAST are mapped on the same PCA as the ecosystems.

This comparison might help identify some of the soil ecosystem unique functional attributes. In order to make the comparison, principal component analysis was generated based on the

distribution of general functional subsystem classes with metagenomes publically available from these ecosystems (Figure 6). Some general functional classifications appear to be relatively more represented in one ecosystem in comparison to the others. Sequences related to RNA and protein metabolism, photosynthesis, fatty acids and lipids, and macromolecular synthesis are more highly represented in ocean metagenomes. In contrast, phosphorus metabolism and virulence are less represented in ocean metagenomes than in those sequenced for soil and human microbiomes. Sulfur and potassium metabolism, membrane transport, stress response and regulation, and cell signaling are more represented, and nucleosides and nucleotides, and RNA and protein metabolism are less represented in soil metagenomes. In human microbiomes, cell division and cell cycle, DNA and phosphorus metabolism, cell wall and capsule, dormancy and sporulation, carbohydrates are more represented than in those of oceans and soils (Figure 6). When comparing the taxonomical structure of these metagenomes, Cyanobacteria and Bacteroidetes appear to be more represented in the oceans. In addition, Eukaryotic sequences were also detected and represent additional specificities of these metagenomes (Figure S2). Actinobacteria, Chloroflexi, Fibrobacteres and Acidobacteria group, Planctomycetes, and Synergistetes are more present in soils. Chlorobi, Firmicutes, Spirochaetes, Fusobacteria and the Bacteroidetes Chlorobi group are clearly relatively dominant in human digestive tracts. In contrast, Proteobacteria are more present in oceans and soils. The metagenomes are clearly grouped as a function of the environment based on both general functional and taxonomical distributions. So in spite of important DNA extraction biases and sequencing technology differences (Illumina, Pyrosequencing and Sanger), global metagenomic comparisons are possible and provide unique information about the functional and taxonomical differences of each environment (Delmont et al., 2011a). As an example, sequences related to metabolism of aromatic compounds are more abundant in soils possibly due to the presence of these compounds in this environment. However, additional comparisons, such as qPCR and metatranscriptomics, need to be performed to confirm which taxa and functions are unusually active in soil to gain a better understanding of soil microbial community function.

The relative percentage of orphan reads decreased continually when accumulating pyrosequences. Therefore, an estimate of the number of reads needed to avoid having orphan reads would possibly provide the absolute minimum number of reads needed to sequence the entire soil metagenome. Rarefaction analysis of this sequencing effort (Figure 5) indicated that the equivalent of about 320 Titanium runs would be required to create contigs from all of

the soil pyrosequence reads generated. Of course, chimeras might be generated due to the complexity of communities, and a much larger effort would be needed to assemble the soil metagenome, but as new efficient high-throughput sequencing technologies and valuable assembling tools are developed, this goal will become less utopic. Genomes from Proteobacteria might be assembled more rapidly than those from Firmicute or Verrucomicrobia phyla. The presence of regions that limit assembly (e.g., insertion sequences regions) and the complexity of diversity among taxa might explain in part the efficiency differences observed between these phyla (4.5x and 30x), but additional experiments are needed to understand the two trends observed in the figure 5B.

Conclusion:

In this study, more than 12 million reads were generated from the soil of the Rothamsted Research Park Grass experiment. These sequences were generated in 13 separate sequencing runs producing over $4x10^9$ bp. The results demonstrated both some DNA extraction biases and relatively low seasonal (when comparing March and July months) and vertical soil metagenomic functional class fluctuations. In addition, this approach provided a statistical view of functional distributions in this soil. This metagenomic study increased our knowledge about soil microbial communities at a metagenomic level by integrating both natural and methodological fluctuations. The metagenomic variance so generated represents a global picture of the Rothamsted soil metagenome that can be used for specific questions and future inter-environmental metagenomic comparisons. However, only 34.5 % of the reads were assigned to functions and less than 1% of annotated sequences correspond to already sequenced genomes (at 96% similarity), therefore, many soil microorganisms remain elusive and genome constructions are needed.

Acknowledgements: T.O.D. was supported by the Rhône-Alpes Region. We want to thank the French National Research Agency (ANR) for financing Metasoil (Projet ANR-08-GENM-025) and the European Union (7^{th} Framework KBBE-2007-3-3-05) funding for Metaexplore (22625) project.

Supplement figures:

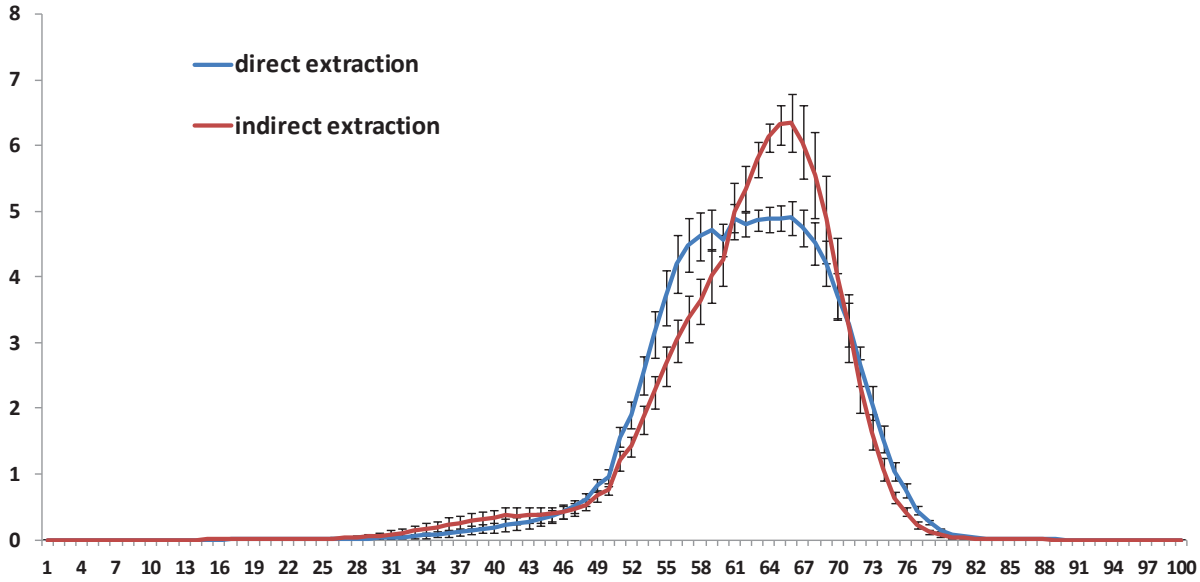


Figure S1: The average and standard deviation of the GC percentage profiles of the pyrosequence reads from all the direct and indirect DNA extractions of the Rothamsted Park Grass soil. There are 6 direct and 7 indirect DNA extractions and associated pyrosequencing runs used.

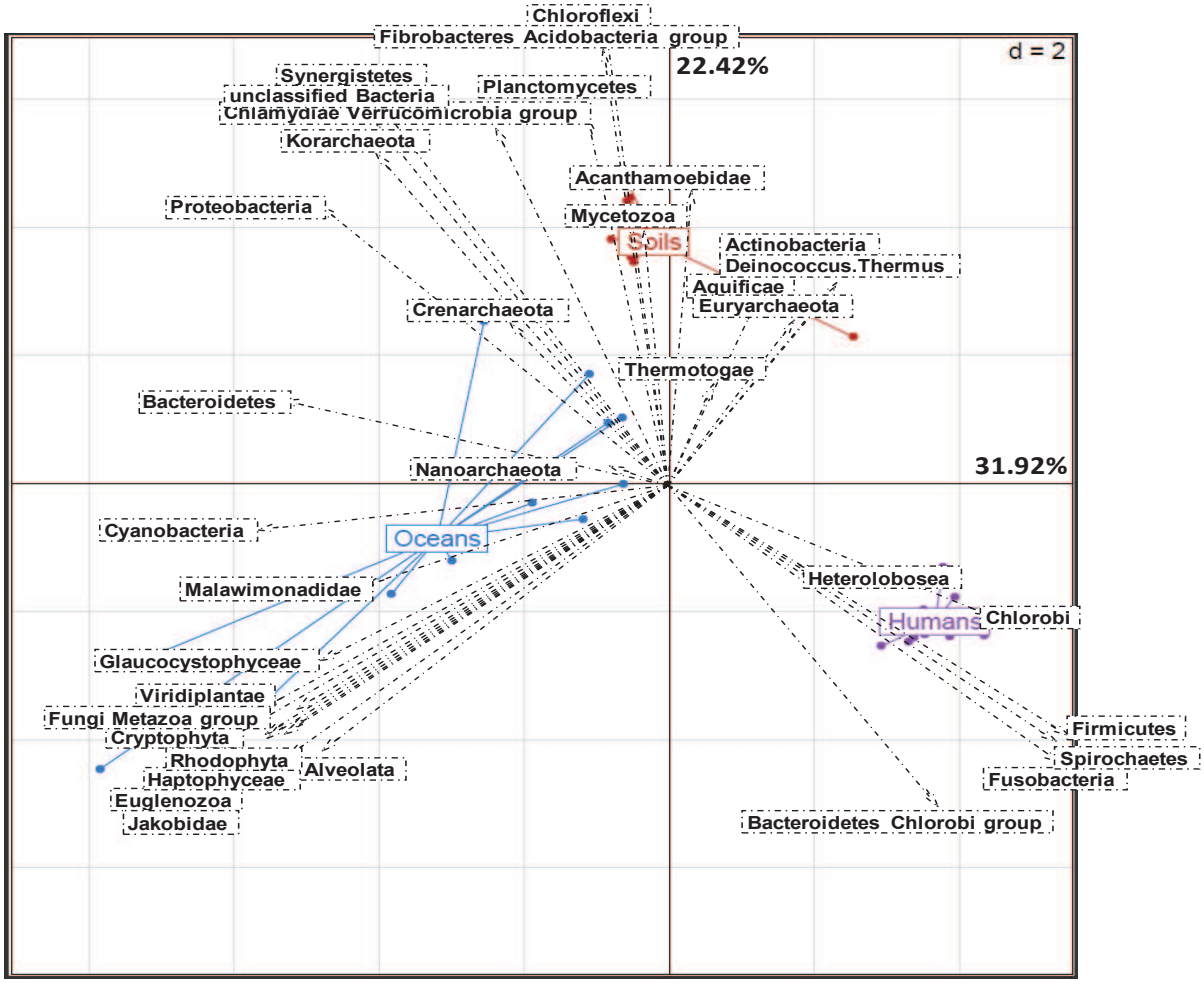


Figure S2: The principal component analysis of three ecosystems using the relative distribution of reads in the different phylogenetic data based on the SEED annotation for all the metagenomic sequences available in the public database in addition to those produced here.

Table S2. The relative proportion of annotated reads (and their standard deviation between the 13 DNA pools) that are classified within a functional class with an E value of 10-5 or less by the MG-Rast annotation system.

Functions	Distribution (%)	SD (%)
cAMP signaling in bacteria	3.243	0.267
Ton and Tol transport systems	1.686	0.111
Phosphate metabolism	1.421	0.038
Bacterial Chemotaxis	1.321	0.099
Cobalt-zinc-cadmium resistance	1.306	0.055
CBSS-258594.1.peg.3339	1.215	0.036
tRNA aminoacylation	1.104	0.037
DNA-replication	1.074	0.064
Serine-glyoxylate cycle	1.065	0.085
CBSS-269799.3.peg.2220	1.056	0.595
CBSS-83332.1.peg.3803	1.009	0.101
CBSS-246196.1.peg.364	0.931	0.070
Glutathione-regulated potassium-efflux system*	0.860	0.047
CBSS-316057.3.peg.1308	0.824	0.044
ABC transporter branched-chain amino acid*	0.780	0.135
Flagellum	0.778	0.124
Maltose and Maltodextrin Utilization	0.759	0.079
Methionine Biosynthesis	0.752	0.035
DNA repair, bacterial	0.744	0.019
Iron-sulfur experimental	0.726	0.033
CO2 uptake, carboxysome	0.704	0.422

Table S3. Species with sequenced genomes that were the closest match with annotated 454 reads from Rothamsted Park Grass soil. Percentage of annotated reads and standard deviation among the 13 pyrosequenced DNA pools that matched with each genome at E= 10^{-5} or less.

Species	Distribution (%)	SD (%)
Solibacter usitatus Ellin6076	6.720	0.292
Blastopirellula marina DSM 3645	4.961	2.878
Bradyrhizobium japonicum USDA 110	4.890	0.635
Acidobacteria bacterium Ellin345	3.642	0.943
Pirellula sp. 1	2.522	1.397
Bradyrhizobium sp. BTAi1	2.499	0.447
Sorangium cellulosum So ce 56	1.875	0.280
Mesorhizobium loti MAFF303099	1.194	0.165
Akkermansia muciniphila ATCC BAA-835	1.181	0.467
Nitrobacter hamburgensis X14	1.128	0.235
Rhizobium leguminosarum bv. viciae 3841	1.056	0.099
Rubrobacter xylanophilus DSM 9941	0.966	0.225
Frankia sp. EAN1pec	0.945	0.314
Rhodopseudomonas palustris BisB18	0.920	0.156
Nostoc punctiforme PCC 73102	0.862	0.184
Gloeobacter violaceus PCC 7421	0.826	0.072
Anaeromyxobacter sp. Fw109-5	0.825	0.033
Roseiflexus sp. RS-1	0.779	0.159
Herpetosiphon aurantiacus ATCC 23779	0.749	0.071
Rhodopseudomonas palustris BisA53	0.739	0.151
Sinorhizobium meliloti 1021	0.729	0.092

Table S4. Species with sequenced genomes that were the closest match with annotated 454 reads from Rothamsted Park Grass soil. Percentage of annotated reads and standard deviation among the 13 pyrosequenced DNA pools that matched with each genome with a similarity at 96% or better and an E value of 10^{-5} or smaller).

Class	Species	Distribution (%)	SD (%)
Alphaproteobacteria	Bradyrhizobium japonicum USDA 110	0.086	0.041
Alphaproteobacteria	Bradyrhizobium sp. BTAi1	0.031	0.008
Acidobacteria	Solibacter usitatus Ellin6076	0.022	0.009
Alphaproteobacteria	Rhodopseudomonas palustris BisB18	0.019	0.006
Gammaproteobacteria	Pseudomonas fluorescens PfO-1	0.013	0.010
Alphaproteobacteria	Nitrobacter hamburgensis X14	0.013	0.003
Alphaproteobacteria	Rhodopseudomonas palustris BisA53	0.011	0.003
Betaproteobacteria	Burkholderia fungorum	0.009	0.008
Alphaproteobacteria	Rhodopseudomonas palustris BisB5	0.008	0.002
Gammaproteobacteria	Pseudomonas fluorescens SBW25	0.007	0.006
Alphaproteobacteria	Mesorhizobium loti MAFF303099	0.007	0.005
Bacteroidetes	Flavobacterium johnsonia johnsoniae UW101	0.007	0.018
Alphaproteobacteria	Nitrobacter winogradskyi Nb-255	0.006	0.002
Alphaproteobacteria	Nitrobacter sp. Nb-311A	0.005	0.001
Alphaproteobacteria	Rhodopseudomonas palustris HaA2	0.005	0.001
Gammaproteobacteria	Pseudomonas fluorescens Pf-5	0.005	0.005
Alphaproteobacteria	Rhizobium leguminosarum bv. viciae 3841	0.005	0.003
Actinobacteria (class)	Streptomyces avermitilis MA-4680	0.005	0.006
Acidobacteria	Acidobacteria bacterium Ellin345	0.004	0.003
Alphaproteobacteria	Rhodopseudomonas palustris CGA009	0.004	0.001
	Total 20 first detected species	**0.273**	**0.070**
	Total annotated sequences	**0.354**	**0.089**

Table S5. Newbler assembler results for the ten independent runs using randomly selected reads from the total of 13 pyrosequencing runs for DNA extracted from soil from the Park Grass experiment plot 3d at Rothamsted Research. Newbler parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed step : 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap identity : 90%; Alignment identity score: 2; Alignment difference score: -3.

Assembler runs:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
runMetrics										
totalNumberOfReads	1257242	2514489	3771729	5028954	6286181	7543423	8800651	10057895	11315094	12572342
totalNumberOfBases	487554794	974670636	1461941647	1949323461	2436671088	2.924E+09	3411376290	3898801814	4386462616	4874169257
Reads Status										
numAlignedReads	57556	185373	352696	554010	788063	1048430	1330404	1634721	1959374	2303182
numAlignedBases	18266373	59672888	114246594	180909939	259220082	346814643	442802556	547016473	658672541	777402244
inferredReadError	5.06%	5.35%	5.43%	5.42%	5.36%	5.31%	5.24%	5.17%	5.11%	5.05%
numberAssembled	31175	104814	200774	319201	459853	620159	798751	994756	1206256	1432866
numberPartial	26381	80500	151773	234651	328066	427926	531532	639829	752966	870034
numberSingleton	1163529	2236204	3260115	4244122	5192433	6112877	7011630	7884093	8738366	9572485
numberRepeat	21	103	268	408	763	1146	1548	1968	2487	3074
numberOutlier	30890	82472	143156	209757	278972	350046	420777	495652	568277	642021
numberTooShort	5246	10396	15643	20815	26094	31269	36413	41597	46742	51862
Total	1257242	2514489	3771729	5028954	6286181	7543423	8800651	10057895	11315094	12572342
largeContigMetrics										
numberOfContigs	3702	9138	17233	27641	40104	53850	68470	84705	101154	118800
numberOfBases	3757449	9786251	17197015	26537742	37785216	50417029	64287965	79618149	95770632	113089732
avgContigSize	1014	1070	997	960	942	936	938	939	946	951
N50ContigSize	1076	1094	998	964	951	949	954	956	965	971
largestContigSize	6361	15448	22645	21204	14380	15875	13912	15752	15426	15425
Q40PlusBases	3208918	8414019	14785259	22817145	32526441	43456627	55545708	68928952	83085358	98268346
Q39MinusBases	548531	1372232	2411756	3720597	5258775	6960402	8742257	10689197	12685274	14821386

Table S5. Newbler assembler results for the ten independent runs using randomly selected reads from the total of 13 pyrosequencing runs for DNA extracted from soil from the Park Grass experiment plot 3d at Rothamsted Research. Newbler parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed step : 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap identity : 90%; Alignment identity score: 2; Alignment difference score: -3.

Assembler runs:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
allContigMetrics										
numberOfContigs	7478	21418	41388	65598	94069	124842	157780	192206	228651	266600
numberOfBases	4738744	12993575	23467908	36508692	51965915	69162343	87893864	108053114	129548168	152337226
alignmentDepths										
1	254115	562821	1047133	1689799	2431319	3210275	4011784	4837438	5731987	6632746
2	1203570	3072382	5787534	9165657	13092802	17334645	21793710	26405044	31238812	36348350
3-4	2093933	4747497	8409968	13474365	19616431	26379378	33651362	41544235	49746108	58339992
5-6	1110246	2960407	4631690	6921263	9876485	13396813	17371792	21667238	26454007	31525166
7-8	332199	1556836	2402579	3184272	4289997	5675685	7371741	9243575	11315472	13600152
9-10	94034	788301	1541061	1954878	2376906	2973868	3721430	4618985	5621952	6710813
11-13	36453	424071	1292575	1901405	2252372	2648259	3078596	3676553	4324585	5059861
14-16	10742	116673	520348	1113087	1457370	1641430	1849416	2092694	2379250	2673058
17-19	5515	40973	174771	542959	932733	1128344	1237370	1340651	1458570	1609141
20-22	2440	15876	63147	213932	537004	812516	931154	964149	988136	1061923
23-25	1773	6433	30899	84046	248247	499440	695566	770823	781784	791998
26-28	1125	4689	16783	39095	108385	257508	463531	605929	674295	663610
29-31	893	3213	7901	21079	46572	120284	254402	431273	544012	592207
32-34	471	2465	4487	12294	27656	56103	127353	254512	396612	484982
35-38	291	1395	3732	9901	19892	41132	78045	164154	303232	460025
39-42	200	1135	2169	4596	11067	21353	36487	68246	143552	263660
43-46	50	1150	1570	3449	5774	13134	21329	36671	65517	123733
47-50	132	451	1102	1567	3299	7786	15037	21567	31831	58094
51-55	304	367	942	1408	2421	4825	11218	16476	25135	37157
56-60	90	323	517	1117	1235	3105	5834	11695	17448	24615
61-70	10	368	486	1510	2142	3085	4382	10797	20884	28372
71-80	6	383	282	722	1350	1875	2397	4266	8248	13665

Table S5. Newbler assembler results for the ten independent runs using randomly selected reads from the total of 13 pyrosequencing runs for DNA extracted from soil from the Park Grass experiment plot 3d at Rothamsted Research. Newbler parameters: Expected depth: 0 (i.e. undefined); Minimum read length: 20; Seed step : 12; Seed length : 16; Seed Count : 1; Minimum overlap length : 40; Minimum overlap identity : 90%; Alignment identity score: 2; Alignment difference score: -3.

Assembler runs:	1	2	3	4	5	6	7	8	9	10
91-100	0	102	191	245	406	670	1227	1460	1770	2513
101-140	0	15	942	580	508	1155	2609	4075	3786	5084
141-180	0	0	0	1068	699	313	291	892	1964	2207
181-240	0	0	0	41	676	814	406	463	500	921
241-300	0	0	0	0	12	500	981	606	469	400
301-400	0	0	4	0	0	3	33	652	1063	1178
401-500	0	0	20	0	0	7	1	1	15	84
501-600	8	0	5	0	1	5	0	0	12	4
601-700	0	0	1	0	0	1	0	0	1	2
701-850	0	0	0	0	0	3	0	0	5	10
851-1000	0	0	0	0	0	9	0	0	0	14
1001+	0	20	23	26	13	23	29	16	26	38

Read Status – status of the read in the assembly, which can be one of the following:

a. Assembled – the read is fully incorporated into the assembly

b. PartiallyAssembled – only part of the read was included in the assembly, the rest was deemed to have diverged sufficiently to not be included

c. Singleton – the read did not overlap with any other reads in the input

d. Repeat – the read was either:

i. Inferred to be repetitive early in the assembly process. A read can be inferred to be repetitive if >70% of the read's seeds hit to at least 70 other reads. Such reads are excluded from the assembly.

ii. Determined to partially overlap a contig. The portions of such reads that overlap unique contigs are still included in the assembly results.

e. Outlier – the read was identified by the GS De Novo Assembler as problematic, and was excluded from the final contigs (one explanation of these outliers are chimeric sequences, but sequences may be identified as outliers simply as an assembler artifact)

f. TooShort – the trimmed read was too short to be used in the computation (shorter than 50 bases and longer than the value of the minlen parameter, unless 454 Paired End Reads are included in the dataset, in which case, all reads at least "minlen" bases are used).

References

Agarwal N and Bishai WR. (2009). cAMP signaling in Mycobacterium tuberculosis. Indian J Exp Biol 47:393-400.

Akhter Y, Yellaboina S, Farhana A, Ranjan A, Ahmed N, Hasnain SE. (2008). Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis. Gene 407:148-58.

Allwood AC, Walter MR, Kamber BS, Marshall CP, Burch IW. (2006). Stromatolite reef from the Early Archaean era of Australia. Nature 441:714-718.

Davis KE, Sangwan P, Janssen PH. (2011). Acidobacteria, Rubrobacteridae and Chloroflexi are abundant among very slow-growing and mini-colony-forming soil bacteria. Environ Microbiol 13:798-805.

Delmont TO, Malandain C, Prestat E, Larose C, Monier J-, Simonet P, et al. (2011a). Metagenomic mining for microbiologists. ISME Journal. DOI: 10.1038/ismej.2011.61

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P et al. (2011b). Accessing the soil metagenome for studies of microbial diversity. Appl Environ Microbiol 77:1315-24.

Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. (2011c).Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods. 2011;86(3):397-400.

Demaneche S, Sanguin H, Poté J, Navarro E, Bernillon D, Mavingui P et al. (2008). Antibiotic-resistant soil bacteria in transgenic plant fields. Proc Natl Acad Sci USA 105: 3957-3962.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM et al. (2008). Functional metagenomic comparison profiling of nine biomes. Nature. 452:629-632.

Falkowski PG. (2001). Biogeochemical cycles. Encyclopedia Biodivers. 1:437-453.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF. (2008). Microbial community gene expression in ocean surface waters. Proc Natl Acad Sci USA 105:3805-10.

Ginolhac A, Jarrin C, Gillet B, Robe P, Pujic P, Tuphile K et al. (2004). Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. Appl. Environ. Microbiol 70:5522-5527.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. (1998). Molecular Biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5:245-249.

Huber JA, Pointing SB, Chan Y, Lacap DC, Lau MC, Jurgens JA, Farrell RL. (2009). Highly specialized microbial diversity in hyper-arid polar desert. Proc Natl Acad Sci USA 106:19964-9.

Kahvejian A, Quackenbush J, Thompson JF. (2008). What would you do if you could sequence everything? Nat Biotechnol 2610:1125-1133.

Knietch A, Waschkowitz T, Bowien S, Henne A, Daniel R. (2003). Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on Echerichia coli. J Microbiol Biotechnol 5: 46-56.

Larose C, Berger S, Ferrari C, Navarro E, Dommergue A, Schneider D, Vogel TM. (2010). Microbial sequences retrieved from environmental samples from seasonal Arctic snow and meltwater from Svalbard, Norway. Extremophiles 14:205-12.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-80.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M et al. (2008). The Metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes BMC Bioinformatics 19-9:386.

Morales SE, Holben WE. (2009). Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by *in silico* analysis. Appl Environ Microbiol 75:2677-83.

Nealson KH, Venter JC. (2007). Metagenomics and the global ocean survey: what's in it for us, and why should we care? ISME J 1:185-7.

Niu B, Fu L, Sun S, Li W. (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11:187.

Ortíz-Castro R, Contreras-Cornejo HA, Macías-Rodríguez L, López-Bucio J. (2009). The role of microbial signals in plant growth and development. Plant Signal Behav 4:701-12.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H-Y, Cohoon M et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Research, 33:5691-5702.

Parks DH, Beiko RG. (2010). Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26:715-21.

Rajendhran J, Gunasekaran P. (2008). Strategies for accessing soil metagenome for desired applications. Biotech Adv 26: 576-90.

Ranjard L, Richaume AS. (2001). Quantitative and qualitative microscale distribution of bacteria in soil. Res Microbiol 152: 707–716.

Ramirez KS, Lauber CL, Knight R, Bradford MA, Fierer N. (2010). Consistent effects of nitrogen fertilization on soil bacterial communities in contrasting systems. Ecology91:3463-70.

Roesch LL, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD et al. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1: 283-290.

Rousk J, Bååth E, Brookes PC, Lauber CL, Lozupone C, Caporaso JG, Knight R, Fierer N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. ISME J 4:1340-51.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59-65.

Schloss PD,Handelsman J. (2003). Biotechnological prospects from metagenomics. Curr Opin Biotechnol 14: 303-310.

Shendure J, Ji H. (2008). Next-generation DNA sequencing. Nat Biotech 26: 1135-1145.

Silvertown, J, Poulton P, JohnstonE, Edwards G, Heard M, Biss PM. (2006) The Park Grass Experiment 1856-2006: its contribution to ecology. (2006). Journal of Ecology 94: 801–814.

Tringe SG, Mering CV, Kobayashi A, Salamov AA, Chen K, Chang HW et al. (2005). Comparative Metagenomics of Microbial Communities. Science 308:554-557.

Van Elsas JD, Jansson JK, Trevors JT. (2006). Modern Soil Microbiology II, CRC press.

Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD et al. (2009). TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Microbiol 7:252.

Willner D, Thurber RV, Rohwer F. (2009). Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol 11:1752-66.

Chapter 3. Rothamsted Park Grass soil study of disturbed microbial communities: a richness vision

1. **A microbial communities' travel between soils**
2. **Stressing complex microbial communities for metagenomic discoveries: one designed evenness at the time**

In chapter 2, section 4, the natural microbial evenness of the Rothamsted Park Grass soil was defined using a relatively deep sequencing effort. However, this metagenomic approach was only able to access a part of the genetic richness of predominant microorganisms. The majority of microorganisms is lowly represented, and therefore, cannot be studied using classical metagenomic approaches.

Due to the complexity of soil microbial communities, it is easier to study their evenness (relative distribution of the diversity) than their richness (number of distinct taxa present). However, I tried to apply different approaches to uncover certain aspects of soil microbial richness.

In the section 1, microbial communities from two distinct soils were transferred into their native soil and in the other soil after soil sterilization. Two hypotheses can be proposed:

1/ the richness of the two communities is different (some taxa are present only in one soil). In this case, the two communities will probably be unable to react similarly when colonizing the same sterilized habitat.

2/ the richness of the two communities is (mainly) similar and only the evenness of microorganisms varies due to particular physico-chemical soil characteristics. In this case, the two communities could react in the same direction when colonizing the same sterile habitat. Then, the evenness of the two distinct communities might become identical.

Based on the obtained results, the two communities reacted equally when colonizing the first soil and equally when colonizing the second soil. Thus, soil characteristics appeared to be the predominant parameters in comparison to the structure of the two communities before the colonization. As a consequence, the large majority of taxa present in one of these two soils is probably also present in the second one. However, additional experiments have to be done to conclude about the similarity of the richness between soils from across the planet.

In the section 2, the natural evenness of the Rothamsted soil was modified using controlled environmental modifications in microcosms to access other diversities than those already sequenced (chapter 2, section 4) and to stimulate the reconstruction of genomes from a soil metagenome. The main objective of this approach was to reach the objective of the Terragenome consortium: to sequence and assemble a soil metagenome.

Since soil richness is high (see chapter 2, section 3 and 4, and chapter 3, section 1), we need new experimental designs to divide the entire soil metagenome into sub-groups that are easily reconstructed (presence of few predominant microorganisms) to better conquer this genetic gold mine.

The strategy proposed in this section succeeded in constructing distinct draft genomes from simplified soil metagenomes and mining their genomic structures.

Following the colonization effect of distinct communities in sterilized environments: a soil microbial richness study

Tom O. Delmont, Davide Francioli, Sophie Jacquesson, Sandra Laoudi, Pascal Simonet and Timothy M. Vogel.

Environmental Microbial Genomics Group, Laboratoire AMPERE, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France

Abstract: While soil is considered as the most biodiverse environment on Earth, the amplitude and role of it rare biosphere is largely unknown and debatable. Due to the unsuitability of 16S rRNA gene amplification and sequencing approaches to study soil richness, other approaches have to emerge. Here two sterilized soils (from England and Italy) were inoculated with a subsample of their initial microbial communities and those from the other soil to study their microbial community evolution. This original approach aims to compare driving factors (original community and soil physico-chemical characteristics) for microbial community definition. After two months of incubation and based on direct environmental DNA sequencing, the two communities possessed similar functional and taxonomical structures when inoculated in the same soil. Thus, similar lowly represented microorganisms from the two distinct communities emerge in each sterilized soil, emphasizing the importance of the original soil microbial community. Microbial communities from different locations could mainly have the same richness and a different evenness due to differences in environmental characteristics.

Key words: Soil, metagenomic, rare biosphere, richness, core.

Running title: A soil microbials richness study

Introduction:

Soil is a highly biodiverse environment that is now largely studied using metagenomic approaches [1]. However, due to the magnitude of its microbial populations (between 10^8 and 10^{11} cells per gram), the number of different species (from 10^4 to 10^7 as a function of the estimations, [2, 3]) and the difficulties to access a soil metagenome [4], its microbial communities represent a black box for ecologists.

In spite of evident limits, it is generally accepted that the evenness of microbial communities varies as a function of soil properties (e.g., pH parameter, [5, 6]). To study soil microbial evenness, 16S rRNA amplification and sequencing techniques were largely used (e.g., [7-9]). Results generally emphasize the prevalence of Alphaproteobacteria, Actinobacteria and Acidobacteria followed by Beta, Gamma and the delta-epsilon subdivision of Proteobacteria. In addition, the majority of species appeared to be lowly represented [10], what could emphasize important amplitude of the rare biosphere.

Unfortunately, this strategy is probably inadequate to study soil richness and it related rare biosphere due to both DNA extraction and PCR steps (in particular primer designing and DNA amplification) that limit the detection of species [11] and sequencing errors generated that overestimate soil biodiversity [12]. Thus the fact that almost all phylogenetic probes (>99.7%) from a microarray were lighted when fractionating a single soil metagenome prior amplifying 16S rRNA genes [4] is not sufficient to conclude about the ubiquity of bacterial and archaeal species in this environment. Obviously, it is actually difficult to study soil richness which can potentially be the same in all natural soils and other approaches have to be developed for more suitable studies.

Here we studied the colonization effect of two distinct soil communities on their soil of origin or not as an alternative strategy to appreciate microbial core between these soils. Soils were sampled from England grassland and Italian forest. Communities were inoculated both separately and mixed in the two soils subsequently to their sterilization (six experiments). In our hypothesis and if the two communities possess the same richness, they should developed and became similar in term of evenness after colonizing the same soil. In the other hand, if several species are present only in one community, their evenness should not be identical at the end of the experiment.

Interestingly, the two inoculated communities evolved considerably during the soil colonization and reacted equally in each sterilized soil. In addition, the two soils appear to possess relatively similar communities before their sterilization and colonization. Thus three distinct groups were present and named “Controls” (soils not sterilized), “England” (the England soil sterilized and colonized) and “Italy” (the Italian soil sterilized and colonized). Based on these observations, soils characteristics influence evenness (the environment selects) and after two months the two communities were stable, similar in the same colonized soil and different between soils.

Material and methods:

Soil management:

The England soil was collected from the untreated control plot (3d) of Park Grass Experiment, Rothamsted Research, Hertfordshire, UK (Silvertown et al., 2006) in July 2010. The Italian soil was collected from Vallombrosa forest, defined as a Cambric Umbrisol.
Soil samples from the top 21 centimeters were collected (Delmont et al., 2011) by sterile manual corers (10 cm diameter) and were placed in sterile plastic bags, sealed and placed on ice 24 hours till and processing. After sieving soils, a part of these soils was then sterilized using an autoclave (120°C during 60 minutes) in microcosms of 30g during six days (two sterilizations per day each two days).

Then nine experiments were performed in triplicats:

"Controls": C1/ England soil (30g); C2/ Italian soil (30g); C3/ A mix of England and Italian soils (15g each).

"England" sterilized soil: E1/ 2g of England soil into 30g of England sterilized soil; E2/ 2g of Italian soil into 30g of England sterilized soil; E3/ 2g of both England and Italian soils into 30g of England sterilized soil.

"Italian" sterilized soil: I1/ 2g of England soil into 30g of Italian sterilized soil; I2/ 2g of Italian soil into 30g of Italian sterilized soil**;** I3/ 2 g of both England and Italian soils into 30g of Italian sterilized soil.

For all microcosms, 3ml of sterilized water was added in spray.

DNA extraction and quantification:

Environmental DNA was extracted from 0.5g of soil and using the MP BIO 1O1 fast prep (Biomedical, Eschwege, Germany) (Griffiths et al., 2000). Samples were purified using GFX columns (final volume of 40 microliters) and the DNA was finally quantified using the Qubit technology. No DNA was quantified on the control sterilized soils during the experimentation (six months) (detection limit of 200 pg of double strand DNA per well).

Ribosomal intergenic spacer analysis (RISA).:

The intergenic spacer (IGS) region between the small (16S) and the large (23S) subunits of ribosomal sequences were amplified by PCR using primers 5_-TGCGGCTGGATCCCCTC CTT-3_ (forward) and 5_-CCGGGTTTCCCCATTCGG-3_ (reverse) (29). For the PCR mix, 2 _l of DNA (10 _M) was mixed with 1.25 _l of reverse and forward primers (10 _M) and 20.5 _l of distilled water (DH2O). PCR cycles consisted of 95°C for 10 min and then 30 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 1 min, followed by 72°C for 15 min, with a Biometra thermocycler. One microliter of the PCR mix was then loaded into an Agilent DNA 7500 Lab

on a Chip, and electropherograms were analyzed and data were normalized by using an Agilent 2100 Bioanalyzer.

Pyrosequencing runs:

A minimum of 10 µg of DNA were used for each Roche/454 pyrosequencing run on a 454 pyrosequencer (GS FLX Titanium Series Reagents ; Roche 454; Shirley, NY, USA). Processing of samples (prior to sequencing) did not involve prior amplification step. An egal quantity of each triplicat was used for each tag. The sequence data are publically available (http://www.genomenviron.org/Projects/METASOIL.html).

Data analyses:

Artificial duplicates were deleted using cd-hit-454 with default parameters [13]. A total of 1.548.028 distinct reads (average length of 425.7 pb (±4.0%)) were generated and metagenomes of 172 003 sequences (±22.2%) represent each condition. Sequences were then annotated on the MG RAST online software [14]. Reads were distributed into different metabolic subsystems. Similarity search between pyrosequencing reads and the SEED database [15] have been processed with a maximum e-value of 10^{-5}. All compared distributions were normalized as a function of the number of annotated sequences for each metagenome. Data corresponding to both functional and taxonomical distributions were then statistically analyzed within the STAMP software [16]. When comparing two groups, Fisher's exact tests were performed and annotated functions and taxa with p-values < 0.05 were considered to be significantly different between the different experiments. When comparing three groups, the ANOVA test integrating Bonferroni correction was applied to the subsystems. Again, the p-value threshold was 0.05.

Results:

Microbial communities versus soil characteristics:

In this experiment, nine distinct conditions were applied (see material and method section) to test the impact of microbial communities and soil characteristics on the colonization of sterilized soil samples.

After two months of inoculation and based on the quantity of DNA extracted, communities were relatively stable in term of population density (figure 1). The quantity of DNA extracted was about 0.84µg/g (±0.4) after two weeks and 10.11µg/g (±2.6) after two months, so representing an augmentation of more than 12 times in six weeks. In addition and based on RISA profiles, replicates were reproducible after 8 weeks (Figure S1). Thus metagenomes corresponding to the nine different experiments (C1, C2, C3, E1, E2, E3, I1, I2, I3) were generated after two months of incubation to study the colonization effect of microbial communities into sterilized soils.

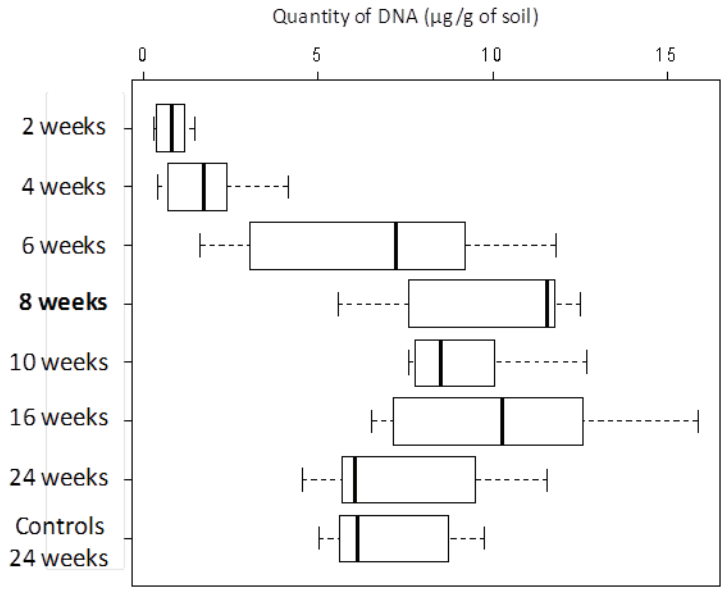


Figure 1: Quantity of DNA extracted in the 6 conditions (E1, E2, E3, I1, I2 and I3 in replicates) of inoculated soils after 2, 4, 6, 8, 10, 16 and 24 weeks. The "Controls" box plot corresponds to the soils not sterilized (C1, C2 and C3) after 24 weeks in microcosms. Box plots represent the variation of DNA extracted among the different microcosms.

Using the MG-RAST annotation platform [14], a total of 809 functional subsystems and 998 sequenced species (SEED) were detected. Their distributions were then confronted among the nine datasets by performing principal component analyses (PCA) (figure2). Interestingly, metagenomes are positioned in three distinct groups based on both functional and taxonomical distributions (panels A and B): C1, C2 and C3 (Controls cluster), E1, E2 and E3 (England cluster), and finally I1, I2 and I3 (Italian cluster). Thus the natural microbial communities from the England grassland and Italian forest soils appear to possess a relatively similar functional and taxonomical distribution when comparing them to the same communities colonizing the two sterilized soils. In addition, communities corresponding to the Italian group evolved more than the "England" group based on the functional and taxonomical relative distributions. Due to the important similarity shown among each of these three clusters, metagenomes were grouped in three (Controls, England and Italy) for the next analyses.

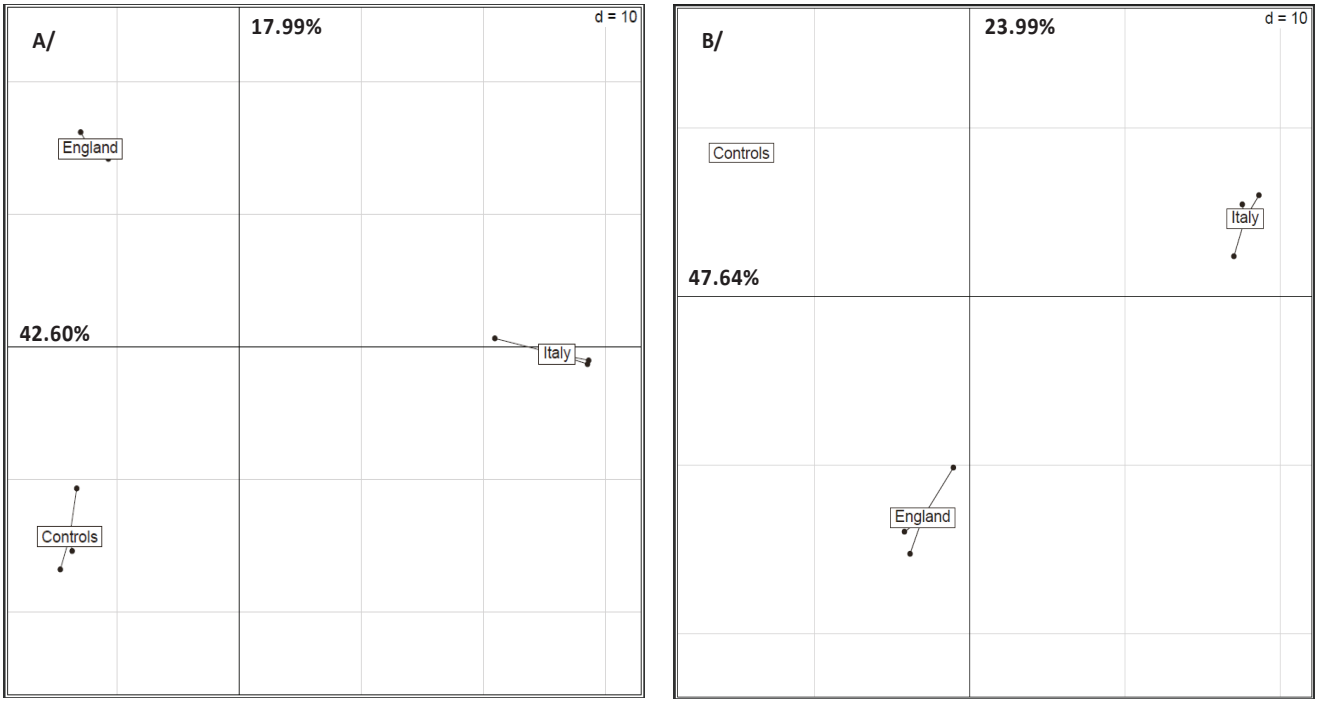


Figure 2: Principal component analysis based on the relative distribution of 809 functional subsystems (panel A/) and 998 species (SEED annotation, panel B/) among metagenomes corresponding to C1, C2, C3 ("Controls"), E1, E2, E3 ("England") and I1, I2 I3 ("Italy") experiments after two months of incubation.

Functional and taxonomical groups' specificities:

At the domain level, only the proportion of Archaea varied significantly between groups (p-value= 2.68e-6). In fact, it distribution appears to be relatively more represented in the controls (1.00%, ±0.03) than in the Italian (0.42%, ±0.02) and England (0.44%, ±0.02) groups.

Among the principal phyla and classes, important distribution differences were observed between the three groups (figure 3, panel A). In particular, sequences related to Alpha (p-value= 1.37e-3), Beta (p-value= 3.59e-5), Gamma (p-value= 0.011), and the delta/epsilon division (p-value= 0.011) of Proteobacteria are in lower relative proportion in communities inoculated in the Italian soil. In contrast, sequences related to the Actinobacteria class (p-value= 1.85e-5) appear to increase considerably during it colonization (66.89%, ±2.72) whatever the inoculated community in comparison to the controls (14.74%, ±1.36) and the England sterilized soil inoculated (14.66%, ±1.65). In addition, sequences related to Bacteroides are more distributed in the England inoculated soil (p-value= 0.037). Finally, The Acidobacteria phyla appears to be less represented in the two colonized soils in comparison to the controls (p-value= 0.05).

When comparing the general functional subsystems in the different groups (figure 3, panel B), differences are less marked. However, some subsystems vary significantly between metagenomes. Especially, sequences related to Carbohydrates (p-value= 6.89e-4) are more represented in communities inoculated in the Italian soil (18.56%, ±0.43) than in the other experiments (13.61 (±0.12) and 14.41 (±0.22) respectively in the controls and England

groups). In contrast, sequences related to virulence (p-value=7.59e-3), regulation and cell signaling (p-value=1.74e-4), motility and chemotaxis (p-value=9.34e-3), and cell wall and capsule (p-value=0.011) are less represented in this condition in comparison to the two other.

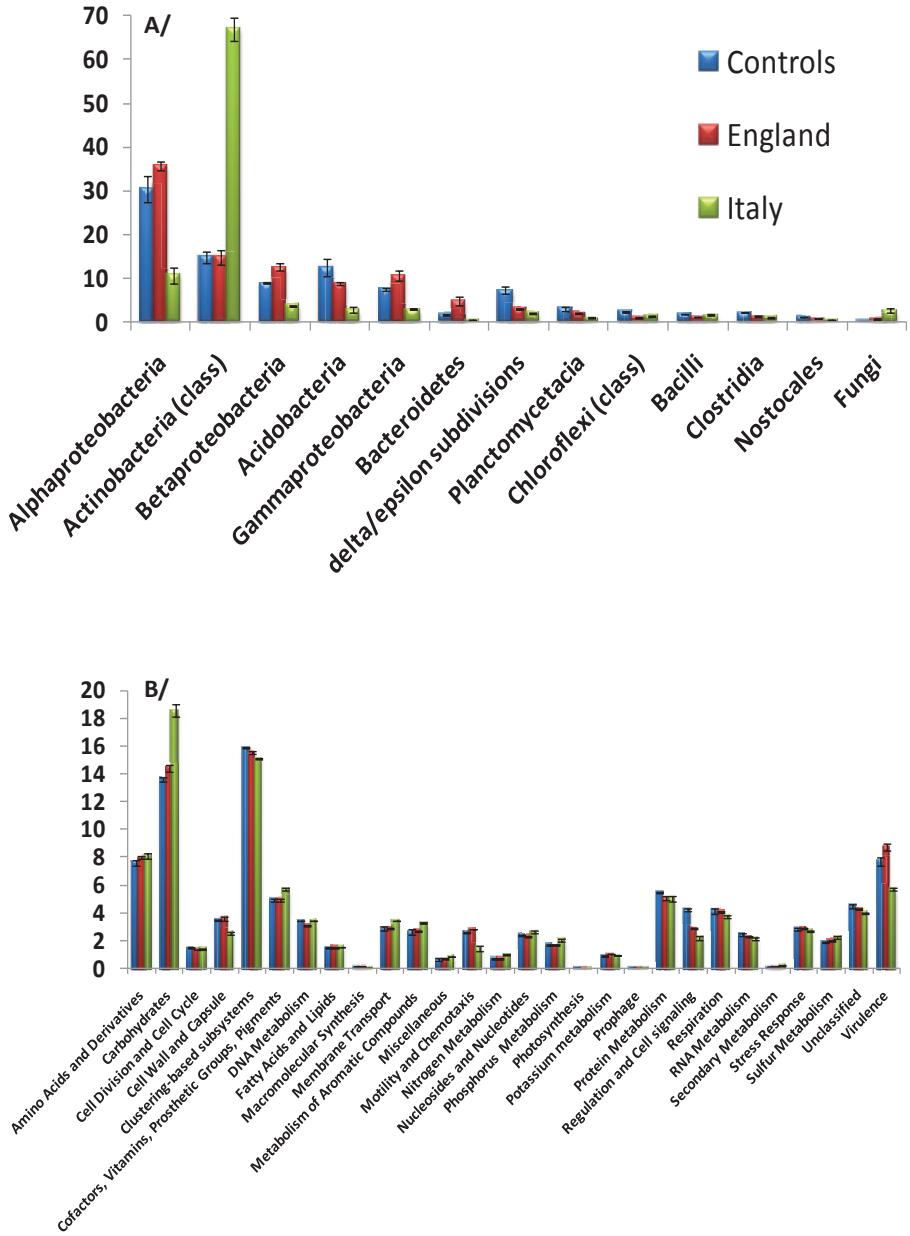


Figure 3: Relative distribution of principal phyla (panel A/) and functions (panel B/) among metagenomes corresponding to C1, C2, C3 ("Controls"), E1, E2, E3 ("England") and I1, I2 I3 ("Italy") experiments after two months of incubation.

Among functions that vary significantly between the different groups, sequences related to Ton and Tol transport system, Heme, hemin uptake and utilization systems in Gram Negatives and Hemin transport system appear to be more represented in the England group (figure S2). Sequences related to cAMP signaling in bacteria and Iojap (a family of small

proteins) are more distributed in the control group (figure S2). In contrast, sequences related to Alkanesulfonates Utilization are less detected.

On the other hand, sequences related to Chitin and N-acetylglucosamine utilization, Ectoine biosynthesis and regulation, Spore pigment biosynthetic cluster in Actinomycetes and L-rhamnose utilization for example are more represented in the Italian group (figure S3). Finally, sequences related to Lipopolysaccharide assembly and Multidrug Resistance Efflux Pumps are lowly detected in this group.

Discussion:

The concept of soil rare biosphere sharpens the passion of several microbial ecologists and it notion is still largely debatable in scientific articles and international conferences. The principal reason of this disagreement is the presence of insoluble limits and biases on techniques generally used to access environmental microbial diversity. In fact, DNA extraction approaches, probes design, DNA amplification steps and sequencing errors perturb our vision of soil evenness and richness when studying 16S rRNA gene amplicons (e.g., [4, 11, 12, 17]). In addition, it substance is highly conceptual (no defined frontiers between normal and rare biosphere) and the part of microorganisms placed in one or the other collection is selected arbitrarily. Thus to define the importance and role of the soil rare biosphere is a headache for scientists and a considerable effort has to be done to stimulate it study and to create a concrete definition.

Due to the importance of soil biodiversity and difficulties to study it related microorganisms, microbial ecologists, geneticists and bioinformaticians created an international consortium (Terragenome) to focus on a unique soil metagenome (Park Grass, Rothamsted, UK). But obviously, to sequence again and again environmental DNA corresponding to natural soil microbial communities using the actual sequencing technologies appeared to be an unsuitable approach to access lowly represented microorganisms. In fact, even predominant genomes cannot be assembled after the sequencing of tens of millions of pyrosequences (Delmont et al., in press). Thus other methods have to be developed to study soil rare biospheres without systematic metagenome sequencing or 16S rRNA amplification steps.

Here and to by-pass some of the involved biases in classical methods to study soil richness, an original approach is proposed to test soil rare biosphere amplitude. Using two distinct communities and sterilized soils, the impact of both community structures and soil characteristics on the colonization of no bacteria's land soil environments was studied using both fingerprint and metagenomic approaches. At the end of the experiment and based on both functional and taxonomical distributions, the nine generated metagenomes were placed in three distinct groups: the controls (not sterilized soils, C1, C2, C3) and communities that colonized the England (E1, E2, E3) and Italian (I1, I2; I3) sterilized soils (see figure 2).

These groups were also defined using the fingerprint approach and true biological replicates (figure S1).

Interestingly, the two sterilized soil characteristics appear to impact more the final community structure than the inoculated community (figure 4). The fact that two distinct communities (from England grassland and an Italian forest soil) possess identical potentials to evolve similarly when colonizing the same no life's land environment emphasizes an important microbial core between the two soils. Without confirming it, these results support the idea of a common richness between soils.

While only few differences were detected between the two natural communities in spite of distinct localizations, considerable structural and functional differences were present between the three groups after two months of experiments (figure 3 and S2, S3, S4) so emphasizing both soil differences and communities' evolution capacities when colonizing a new habitat.

Unexpectedly, the distribution of the Actinobacteria phylum increased considerably during the colonization of the Italian sterilized soil (see figure 3, panel A). This structural modification impacted the functional distribution of the related metagenomes (I1, I2 and I3) and thus these metagenomes are a good model to link structure and function in soil microorganisms (e.g., negative correlation with the subsystem corresponding to Multidrug Resistance Efflux Pumps). Globally, communities that possess more of microorganisms related to Actinobacteria possess also more of genes related to Carbohydrate functions (figure 3 and S3). In contrast, sequences related to virulence for example are more present in communities for which the Proteobacteria phylum is more represented. However, functions vary less than community structures in these nine metagenomes due to an important functional redondancy in soil microorganisms. Thus other methods to study not only functional potentials but also microbial activities have to be done (e.g., metatranscriptomic and metaproteomic approaches) for a deeper analysis of these communities' transfer experiments.

The fact that two distinct communities evolve considerably and similarly when colonizing a sterilized soil add information on the importance of their richness. Thus this new approach appears to be suitable to study soil richness and to access normally undetected microorganisms. In fact, the England soil microbial community (reference of the Terragenome consortium) has to be modified to stimulate the study of the entire metagenome [4] and to transfer it communities into other soils appears to modify considerably it structure (figure 4). A more global experiment based on several natural and polluted soils and even other environments as matrices for the Rothamsted soil community has now to be performed to break the soil black box and access it rare biosphere.

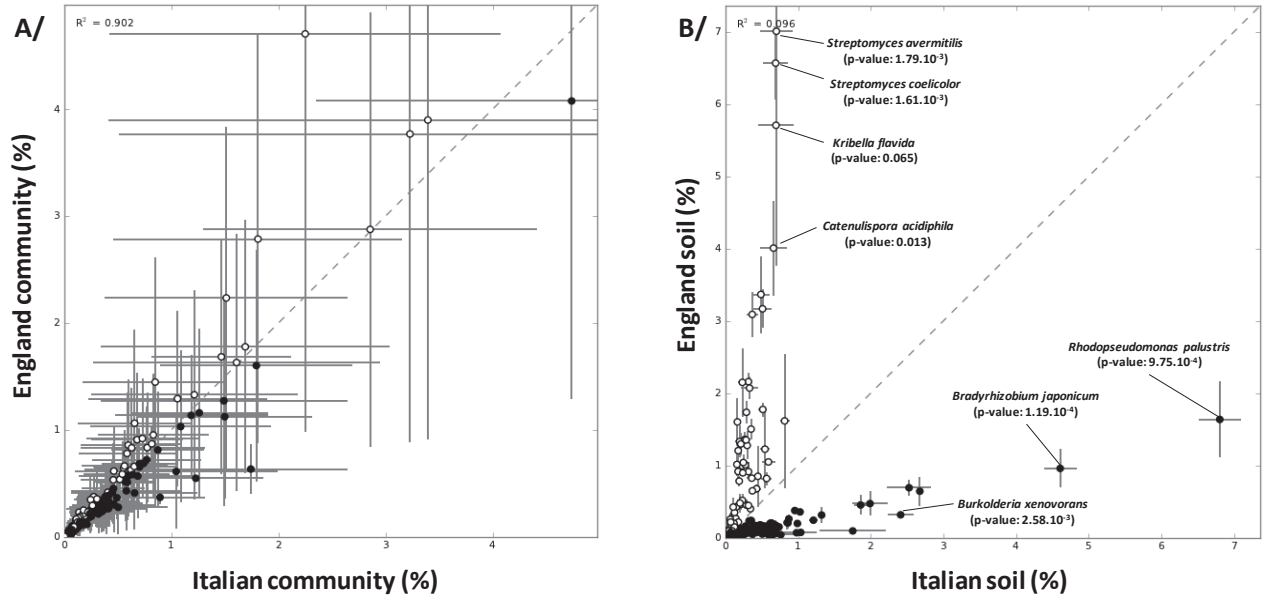


Figure 4: Relative distribution in percentage of species (SEED annotation) between two groups. Panel A: groups correspond to the Italian (E2 and I2) and England inoculated community (E1 and I1). Panel B: groups correspond to the Italian (I1, I2, I3) and England (E1, E2, E3) soil sterilized.

Conclusion:

Studying soil microorganisms is a challenge for microbial ecologists and to access it rare biosphere is probably a considerable defy of science. To illustrate this challenge, the DNA contained in 100 grams of soil possessing a population of 10^{10} cells per gram is equivalent to twice the distance between earth and moon, when all DNA is put end to end. In contrast, pyrosequencing Titanium and HiSeq runs correspond respectively to 0.13 and 150 meters of DNA approximately. Obviously and in spite of considerable advances in the field, classical approaches revealed considerable limits to study soil richness and other approaches have to emerge. Here the effort of understanding the impact of communities' transfers into two distinct sterilized soils in term of both structure and function provides new insights about soil richness. While these experiments are insufficient to conclude about a hypothetical common richness among soils, results emphasize a considerable microbial core between the two tested soils. By focusing on one soil and making efforts on modifying the natural structure of it microbial communities, the black box would finally be cracked and to transfer communities into other matrices is one strategy among others to reach this objective.

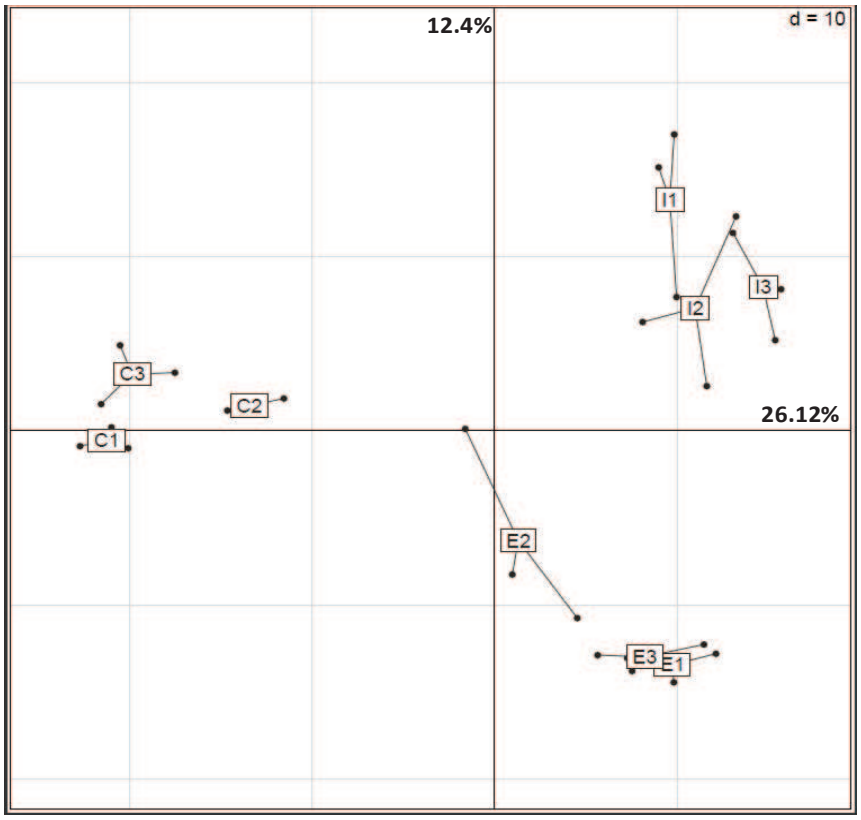


Figure S1: PCA based on RISA profiles corresponding to the nine experiments (see material and method section) in triplicats and after 8 weeks in microcosms. Percentages correspond to the variance explained in each axe.

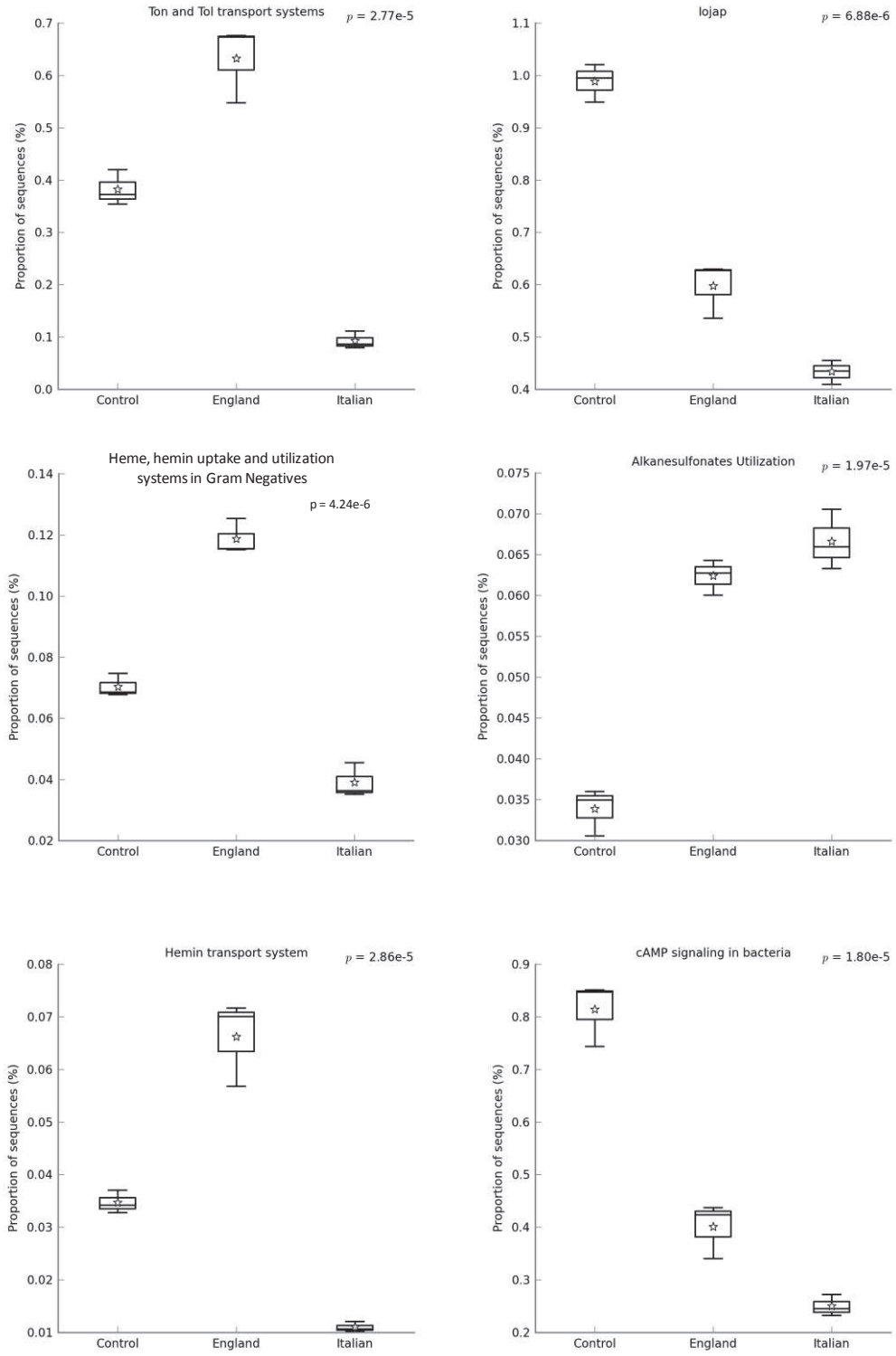


Figure S2: Relative distribution of functional subsystems in metagenomes corresponding the control, England and Italian groups, using MG-RAST and a E-value cut-off of 10^{-5}.

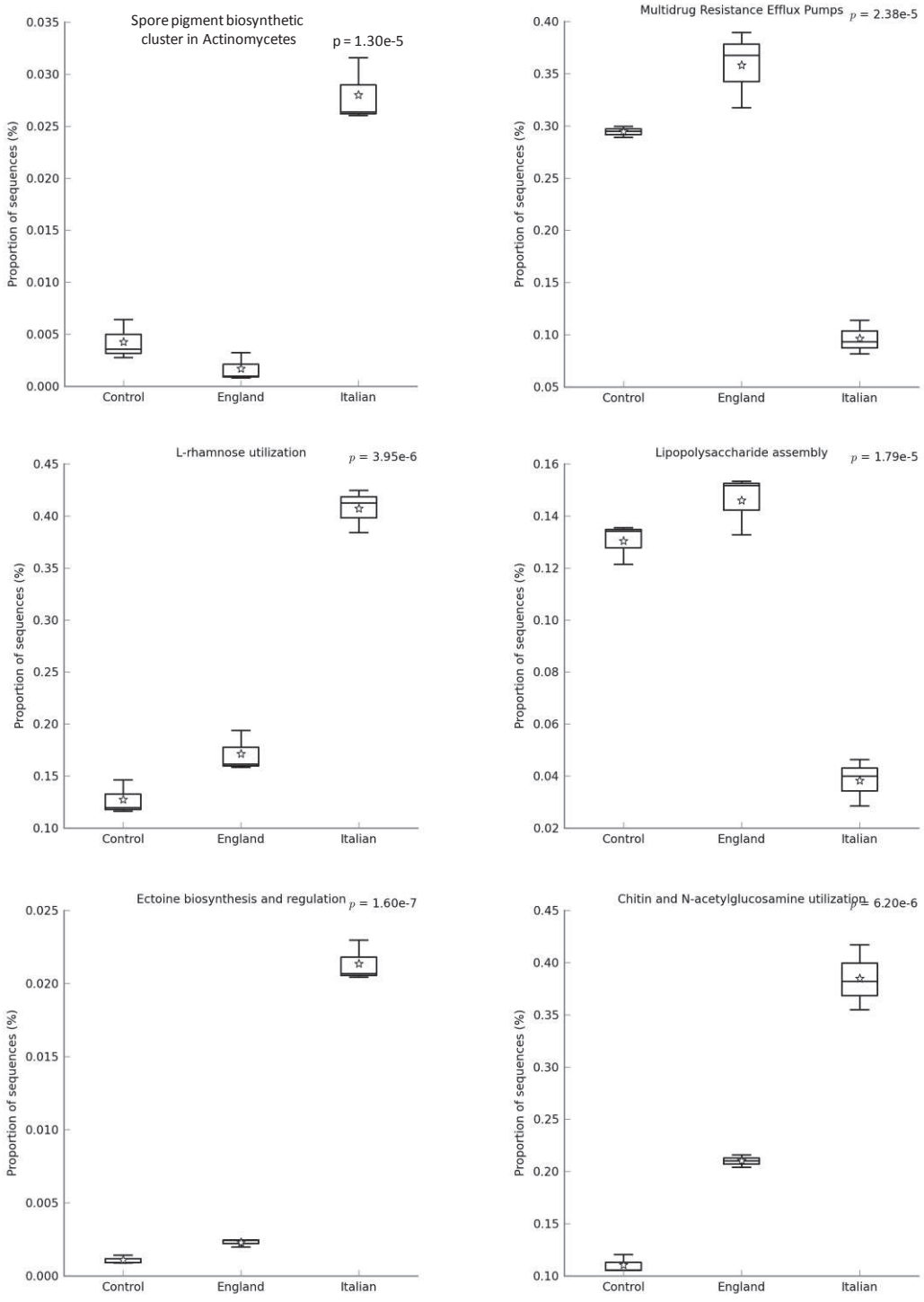


Figure S3: Relative distribution of functional subsystems in metagenomes corresponding the control, England and Italian groups, using MG-RAST and a E-value cut-off of 10^{-5}.

References:

1. Vogel, T.M., et al., *TerraGenome: a consortium for the sequencing of a soil metagenome*, in *Nat Rev Microbiol*2010. p. 252.
2. Torsvik, V., L. Ovreas, and T.F. Thingstad, *Prokaryotic diversity--magnitude, dynamics, and controlling factors.* Science, 2002. **296**(5570): p. 1064-6.
3. Gans, J., M. Wolinsky, and J. Dunbar, *Computational improvements reveal great bacterial diversity and high metal toxicity in soil.* Science, 2005. **309**(5739): p. 1387-90.
4. Delmont, T.O., et al., *Accessing the soil metagenome for studies of microbial diversity.* Appl Environ Microbiol, 2011. **77**(4): p. 1315-24.
5. Lauber, C.L., et al., *Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale.* Appl Environ Microbiol, 2009. **75**(15): p. 5111-20.
6. Rousk, J., et al., *Soil bacterial and fungal communities across a pH gradient in an arable soil.* ISME J, 2010. **4**(10): p. 1340-51.
7. Morales, S.E., et al., *Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters.* Appl Environ Microbiol, 2009. **75**(3): p. 668-75.
8. Roesch, L.F., et al., *Pyrosequencing enumerates and contrasts soil microbial diversity.* ISME J, 2007. **1**(4): p. 283-90.
9. Elshahed, M.S., et al., *Novelty and uniqueness patterns of rare members of the soil biosphere.* Appl Environ Microbiol, 2008. **74**(17): p. 5422-8.
10. Ashby, M.N., et al., *Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities.* Appl Environ Microbiol, 2007. **73**(14): p. 4532-42.
11. Hong, S., et al., *Polymerase chain reaction primers miss half of rRNA microbial diversity.* ISME J, 2009. **3**(12): p. 1365-73.
12. Kunin, V., et al., *Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates.* Environ Microbiol, 2010. **12**(1): p. 118-23.
13. Niu, B., et al., *Artificial and natural duplicates in pyrosequencing reads of metagenomic data.* BMC Bioinformatics, 2010. **11**: p. 187.
14. Meyer, F., et al., *The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes.* BMC Bioinformatics, 2008. **9**: p. 386.
15. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Res, 2005. **33**(17): p. 5691-702.
16. Parks, D.H. and R.G. Beiko, *Identifying biologically relevant differences between metagenomic communities.* Bioinformatics, 2010. **26**(6): p. 715-21.
17. Hamp, T.J., W.J. Jones, and A.A. Fodor, *Effects of experimental choices and analysis noise on surveys of the "rare biosphere".* Appl Environ Microbiol, 2009. **75**(10): p. 3263-70.

Stressing complex microbial communities for metagenomic discoveries: one designed evenness at the time

Tom O Delmont, Emmanuel Prestat, Eric Pelletier, Denis LePaslier, Pascal Simonet and Timothy M. Vogel

Abstract: One of the biggest challenges in metagenomics is the reconstruction of genomes directly sequenced from complex environments. The evenness of microbial communities from an undisturbed grassland soil was highly modified in microcosm conditions prior deep sequencing efforts to access other genetic diversities and stimulate assembly efficiency. Soil communities were incubated at 37°C, under anaerobic condition or under aerobic conditions with specific enrichments (different concentrations of ethanol, salt, mercury, heavy metals or diesel) during four months and then sequenced. Generated datasets represent distinct parts of the entire metagenome. Results provided valuable information about how complex communities respond during controlled environmental modifications (*e.g.*, unusual distribution of key functions that respond directly to the induced stress). Interestingly, assembly efficiency was improved under several conditions (up to 80% of reads were assembled) and led to several draft genomes. Additional sequencing efforts were done to complete genomes. Constructed genomes were then used to help the annotation of the undisturbed soil metagenome and tracked in datasets from several environments to study their distribution across the planet. Paradoxically, we are currently able to reconstruct lowly represented genomes selected by relatively harsh or specific conditions but not those from the small amount of predominant native soil microorganisms. However, when considering microbial biodiversity estimations and the number of conditions that can be applied, soil metagenomic discoveries can potentially be equivalent to what can be found in other environments.

Key words: metagenomic, microcosm, evenness, assembly

Introduction:

From deep oceans to hyper-arid deserts (Grzymski, Murray et al. 2008; Pointing, Chan et al. 2009), microorganisms colonized almost all conceivable ecosystems after their emergence on our planet more than three billion years ago (Allwood, Walter et al. 2006). With an estimated 10^{30} microorganisms on Earth (Whitman, Coleman et al. 1998), they impact all aspects of life, from health and agriculture to environmental depollution and climate stability (*e.g.*, O_2/CO_2 ratio in the atmosphere). While microorganisms have been studied for centuries using cultural approaches, they are now often examined by extracting and sequencing their nucleic diversity directly from the environment (*e.g.*, (Tyson, Chapman et al. 2004; Venter, Remington et al. 2004). By comparing generated sequences to already known genes (reference databases), this strategy can describe complex microbial communities without any cultural step. In addition, the relative distribution of taxa and functions can be compared among and between environments to help understand the life style variation of microorganisms across the planet (Delmont, Malandain et al.; Tringe, von Mering et al. 2005; Dinsdale, Edwards et al. 2008).

However, due to the high diversity of metagenomes extracted from the majority of microbial habitats (*e.g.* oceans and soils), only a small part of the environmental genetic diversity is currently sequenced. When considering the extraction and sequencing of DNA, generated datasets are mainly influenced by the evenness of its represented communities and correspond mainly to predominant microorganisms. Thus, when using metagenomic experimental designs to extract and sequence DNA directly from the environment, the best strategy to access the diversity of the 10^{30} microorganisms is probably to track original evenness from across the planet than to focus on one complex metagenome. As an example, the Earth microbiome project (EMP) aims to sequence a large number of DNA samples extracted from various environments (Gilbert, Meyer et al. 2010) and will access distinct evenness, so stimulating considerably the number of taxa and functions detected.

In addition, genomes can potentially be reconstructed from metagenomes that represent extreme environments or specific habitats. The interest of the approach was already demonstrated with datasets from an acid mine drain (Tyson, Chapman et al. 2004; Bertin, Heinrich-Salmeron et al. 2011), a hypersaline lake (Narasingarao, Podell et al. 2011), cow rumen (Hess, Sczyrba et al. 2011) and human feces (Qin, Li et al. 2010). Thus, while the majority of metagenomes will not be efficiently assembled, several genomes could be reconstructed from relatively lowly complex datasets generated during the EMP, so improving our understanding about genomic diversity.

In contrast, the Terragenome initiative (Vogel, Simonet et al. 2010) that aims to sequence and assemble an entire soil metagenome appears to be difficult due to the extensive evenness. As an example, the already generated datasets from the Rothamsted Park Grass soil were sufficient to characterize globally predominant microbial communities and perform inter-environmental metagenomic comparisons but were largely unable to reconstruct

genomes (Delmont et al., in press). This metagenomic assembly efficiency limit can easily be explained by the absence of highly predominant microorganisms and an important similarity among phylogenetic groups of genomes. Thus, generated datasets are too complex and the large majority of reads are unassembled.

In spite of considerable difficulties accessing the entire soil metagenome (Delmont, Robe et al. 2011), the estimated soil microbial biodiversity is extensive (Torsvik, Ovreas et al. 2002; Gans, Wolinsky et al. 2005; Roesch, Fulthorpe et al. 2007). Therefore, when considering the biological potential of soil and other diverse microbial communities, defining alternative strategies to access its metagenomic diversities is of high interest for both the scientific community and a wide range of industries. Here, we propose to modify the evenness of the microbial community under controlled environmental conditions prior deep sequencing efforts to access distinct parts of its metagenome and stimulate assembly efficiency.

The relative distribution of species can be modified without destroying all microorganisms, so emphasizing a considerable flexibility of the taxonomical structure of microbial communities and an opportunity for microbiologists. A typical example is the Mexico gulf oil spill which took place in 2010, and where deep-ocean and beach sand microbial communities were rapidly modified with the emergence of predominant oil-degrading microorganisms (Hazen, Dubinsky et al. 2010; Kostka, Prakash et al. 2011). These species would be difficult to access without this particular environmental perturbation. Of course, it is not necessary to apply an oil spill to the environment to access other diversities. It could be possible to enrich a few liters of ocean water with oil prior to sequencing extracted DNA to reconstruct these particular genomes.

Thus, this flexibility can be used to design the evenness of a community but suitable experiments have to be performed. Communities have to incubate under relatively extreme microcosm conditions to generate unusual datasets and success on metagenome assembly. Each condition can potentially increase the relative distribution of specific microorganisms due to the presence of key functions in their genomes, so opening new possibilities to access the genomic structure of normally lowly represented taxa and to stimulate metagenomic discoveries. In addition, this approach can potentially be apply to culture and single cell techniques to reconstruct other genomes and libraries constructions to study the role of new functions. The overall strategy is to sequence and mine complex microbial communities at one designed evenness at the time. A major advantage of this approach is the possibility to create evenness that does not exist.

Thus, the number of lowly represented species accessible using metagenomic approaches should be positively correlated to the stress of the applied incubation conditions (*e.g.*, sugar, toxic compounds or antibiotics enrichment, temperature, pressure, modified atmosphere). We hypothesized that applying a high number of extreme incubation conditions to a complex environment followed by a deep sequencing effort is a suitable approach to improving genomic recovery.

To demonstrate the interest of this stratagem, ten distinct conditions and a control (no modified parameters) were applied to the Rothamsted Park Grass soil in microcosms. This soil has been studied for 150 years (Crawley, Johnston et al. 2005) and is known to possess a considerable biodiversity (Delmont, Robe et al. 2011). In addition, its metagenome was relatively unsuccessfully assembled using direct sequencing efforts (no contigs longer than 15 kbp; Delmont, Prestat et al., in press). As a consequence, this soil was a suitable environment to test the proposed strategy. After 4 months of incubation, metagenomes were generated, annotated and studied using both specific and global comparisons. Assembly efforts were finally performed for each condition.

Material and methods:

Soil management:

Soil was collected from the untreated control plot (3d) of Park Grass Experiment, Rothamsted Research, Hertfordshire, UK (Silvertown et al., 2006) in July 2010. Soil samples from the top 21 centimeters were collected (Delmont et al., 2011) by sterile manual corers (10 cm diameter) and were placed in sterile plastic bags, sealed and placed at room temperature 24 hours during it transport to Lyon (France). A total of 50 kilograms of soil were then sieved (2 mm) and one fraction was directly used for the microcosms experiment.

Microcosm conditions:

Microcosms were done in triplicates (50g of soil in each microcosm), stored at room temperature (excepted for the high temperature condition) without light, and were closed during all the experiment. In parallel to a control condition, ten distinct conditions were applied by modifying the environment of microorganisms in each microcosm:

Control condition: 5 ml of purified water was sprayed.
Nitrogen condition: 5 ml of purified water was sprayed. Then atmosphere was replaced by Nitrogen gas inside microcosms.
High temperature condition: 5 ml of purified water was sprayed. Then microcosms were incubated at 37°C.
Salt enrichment 1: 5 ml of purified water enriched on NaCl (30g/L) was sprayed.
Salt enrichment 2: 5 ml of purified water enriched on NaCl (300g/L, salt saturation) was sprayed.
Diesel enrichment: 5 ml of purified water enriched on diesel (for a final concentration of 50g/kg of soil) was sprayed.
Ethanol enrichment: 5 ml of purified water enriched on ethanol (20% of the volume) was sprayed.
Metals enrichment 1: 5 ml of purified water enriched on heavy metals (Zinc, Cadmium, Nickel and Cobalt, for a final concentration of 0.2g/kg of soil for each metal) was sprayed.
Metals enrichment 2: 5 ml of purified water enriched on heavy metals (Zinc, Cadmium, Nickel and Cobalt, for a final concentration of 2g/kg of soil for each metal) was sprayed.

Mercury enrichment 1: 5 ml of purified water enriched on inorganic mercury salts (for a final concentration of 0.02g/kg of soil) was sprayed.
Mercury enrichment 2: 5 ml of purified water enriched on inorganic mercury salts (for a final concentration of 0.2g/kg of soil for each) was sprayed.

DNA extraction and quantification:

After four months of incubation, soil samples were extracted from 0.5g of soil using the MP BIO 1O1 fast prep (Biomedical, Eschwege, Germany) (Griffiths, Whiteley et al. 2000). Samples were purified using GFX columns (GE Healthcare) (final volume of 40 microliters) and the DNA was finally quantified using the Qubit® (1.0) Fluorometer. A minimum of six DNA extractions were used to generate a standard deviation of extraction yield for each condition.

Ribosomal intergenic spacer analysis (RISA):

The intergenic spacer (IGS) region between the small (16S) and the large (23S) subunits of ribosomal sequences were amplified by PCR using primers 5'-TGCGGCTGGATCCCCTC CTT-3' (forward) and 5'-CCGGGTTTCCCCATTCGG-3' (reverse) (Ranjard, Brothier et al. 2000). For the PCR mix, 2 µl of DNA (10 µM) was mixed with 1.25 µl of reverse and forward primers (10 µM) and 20.5 µl of distilled water (DH2O). PCR cycles consisted of 95°C for 10 min and then 30 cycles of 95°C for 30 s, 55°C for 30 s, and 72°C for 1 min, followed by 72°C for 15 min, with a Biometra thermocycler. One microliter of the PCR mix was then loaded into an Agilent DNA 7500 Lab on a Chip, and electropherograms were analyzed and data were normalized by using an Agilent 2100 Bioanalyzer.

Pyrosequencing runs:

A minimum of 10 µg of DNA were used for each Roche/454 pyrosequencing run on a 454 pyrosequencer (GS FLX Titanium Series Reagents; Roche 454; Shirley, NY, USA). In some cases, tens of extractions were necessarily to cumulate the sufficient quantity of DNA. Thus, processing of samples (prior to sequencing) did not involve prior amplification step. The sequencing effort was done for biological duplicates of all conditions after four months of incubation. As an exception, the three replicates of the Mercury enrichment 2 condition were sequenced. The sequence data are publically available. (http://www.genomenviron.org/Projects/METASOIL.html).

HiSeq paired ends runs:

For the ethanol enrichment, metals enrichment 2 and mercury enrichment 1, an additional sequencing effort was done with the same DNA samples and using the HiSeq paired ends technology. Tags were done for each duplicate, and approximately 15 million reads were generated for each of the 6 tags. The sequence data are aloso publically available on the same website (http://www.genomenviron.org/Projects/METASOIL.html).

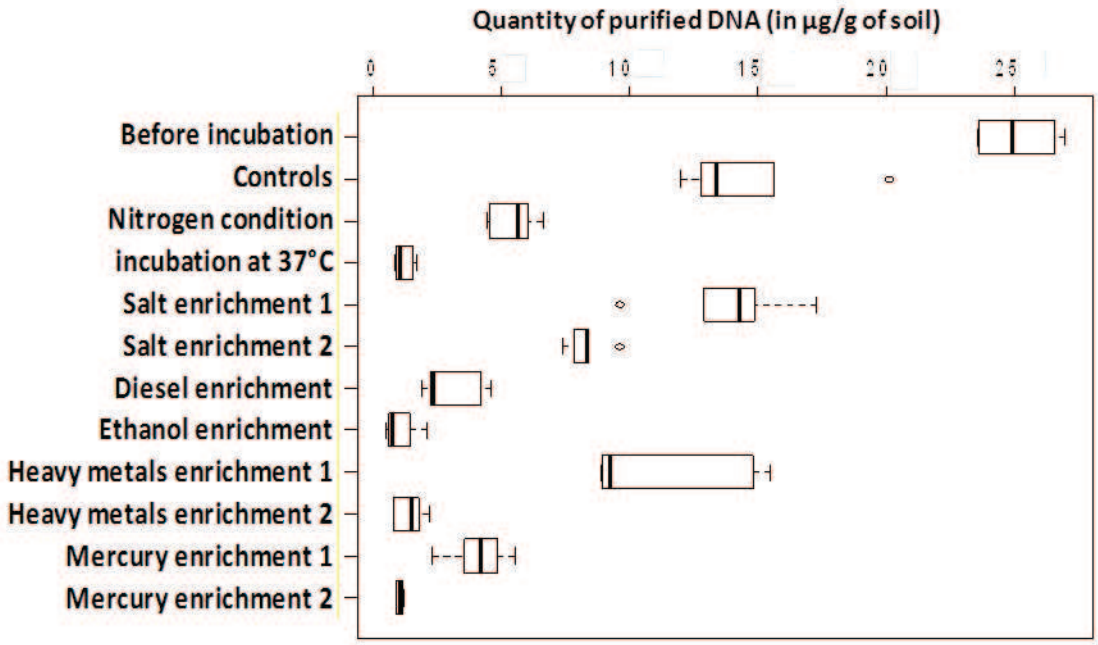
Data analyses:

Artificial duplicates were deleted using cd-hit-454 with default parameters (Niu, Fu et al. 2010). Sequences were then directly annotated on the MG RAST online software (Meyer, Paarmann et al. 2008) or assembled (Newbler software v.05) and annotated in contigs using IMG/M (Markowitz, Ivanova et al. 2008). Similarity search between pyrosequencing reads and the M5NR database (Overbeek, Begley et al. 2005) have been processed with a maximum e-value of 10^{-5}. All compared distributions were normalized as a function of the number of annotated sequences for each metagenome. Data corresponding to both functional and taxonomical distributions were then statistically analyzed within the STAMP software (Parks and Beiko 2010). When comparing all groups, the ANOVA test integrating Bonferroni correction was applied to the subsystems. Annotated functions and taxa with p-values < 0.05 were considered to be significantly different between the different experiments.

Results:

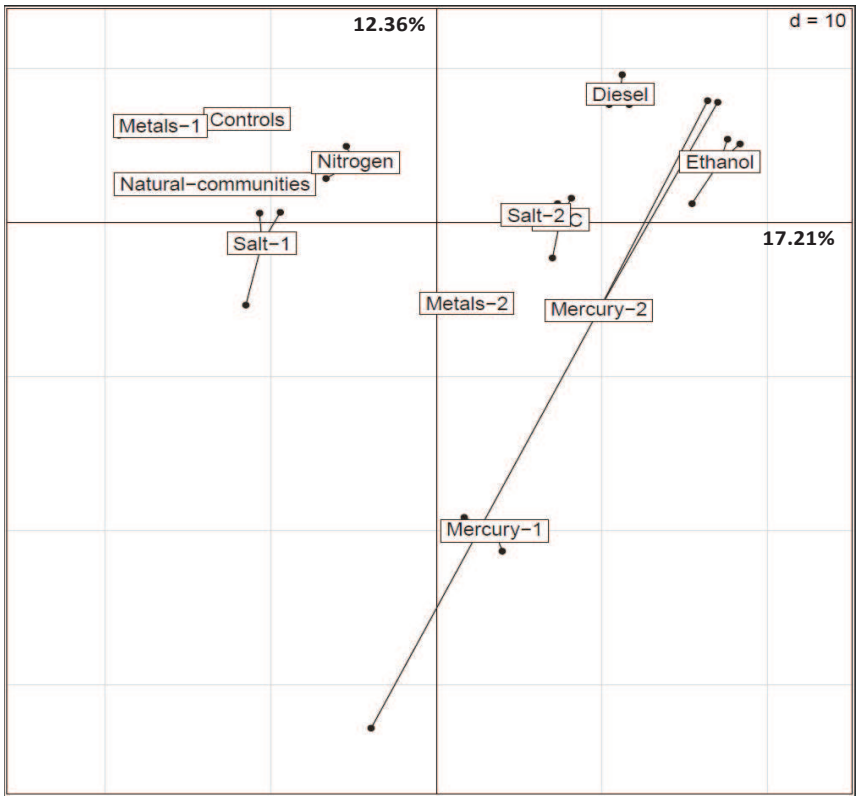
Quantity of extractable DNA among conditions and replicates reproducibility:

After 4 months of incubation DNA was extracted, purified and quantified for all microcosms. The yield of purified DNA per condition is presented in the supplement figure 1. 14.55 (±2.73) micrograms of DNA per gram of soil (µg/g) were extracted from the controls. This yield corresponds to a diminution of 42.03% when comparing to the quantity of DNA extracted from fresh soil before the microcosm incubation. In comparison to the controls, some conditions didn't impact considerably the yield of extracted DNA. It is in particular the case of the salt enrichment 1 (13.89 µg/g (±2.32)) and Metals enrichment 1 (13.83 µg/g (±2.43)). In contrast, the quantity of extracted DNA was highly impacted by specific conditions, often corresponding to extreme conditions (*e.g.*, Mercury enrichment 2 (0.99 µg/g (±0.15), Ethanol enrichment (0.72 µg/g (±0.38), Metals enrichment 2 (1.43 µg/g (±0.54), high temperature condition (1.19 µg/g (±0.34)).



Supplement figure 1: Quantity of DNA extracted and purified (in microgram of DNA per gram of soil) as a function of the applied condition. A minimum of four extractions were done for each condition. Box plots are based on the yield variation between these extractions.

A RISA fingerprint was then applied to all microcosms, and signal intensities were compared in a principal component analysis (supplement figure 2). Excepted for one condition (the Mercury enrichment 2), biological triplicates were highly reproducible in term of RISA profiles generated. Based on these results, communities from Nitrogen condition, salt enrichment 1 and metals enrichment 1 appear to be more similar to communities from fresh soil (T0) and the control condition than all the other conditions.



Supplement figure 2: Principal component analysis based on RISA profiles from the different Rothamsted soil microcosm conditions. DNA was extracted after four month of incubations. Each group represents triplicate variations.

Then, DNA was cumulated from each condition, and 23 pyrosequencing runs were performed. These runs correspond to biological duplicates from all conditions and to triplicates for the mercury enrichment 2. In addition, 13 runs that aim to represent a global picture of the Rothamsted Park Grass soil natural communities (Delmont et al., in press) were used as a control for better comparisons. Datasets were analyzed using MG-RAST-CLOUD and STAMP**.**

Microbial evenness among microcosm conditions:

Taxonomical structure of datasets was analyzed using M5NR, a non-redundant protein database (see http://blog.metagenomics.anl.gov/howto/m5nr-%E2%80%94-the-m5-non-redundant-protein-database/ for more details). Between 925 (with the third replicate of mercury enrichment 2) and 1532 (with the second replicate of salt enrichment 2) genera were detected in each dataset. By compiling the 36 datasets, a total of 4064 distinct genera were detected in the Park Grass Rothamsted soil. The distribution of the ten first genera for each condition was compiled in the table 1. In comparison to natural communities, *Mycobacterium* and *Brucella* genera appear to be stimulated in the controls, *Anaeromyxobacter* with the nitrogen condition, *Burkholderia* in the mercury enrichment 1 or still *Streptomyces* in a majority of conditions. The relative distribution of fungi increased in some conditions (*e.g.*, *Neosartorya* and *Aspergillus* with the ethanol and mercury enrichment 2).

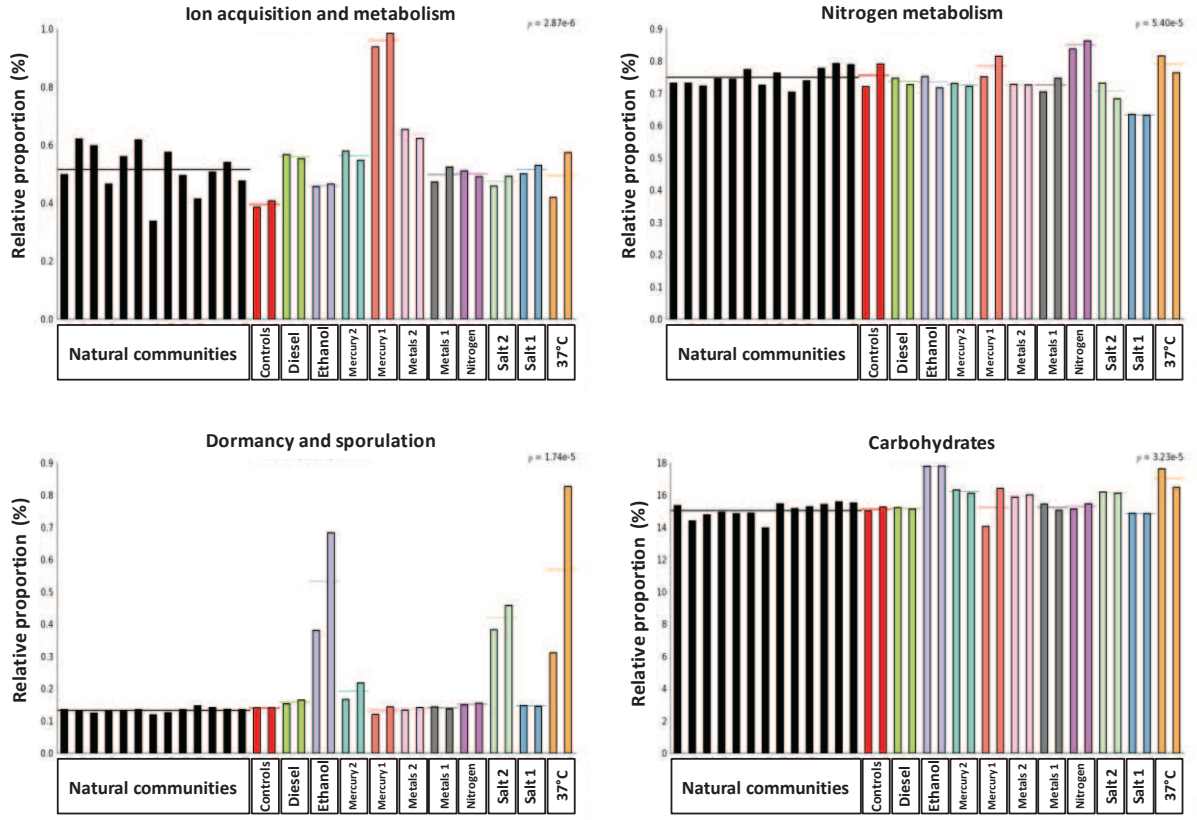
In addition to predominant genera, the distribution of several taxa varies considerably between conditions (a total of 366 genera with a p-value < 0.05, supplement data). Sequences related to *Actinomyces*, *Corynebacterium*, *Sanguibacter* and *Xhylanimonas* are more detected in communities sequenced after the heavy metals enrichment 2, Candidatus *Azobacteroides*, *Methanosarcina* and *Prevotella* with the Nitrogen condition. The ethanol enrichment provided a distinct evenness, with *Dehalococcoides*, *Dehalogenimonas*, *Lactobacillus*, *Lactococcus* and *Oenococcus* for example more represented in comparison to the other conditions. As a last example, sequences related to *Halobacillus* are more detected with the salt enrichment 2 (supplement data).

Natural communities (Delmont et al., 2011)			**Control**			**Nitrogen condition (no oxygen)**		
	Genera (%)	S D (%)		Genera (%)	S D (%)		Genera (%)	S D (%)
Bradyrhizobium	4.10	0.58	*Mycobacterium*	8.19	0.21	*Bradyrhizobium*	3.22	0.66
Streptomyces	3.17	1.26	*Brucella*	6.25	0.59	*Mycobacterium*	3.02	1.81
Burkholderia	3.06	0.32	*Streptomyces*	4.73	0.09	*Brucella*	2.91	2.79
Mycobacterium	2.44	0.86	*Burkholderia*	3.44	0.01	*Candidatus Solibacter*	2.76	0.64
Rhodopseudomonas	2.42	0.35	*Bradyrhizobium*	3.09	0.02	*Streptomyces*	2.74	0.15
Candidatus Solibacter	2.12	0.37	*Frankia*	2.26	0.04	*Burkholderia*	2.41	0.51
Methylobacterium	1.87	0.33	*Rhodopseudomonas*	1.94	0.02	*Candidatus Koribacter*	2.18	0.61
Frankia	1.61	0.57	*Bacillus*	1.74	0.09	*Bacteroides*	1.87	0.96
Pseudomonas	1.40	0.22	*Candidatus Solibacter*	1.67	0.03	*Anaeromyxobacter*	1.74	0.35
Gemmata	1.28	1.02	*Methylobacterium*	1.54	0.03	*Frankia*	1.66	0.26
Temperature (37°C)			**Salt enrichment (30g/L)**			**Salt enrichment (300g/L)**		
	Genera (%)	S D (%)		Genera (%)	S D (%)		Genera (%)	S D (%)
Streptomyces	11.56	0.18	*Mycobacterium*	10.35	0.73	*Bacillus*	6.24	0.76
Clostridium	3.98	2.17	*Streptomyces*	5.44	0.46	*Streptomyces*	5.54	0.25
Bacillus	3.96	1.76	*Xanthomonas*	4.82	0.63	*Mycobacterium*	3.22	0.10
Frankia	3.16	1.04	*Brucella*	4.18	0.19	*Frankia*	2.81	0.13
Ktedonobacter	2.60	1.62	*Burkholderia*	3.72	0.08	*Bradyrhizobium*	2.25	0.07
Mycobacterium	2.50	0.72	*Frankia*	2.68	0.24	*Burkholderia*	2.14	0.04
Micromonospora	2.44	0.47	*Shigella*	2.06	0.34	*Candidatus Solibacter*	1.86	0.06
Salinispora	2.01	0.62	*Escherichia*	1.82	0.39	*Geobacillus*	1.80	0.26
Paenibacillus	1.99	1.18	*Pseudomonas*	1.73	0.17	*Candidatus Koribacter*	1.50	0.05
Burkholderia	1.37	0.16	*Bradyrhizobium*	1.49	0.05	*Neosartorya*	1.29	0.29
Diesel enrichment (50g/kg)			**Ethanol enrichment (20%)**			**Metals enrichment (4 X 0.2g/kg)**		
	Genera (%)	S D (%)		Genera (%)	S D (%)		Genera (%)	S D (%)
Streptomyces	6.12	0.16	*Bacillus*	7.75	2.33	*Streptomyces*	5.39	0.66
Burkholderia	4.06	0.37	*Streptomyces*	6.93	2.22	*Bradyrhizobium*	4.25	0.63
Mycobacterium	3.87	0.06	*Ktedonobacter*	6.19	0.23	*Frankia*	3.03	0.01
Frankia	3.03	0.10	*Frankia*	2.87	0.84	*Mycobacterium*	2.89	0.51
Bradyrhizobium	2.59	0.10	*Mycobacterium*	2.31	0.58	*Burkholderia*	2.60	0.22
Candidatus Solibacter	1.62	0.06	*Geobacillus*	2.19	0.65	*Acidobacterium*	2.12	0.72
Methylobacterium	1.61	0.08	*Clostridium*	2.02	0.49	*Rhodopseudomonas*	1.90	0.13
Rhodopseudomonas	1.42	0.04	*Neosartorya*	1.74	0.73	*Chthoniobacter*	1.85	0.96
Bacillus	1.22	0.07	*Aspergillus*	1.41	0.59	*Candidatus Solibacter*	1.83	0.04
Nocardioides	1.22	0.01	*Paenibacillus*	1.28	0.16	*Methylobacterium*	1.83	0.03
Metals enrichment (4 X 2g/kg)			**Mercury enrichment (0.02g/kg)**			**Mercury enrichment (0.2g/kg) (rep 1 and 2)**		
	Genera (%)	S D (%)		Genera (%)	S D (%)		Genera (%)	S D (%)
Streptomyces	9.38	0.02	*Burkholderia*	17.40	1.95	*Streptomyces*	9.63	1.91
Mycobacterium	5.57	0.11	*Streptomyces*	16.40	6.28	*Neosartorya*	6.06	1.13
Frankia	4.30	0.03	*Xanthomonas*	4.06	1.24	*Aspergillus*	4.94	0.97
Xanthomonas	3.90	0.19	*Mycobacterium*	1.93	0.56	*Penicillium*	3.42	0.67
Burkholderia	2.67	0.11	*Pseudomonas*	1.87	0.34	*Burkholderia*	3.18	0.65
Arthrobacter	2.36	0.11	*Brucella*	1.80	1.56	*Bradyrhizobium*	1.85	0.05
Rhodococcus	1.80	0.01	*Acidobacterium*	1.50	0.33	*Mycobacterium*	1.67	0.18
Pseudomonas	1.57	0.09	*Bradyrhizobium*	1.36	0.05	*Candidatus Solibacter*	1.60	0.06
Bradyrhizobium	1.39	0.01	*Frankia*	1.19	0.36	*Frankia*	1.51	0.21
Nocardioides	1.32	0.02	*Shigella*	1.12	1.04	*Candidatus Koribacter*	1.42	0.04

Table 1: Relative distribution of the ten most detected genera for each condition when using M5NR database and an E-value cut-off of 10^{-5}. The standard deviation represents the variation between the two sequenced duplicates.

Functional distribution among microcosm conditions:

From general to highly specific subsystems, several functions vary significantly in distribution between conditions. As an example, the general functional subsystem related to iron acquisition and metabolism is distinctly more distributed in communities corresponding to the mercury enrichment 1 (supplement figure 3). Nitrogen metabolism appears to be more represented in nitrogen condition, and carbohydrates in the ethanol enrichment condition. Some general functions are more represented in different conditions that have apparently no relationships. In particular, the dormancy and sporulation subsystem is relatively stable in all conditions excepted in ethanol enrichment, salt enrichment 2 and high temperature condition were it distribution increases significantly. In contrast, some general functional subsystems are stable in all conditions. It is in particular the case of RNA metabolism, protein metabolism and amino acids and derivatives (supplement data).



Supplement figure 3: Relative distribution (in percentage of annotated reads) of general functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for datasets generated from the Rothamsted soil. Horizontal lines are the mathematical averages for the metagenomic datasets from each condition. The p-values are the likelihoods that the distribution differences observed between environments are random.

At the more specific functional level (collection of protein families and domains named Pfams, (Bateman, Coin et al. 2004)), a total of 8543 distinct functions were detected. The number of functions detected by dataset varied between 3853 (with the replicate 2 of salt enrichment 1) and 6214 (with the replicate 2 of high temperature condition). A total of 1373 of them vary significantly in distribution between conditions. For example, the Pfam related to Cobalt-zinc-cadmium resistance protein CzcD is around 0.02% in all conditions except in the heavy metals enrichment 2 where it distribution is about 0.065% (figure 2). Iron-containing alcohol dehydrogenase Pfam is more detected when enriching soil with ethanol. Sequences related to the anaerobic cytochrome c552 were more detected in the two metagenomes from the nitrogen condition. In addition, sequences related to mercuric resistance operon regulatory protein and organomercurial lyase (supplement data) were more detected in the mercury enrichment 1 in comparison to all the other conditions.

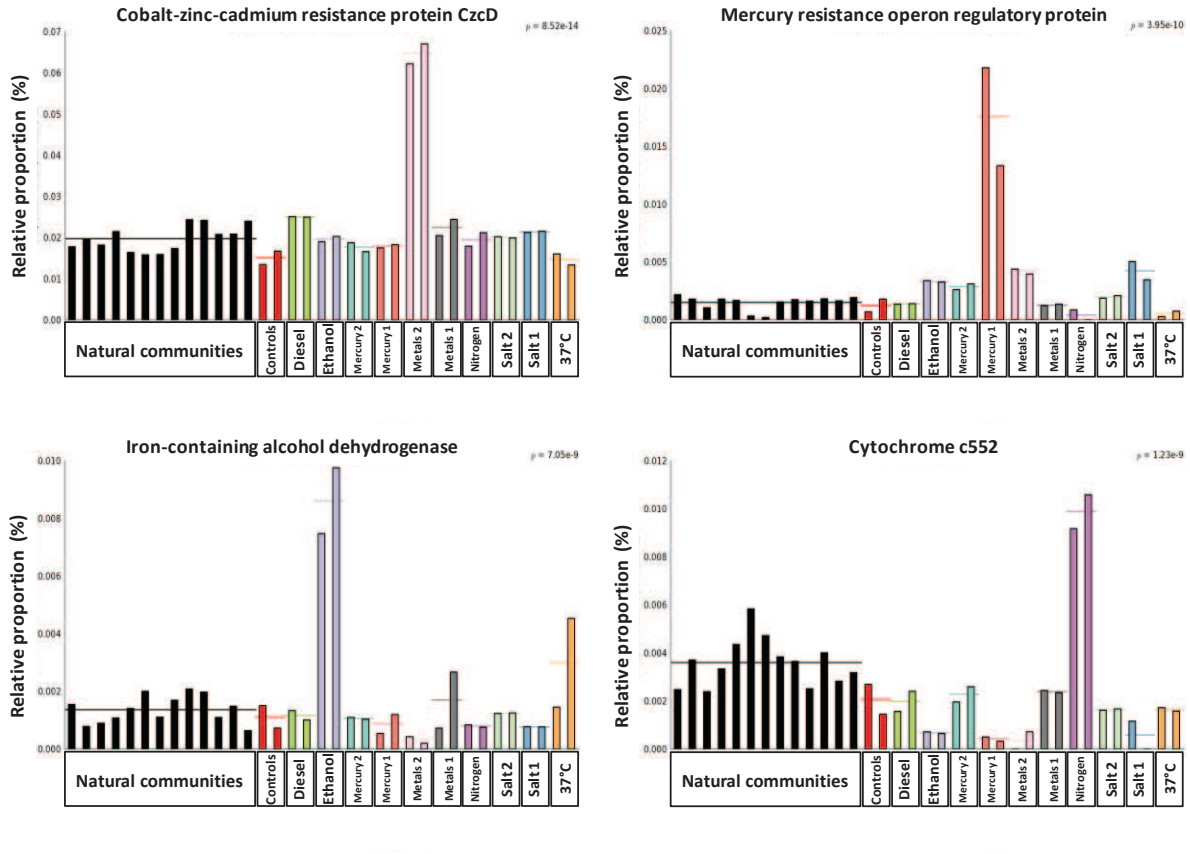
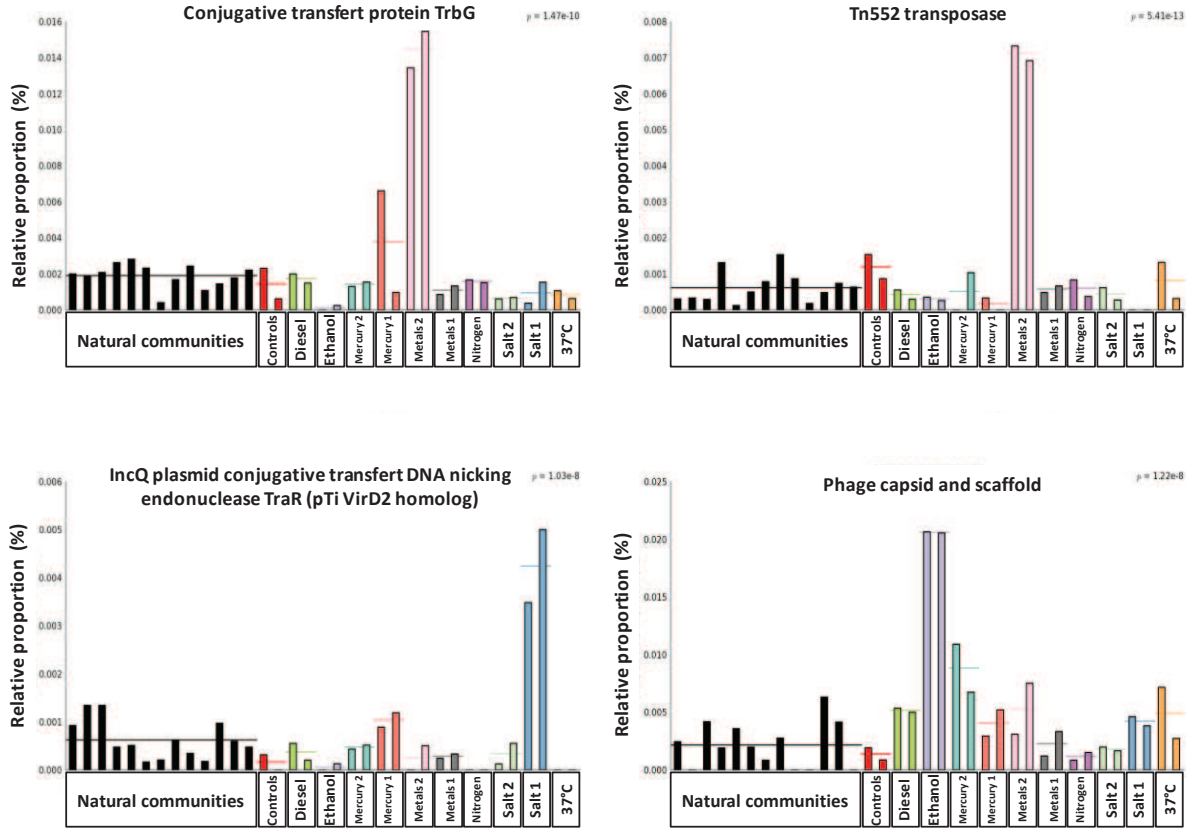


Figure 2: Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for datasets generated from the Rothamsted soil. Horizontal lines are the mathematical averages for the metagenomic datasets from each condition. The p-values are the likelihoods that the distribution differences observed between environments are random.

Moreover, some genetic structures known to be involved in microbial adaptation were more detected in specific conditions. In particular, sequences related to the conjugative transfert protein TrbG and Tn552 transposase appear to be more represented in the heavy metal

enrichment 2 (supplement figure 4). On the other hand, IncQ plasmid conjugative transfer DNA nicking endonuclease TraR (pTi VirD2 homolog) was more detected in the salt enrichment 1, phage capsid and scaffold with the ethanol enrichment.

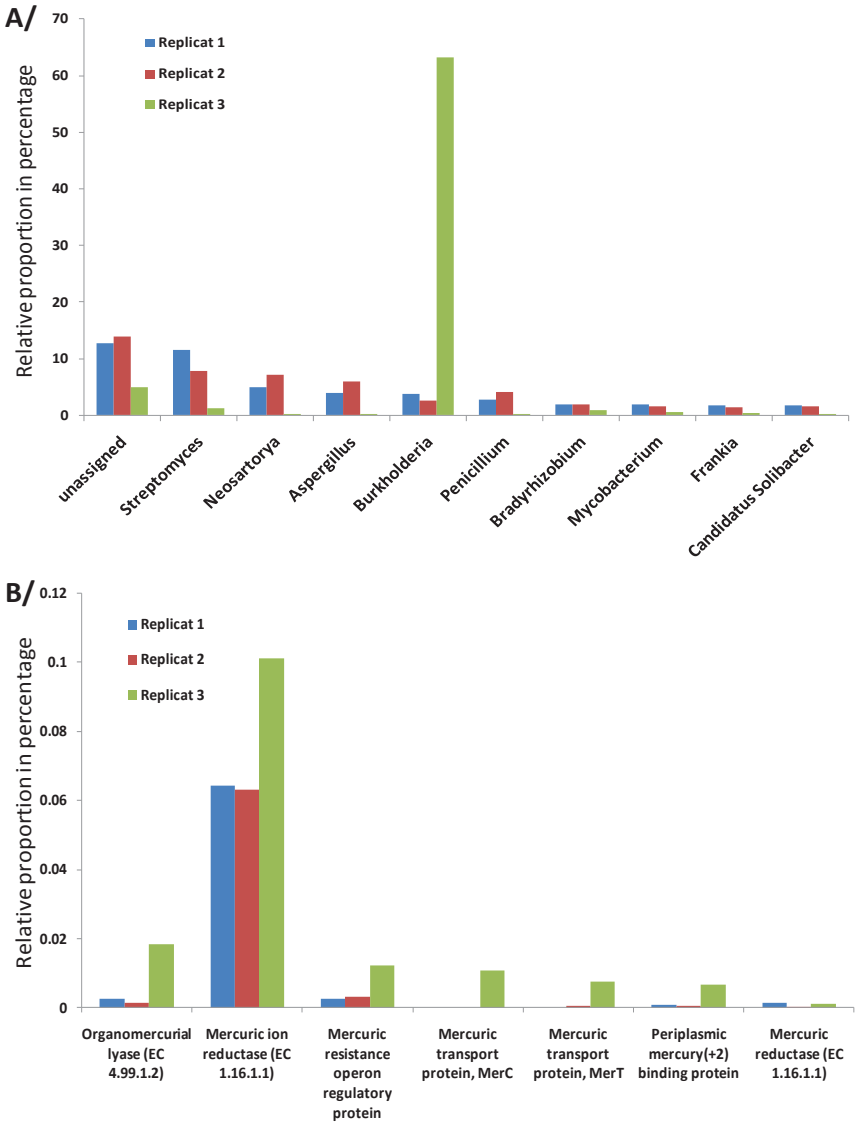


Supplement figure 4: Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for datasets generated from the Rothamsted soil. Horizontal lines are the mathematical averages for the metagenomic datasets from each condition. The p-values are the likelihoods that the distribution differences observed between environments are random.

The case of the mercury enrichment 2:

Due to the low reproducibility of this condition showed by RISA (see supplement figure 2), the three microcosms corresponding to the mercury enrichment 2 were sequenced. Confirming the RISA profiles (supplement figure 2), two of them (called replicates 1 and 2) were similar in term of taxa (see table 1) and functional relative distributions (supplement data), and highly convergent to the third one. In fact, while *Streptomyces*, *Neosartorya*, and *Aspergilus* are predominant in the replicates 1 and 2, *Burkholderia* represents more than 63% of all detected genera in the replicate 3 (supplement figure 5). In contrast, Ascomyceta was more detected in the two first replicates (18.6% and 22.4%) than in the third (less than 0.5%).

When focusing on genes related to mercury resistance, a majority of them are distinctly more represented in the third replicate (supplement figure 4). It is in particular the case of the organomercurial liase, mercuric ion reductase, mercuric resistance operon regulatory protein, mercuric transport proteins MerC and merT and the periplasmic mercury (+2) binding protein.



Supplement figure 5: Relative distribution (in percentage of annotated reads) of functional subsystems (based on SEED assignments of sequenced genomes in the MG-RAST program) for biological replicates corresponding to the mercury enrichment 2.

Global metagenomic comparisons:

The already generated Rothamsted soil datasets correspond to both highly complex (*e.g.*, natural communities, nitrogen condition) and highly simplified metagenomes (*e.g.*, from mercury, heavy metals and ethanol enrichment conditions). The functional distribution of all

these datasets (36 pyrosequencing runs) were compiled and confronted to metagenomes corresponding to various other environments (figure 3).

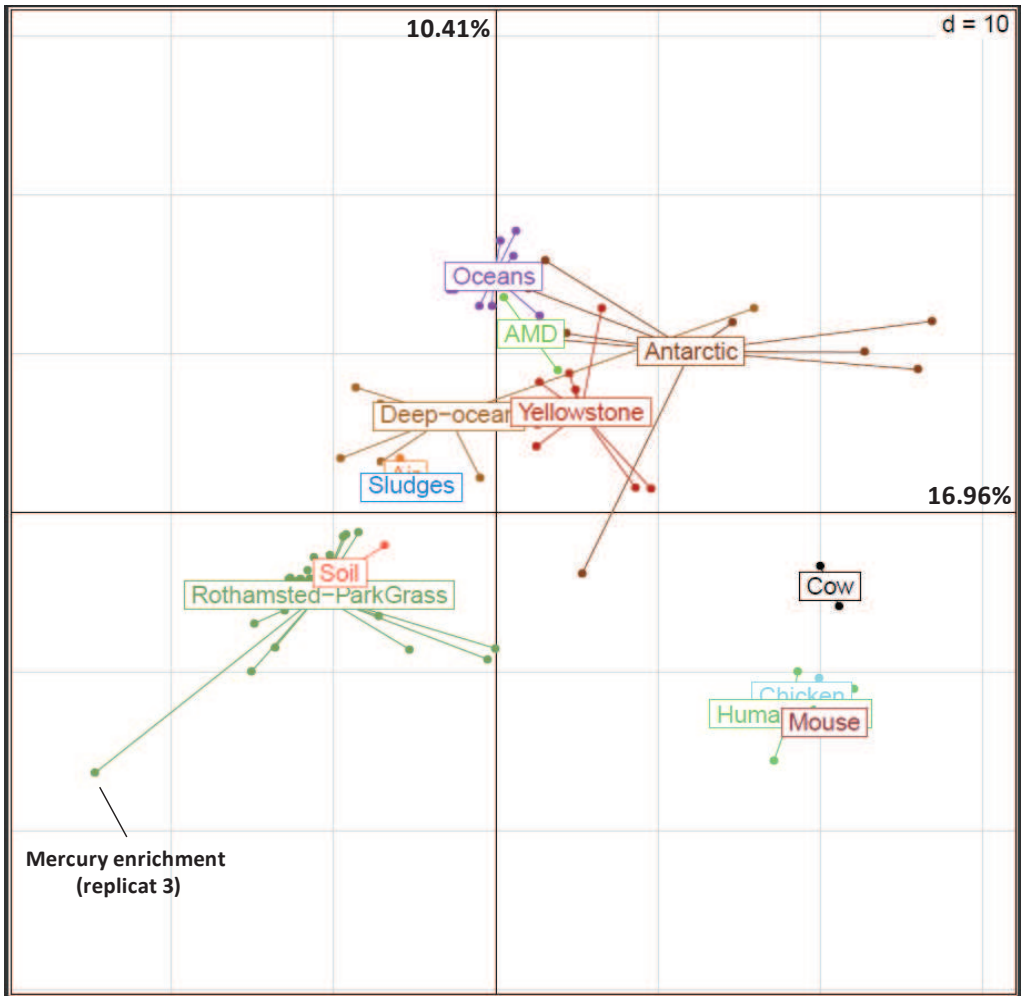


Figure 3: Principal component analysis based on the relative distribution of functional subsystems (level 3) among metagenomes that correspond to various environments. Annotation was done using MG-RAST and an E-value cut-off of 10^{-5}.

In spite of significant structural modifications of natural Rothamsted soil microbial communities during the different incubations, all functional distributions corresponding to this environment are globally grouped together. Thus, at the functional level, all predominant microorganisms from the different conditions possess globally the same functional pool. This functional pool is also similar to soil metagenomes from other places (*e.g.*, from Puerto Rico, North America and Italy; "Soil" group on the PCA).

Assembly efficiency:

To evaluate the assembly efficiency of metagenomes generated from each microcosm condition, assembly software (Newbler v5.0) was independently applied to all condition (using the two duplicates) and the replicate 3 of the mercury enrichment 2. The percentage

of assembled reads provides information about the complexity of datasets and is directly correlated to assembly efficiency. This ratio ranges from 5.67 in control condition (consistent with results from Delmont et al., in press) to 83.05% with the replicate 3 of the mercury enrichment 2. Less than 10% of reads are in contigs with the Nitrogen, diesel and metal enrichment 1. In contrast, this ratio exceeds 50% in ethanol enrichment and mercury enrichment 1.

In the case of the replicate 3 of the mercury enrichment 2 (more efficiently assembled), the dataset was devised into 20 equal fractions of approximately 50 thousands of reads. Then an assembly effort was done using between 1 fraction and 20 fractions. The relative percentage of singletons (unassembled reads) and longest generated contig are presented in the figure 4.

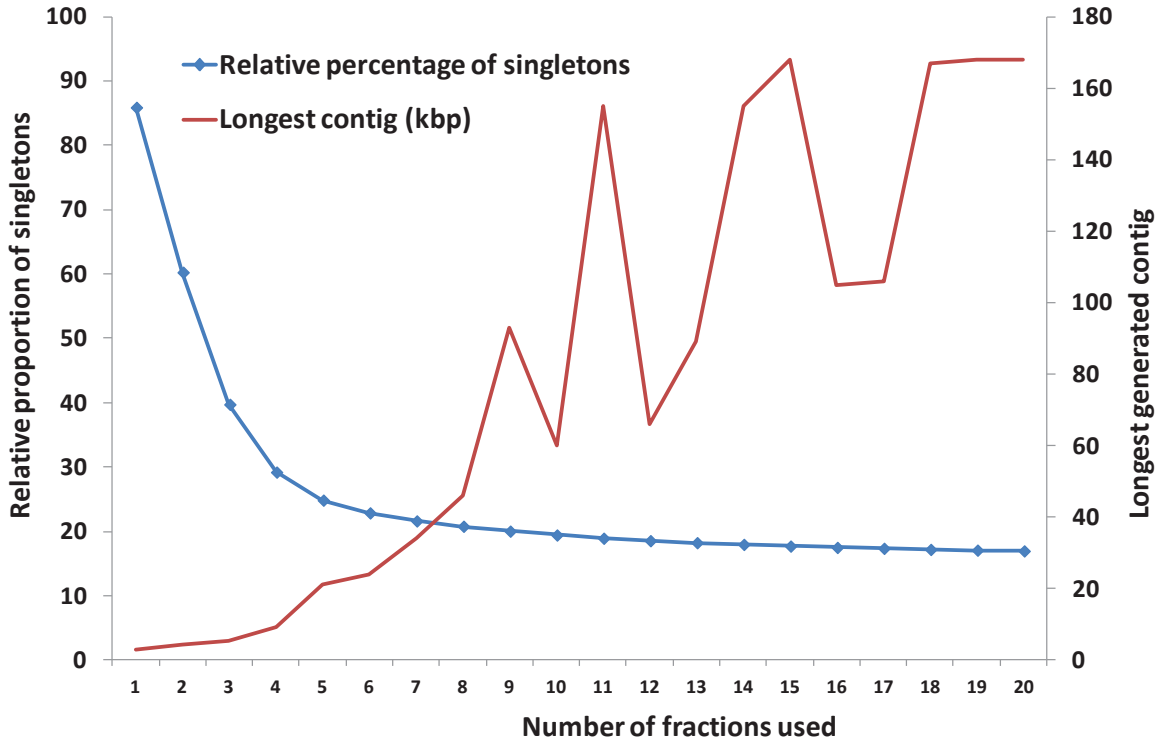


Figure 4: Singleton rarefaction curve using between one and 20 fractions of 50 000 reads from the mercury enrichment 2 (replicat 3) on Newbler v2.5. The second axe represents the longest contig reconstructed (in kbp) as a function of the number of fractions used.

With one fraction, 85.89% of reads are unassembled, and the longest contig is about 2.9 kbp. With the entire dataset (one complete pyrosequencing run), only 16.95% of reads are unassembled and the longest contig is longer than 168 kbp. In contrast, 93% of reads are unassembled using the same number of sequences from natural communities (Delmont et al., in press). While the singleton rarefaction curve in highly homogeneous, the longest contig size fluctuates considerably as a function of the number of used fractions. As an example, it size is longer when using 15 fractions than when using 16 or 17 fractions.

	Control			Nitrogen condition (no oxygen)			Temperature (37°C)		
	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%
contig 1	6.0	8.3	57.4	4.7	6.4	56.6	44.8	9.7	46.7
contig 2	4.9	7.1	57.0	4.6	9.1	56.4	44.3	31.6	42.4
contig 3	4.4	9.6	54.4	4.2	8.9	56.2	41.5	16.6	56.6
contig 4	4.4	9.3	56.3	3.7	7.8	55.9	40.2	10.5	58.6
contig 5	4.3	9.2	54.7	3.7	8.3	55.5	38.0	9.8	43.8
contig 6	4.3	5.0	53.0	3.7	6.1	54.5	36.0	11.1	56.3
contig 7	4.2	6.3	61.9	3.7	6.0	57.0	35.1	10.4	57.1
contig 8	4.2	6.2	54.2	3.7	7.5	56.6	33.0	10.9	57.5
contig 9	4.1	9.7	54.9	3.6	6.6	61.1	29.2	10.6	58.6
contig 10	4.0	8.9	54.5	3.4	5.3	63.8	25.0	10.9	57.0

	Salt enrichment (30g/L)			Salt enrichment (300g/L)			Diesel enrichment (50g/kg)		
	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%
contig 1	20.2	15.5	62.9	93.5	13.6	47.1	8.5	7.2	59.2
contig 2	19.7	15.5	64.9	84.8	13.5	44.2	7.7	5.3	71.4
contig 3	15.1	26.1	64.3	62.9	13.8	46.1	7.2	5.1	54.4
contig 4	13.8	27.0	58.0	60.8	13.7	47.7	7.1	7.2	54.9
contig 5	13.2	24.5	62.9	59.1	13.8	47.1	7.0	6.2	68.5
contig 6	13.2	26.5	61.4	57.4	13.6	46.7	6.9	8.0	56.9
contig 7	13.1	27.4	62.6	57.1	13.5	46.0	6.8	5.2	69.9
contig 8	12.3	25.2	65.1	55.6	13.3	46.2	6.7	6.3	71.1
contig 9	12.1	24.8	63.0	54.9	13.9	43.2	6.2	5.5	74.0
contig 10	12.0	24.8	63.6	52.5	13.7	46.7	6.0	6.6	55.4

	Ethanol enrichment (20%)			Metals enrichment (4 X 0.2g/kg)			Metals enrichment (4 X 2g/kg)		
	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%
contig 1	424.3	21.5	47.6	6.6	8.4	57.0	311.6	23.4	68.9
contig 2	240.9	22.5	46.6	4.3	10.3	55.4	265.2	23.7	68.5
contig 3	236.3	22.7	48.8	4.2	9.3	55.0	207.4	25.2	68.2
contig 4	204.1	21.8	46.3	4.2	9.0	57.0	184.5	24.4	68.0
contig 5	155.1	21.8	48.2	4.1	5.1	66.1	159.6	24.5	67.2
contig 6	151.1	22.3	49.1	3.9	9.4	63.3	155.1	24.2	67.9
contig 7	129.0	21.8	46.5	3.9	9.4	62.6	143.6	25.8	67.6
contig 8	127.6	23.7	47.8	3.8	5.6	68.4	118.1	23.1	69.2
contig 9	124.0	22.8	48.0	3.7	10.5	55.3	110.0	25.1	68.7
contig 10	105.8	22.1	46.1	3.6	7.2	56.6	109.5	25.0	67.2

	Mercury enrichment (0.02g/kg)			Mercury enrichment (0.2g/kg) (rep 1 and 2)			Mercury enrichment (0.2g/kg) (rep 3)		
	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%	Length (kbp)	coverage	GC%
contig 1	254.8	21.7	70.6	73.0	15.1	69.3	168.5	29.8	58.9
contig 2	160.1	11.9	53.1	40.4	24.8	62.7	110.1	19.2	60.7
contig 3	96.8	11.9	56.5	14.2	14.2	69.1	92.3	17.6	60.0
contig 4	87.8	11.5	55.7	11.0	20.2	73.4	76.7	17.6	59.8
contig 5	74.6	11.1	68.8	11.0	8.1	44.9	73.1	18.5	60.2
contig 6	73.0	17.3	69.3	10.5	7.0	48.9	62.7	17.3	62.9
contig 7	69.8	12.8	60.0	9.5	6.5	64.1	62.5	20.7	60.8
contig 8	69.2	11.5	55.7	9.2	7.1	48.9	54.5	17.5	61.8
contig 9	67.0	11.6	58.3	9.0	10.3	66.1	52.3	18.3	61.8
contig 10	65.0	10.8	58.3	9.0	6.8	71.5	49.0	18.2	59.6

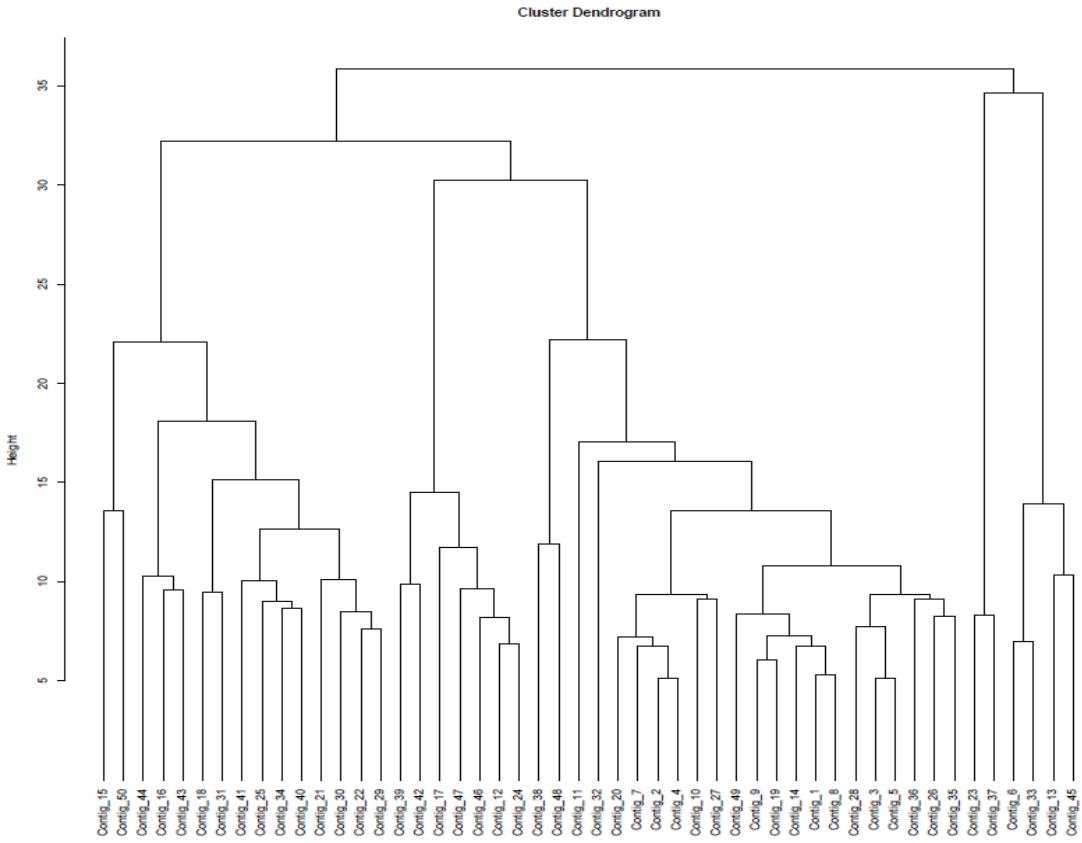
Table 2: Length, coverage and GC content of the ten longer contigs generated with datasets that correspond to different conditions. Except for the mercury enrichment 2 (0.2g/kg) replicat 3, two datasets corresponding to duplicates were used for each assembly effort. Newbler 2.5 was used with default parameters to assemble datasets.

The length, coverage and GC content of the ten longer contigs reconstructed for each condition were summarized in the table 2. The control, Nitrogen, Diesel enrichment and metal enrichment 1 conditions were unable to generate contigs longer than 10 kbp. In contrast, the sequencing effort of DNA pools extracted from high temperature, salt enrichments 1 and 2, ethanol, metal enrichment 2, and finally mercury enrichment 1 and 2 provided quantities of long contigs with a majority of relatively deep coverage (*e.g.*, more than 23 times with the metal enrichment 2). In particular, the ethanol condition provided 252 contigs longer than 10 kbp and 27670 reads (1.47% of the two duplicate datasets) were used to reconstruct the longer one (424.3 kbp). In addition, the GC content of contigs varies considerably between some conditions. In particular, the GC content average of the 10 first contigs is about 47.5% (±1.04) with the ethanol condition and 68.1% (±0.65) in the metal

enrichment 2. This GC content distribution difference can also be observed when comparing unassembled metagenomes from these two conditions. However, important GC content differences were shown among contigs from the same conditions (*e.g.*, high temperature, mercury enrichment 1 and 2 (for replicates 1 and 2) and emphasize the probable presence of distinct predominant genomes in these datasets.

Binning contigs using tetranucleotidique frequency:

The tetranucleotidique frequency (TNF) was compared between contigs generated from the same conditions to help binning them in draft genomes. The Ethanol enrichment condition was used as an example. TNF of the 50 longer contigs was compared in a dendogram (supplement figure 6). 18 contigs were grouped together and represent a total of 2.3 Mbp, with a coverage of 22.49 (±0.62) and a GC content of 47.78% (±1.07). While the TNF can be similar for taxa closely related (Teeling et al., 2004; Woyke et al., 2006), the coherent coverage between contigs emphasizes a similar distribution of these genetic structures in the microbial community. Using the same strategy, contigs were also binned in draft genomes for the salt enrichment 2, metal enrichment 2 and the two mercury enrichments.



Supplement figure 6: Dendogram based on the tetranucleotidic frequency of the 50 longest contigs generated from the ethanol enrichments dataset (using reads generated from the two duplicates).

Draft genomes mining for metagenomic discoveries:

After the reconstruction of long contigs and their binning in draft genomes, they were mined to track genes and genetic structures of interest and to study their genetic environment. An example of reconstructed contig from the mercury enrichment 1 is represented in the figure 5. The GC content of the 261 genes present is provided and shows a relative homogeneity along this contig. Genes related to mercury resistance were studied in contigs generated with the mercury enrichment 1. In the longer contig (figure 5), two mercuric reductase genes and two alkylmercury lyases were detected. In the contig 141, a MerC mercury resistance protein attached to a Hg (II) responsive transcriptional regulator gene were detected. More interestingly, seven genes related to mercury resistance were detected in the same operon in the contigs 160.

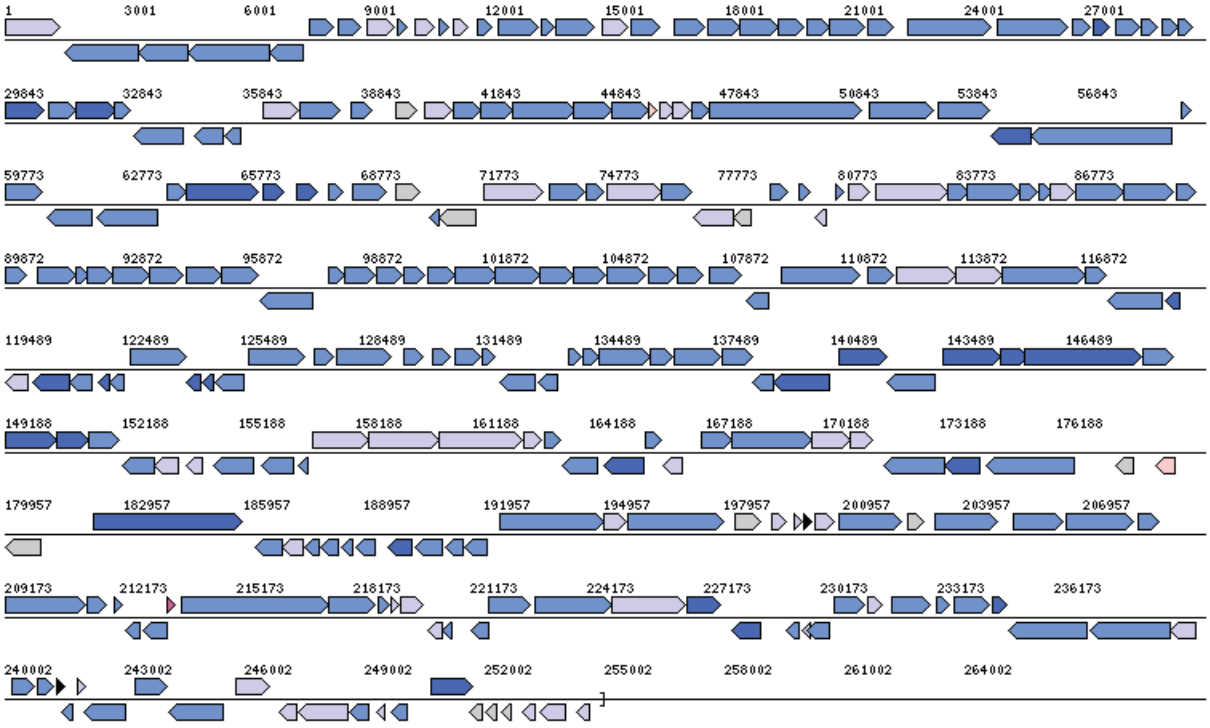


Figure 5: Representation of a reconstructed contig using IMG/M platform from JGI institute.

In the same assembled datasets, four polyketide synthase (PKS) modules and related proteins (from 968 to 1530 amino acids) were detected in one operon of the contig 2. When blasting these genes to the nucleotide collection of NCBI, the maximum of identity with already known polyketide synthases varied between 80 and 84% and only in limited parts of the genes.

Heavy metals related functions were detected in 12 distinct contigs reconstructed with the metal enrichment 2. These genes were often attached to RND family efflux transporters

(MFP subunit), outer membrane proteins, transporters of the NRAMP family and lipoprotein signal peptidase related genes.

Additional sequencing

To reconstruct draft genomes and build additional contigs, a HiSeq paired ends sequencing strategy was performed in tags corresponding to the duplicates from mercury enrichment 1, ethanol enrichment and heavy metals enrichment 2. Approximately 15 million reads were generated for each tag. Assembly efforts are ongoing.

Metagenomes annotation using soil reconstructed draft genomes

After reconstructing draft genomes using both pyrosequences and HiSeq sequences, these genetic structures will be annotated on the RAST server, and integrated on SEED for MR-RAST annotation. The 36 datasets from Rothamsted and those corresponding to other environments (see figure 3) will be re-annotated using reference genomes and these specific draft genomes. The relative distribution of these genomes will be observed on natural soil communities and other environments.

Discussion:

Soil metagenomics access the highly diverse soil microbial community, and therefore, the subsequent reconstructions of their genomes from short reads (150-500pb) are difficult due to the high diversity of the generated datasets. As an alternative approach, microbial communities from the untreated Park Grass Rothamsted soil were incubated under controlled microcosm conditions (see Material and methods section). The main goal of this experimental design was to modify their natural microbial evenness to stimulate metagenomic discoveries and assembly efficiency when applying high throughput sequencing techniques.

After four months of incubation, DNA was extracted using a stringent lysis known to be relatively efficient with this soil (Delmont et al., 2011) and purified. The quantity of extracted DNA was highly influenced by the different incubation conditions (supplement figure 1). While the quantity of extracted and purified DNA cannot be directly correlated to microbial population density due to possible other factors influencing extraction yield, our data shows the advantage of incubation conditions to influence soil microorganisms. For example, the influence of ethanol enrichment, metal enrichment, and mercury enrichment on microbial communities was confirmed using metagenomic approaches (tables 1 and 2). Thus, under a majority of tested conditions, the lower the obtained DNA yield was, the higher modified the designed evenness was. However, while DNA extraction yield did not change with salt enrichment in comparison to controls, these communities were affected. Thus, DNA yield was not sufficient to screen all the efficient conditions prior deep sequencing.

A fingerprint technique (RISA) that provides semi-quantitative information about the evenness of microbial communities was applied to all microcosms (supplement figure 2). Based on the comparison of generated profiles, a majority of conditions impacted considerably soil biodiversity. In some cases, only one pick was observed. However, Salt enrichment 1, metal enrichment 1 and Nitrogen conditions appear to be relatively similar to controls. Due to the apparent high reproducibility between biological replicates and to investigate a maximum of conditions, the relatively massive sequencing effort was done from duplicates only. As an exception and in the main goal of understanding why communities reacted differently in the third replicates, the three microcosms from the mercury enrichment 2 were sequenced.

Generated datasets were annotated using MG-RAST-CLOUD for taxonomical and functional comparisons. Using 13 datasets generated by varying season, depth and DNA extraction protocols to represent a global picture of the natural microbial populations (Delmont et al., in press), the control condition appears to impact the evenness of microorganisms (table 1). In particular *Mycobacterium*, *Brucella* and *Bacillus* genera appear to be stimulated in relative distribution during the incubation. However, this structural modification is relatively limited in comparison to other conditions. For example, the relative distribution of *Burkholderia* increased considerably with the mercury enrichment 1 (17.40%, ±1.95) and in the third replicate of mercury enrichment 2 (63.16%). Interestingly, this genus was not particularly known to resist to high mercury concentrations before this study and could provide new genes and mechanisms involved in this resistance (MerA gene for example was no detected in the draft genome reconstructed from the third replicate of mercury enrichment 2). In addition, S*treptomyces* relative distribution increased in almost all conditions (table 1). Thus species related to this genus appear to resist more than other taxa to a wide range of environmental perturbations (from high temperature to heavy metal enrichment).

As the goal of the experimental design was to access other diversities, most of the tested conditions appear stimulated the detection of particular genus. The strategy that aims to divide a complex system to better conquer it metagenome appears to be realizable and to success on soil environment. In fact, each of these experiments provided access to a distinct evenness (four examples are presented in the figure 6). These taxonomical distributions were more or less reproducible among sequenced duplicates, depending on the condition (highly reproducible with salt and heavy metals enrichment 2 for example, less with mercury enrichment 1 and high temperature).

In addition to structural modifications, studying functional distribution differences between datasets is supposed to help understanding why specific taxa are stimulated in the different condition. However, it is often difficult using metagenomic approaches to separate key functions that help taxa surviving and being more competitive in a specific condition and the associated functions (present in the same genome) that increase in the same level (co-correlation effect) but which are not crucial for adaptation.

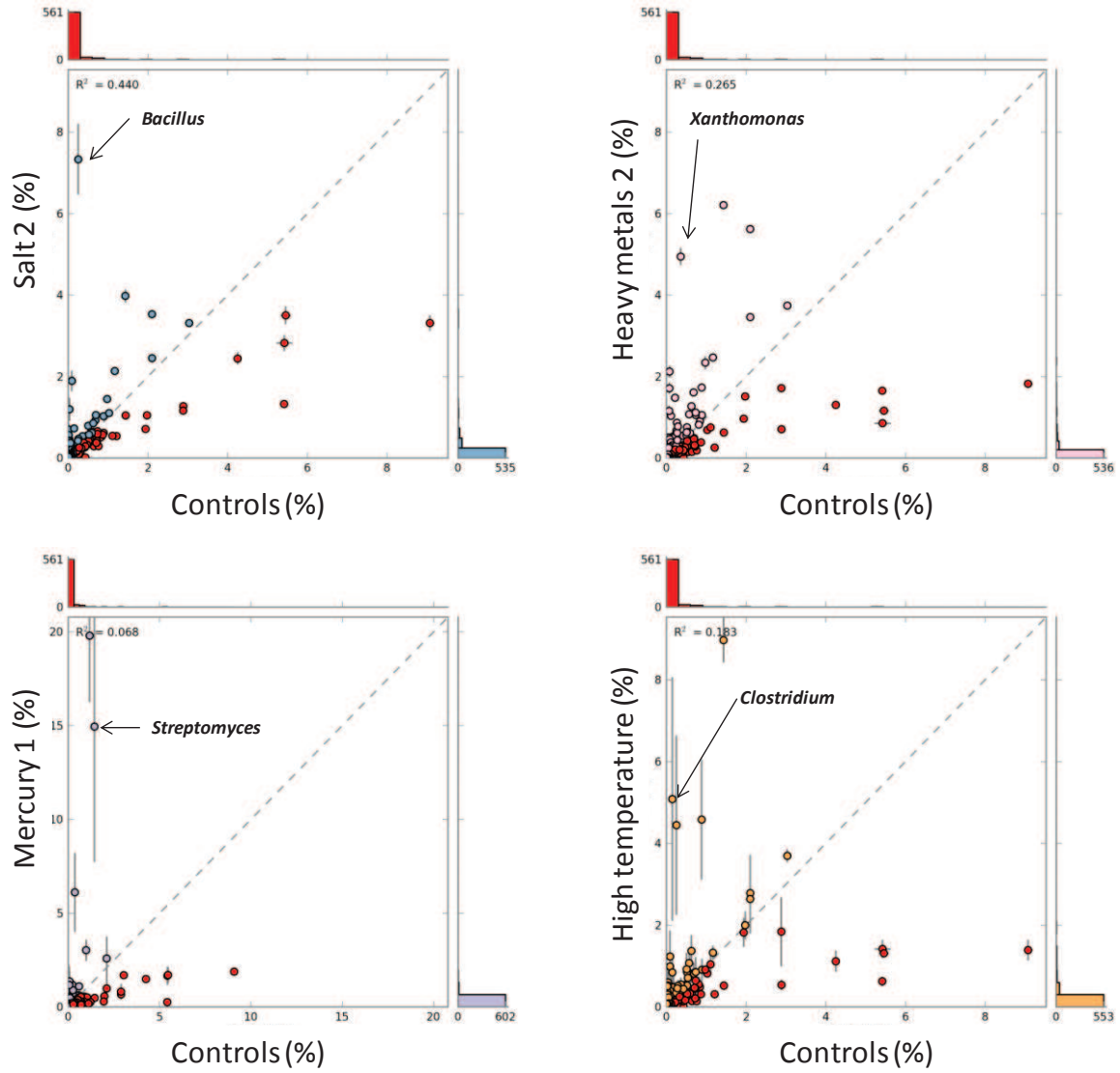


Figure 6: Relative distribution in percentage of detected genera between controls and four distinct conditions: salt enrichment 2, mercury enrichment 1, high temperature (37°C) and heavy metals 2, using MG-RAST-CLOUD and M5NR annotation (e-value<10^{-5}). Lines crossing each genus represent duplicate fluctuations.

Thus, among the hundreds of functions that vary significantly in distribution between conditions, only few of them are probably the main causative effect of taxonomical structure modifications observed. By focusing on processes directly associated to the induced environmental modification, it is possible to select some of these key functions. As an example, the distribution of sequences related to Cobalt-zinc-cadmium resistance protein CzcD increases in the metal enrichment 2 (figure 2). Because the Czc system known to mediate resistance to cobalt, zinc and cadmium through ion efflux (Anton, Grosse et al. 1999) is more present in the condition highly enriched in these three heavy metals (see mat and met section), it is probably a key function to respond to this stress. Thus CzcD probably provides a considerable advantage to microorganisms that possess it in these microcosms.

Genes involved in mercury resistance or alcohol dehydrogenase are other examples of key functions increasing in corresponding conditions (figure 2).

Using both structural and functional differences found between the two first replicates and the third one of the mercury enrichment 2, hypotheses can be proposed to explain this lack of reproducibility (see supplement figure 5). In fact, the third replicate is probably more adapted to mercury concentration with a clear predominance of *Burkholderia* (the same genus that increase with the mercury enrichment 1) and a relative augmentation of almost all genes related to mercury resistance. Thus for an undefined reason, the *Burkholderia* taxa dominating the third replicate was not able to develop two of the three tested microcosms. In the other hand, fungi are considerably more represented in the two other replicates than in natural communities. As a consequence, the lack of bacterial and archaeal resistance in the two first microcosms provided probably an ecological place to specific fungi able to resist to high mercury concentrations. This lack of reproducibility reflects probably the limit of extreme conditions we can apply to soil microorganisms, as only one replicate was able to react efficiently to the induced stress.

Studying the diversity and genetic environment of key functions add information about the genomic structure of microorganisms (*e.g.*, presence of mobile genetic elements that carry specific genes). However, this information cannot be extracted from complex microbial communities when applying classical metagenomic approaches due to limited assembly efficiencies. As an example, the low percentage (5.67%) of assembled sequences was unable to construct contigs longer than 10 kbp with datasets corresponding to the control condition.

In the goal of stimulating the study of microorganisms that predominate in these microcosms, datasets were reconstructed using assembly software (Newbler v5.0) and annotated on IMG/M. The table 2 emphasizes the interest of this approach to stimulate assembly efficiency when studying complex microbial communities like in soil environments. In fact, while some conditions were unable to stimulate assembly efficiency (*e.g.*, Nitrogen condition and metal enrichment 1), a majority of them provided long contigs and in some cases draft genomes (*e.g.*, with the ethanol enrichment and based on coverage, GC content and TNF). These results represent the first successful assembly of a soil metagenome. More interesting, each condition provided a distinct genetic richness of the Rothamsted soil metagenome (table 1). As a consequence, when assembling datasets from all the conditions in one run, the longer contig was only about 292 kbp. The entire dataset was just more complex. Additional sequencing effort were done for duplicates corresponding to ethanol enrichment, mercury enrichment 2 and heavy metals enrichment 2 to stimulate the reconstruction of contigs, and not only for the few number of predominant genomes. Bioinformatics analyze of these new datasets is ongoing.

Because soil metagenome is an untapped genetic gold mine, several genes of interest were discovered from the first reconstructed draft soil genomes. For example, PKS were detected

in a draft genome sequenced from to the mercury enrichment 1. These enzymes are known to produce secondary metabolites (*e.g.*, antimicrobial metabolites, (Cane, Walsh et al. 1998) and were already tracked in soil using metagenomic approaches (*e.g.*, using fosmid library screening, (Ginolhac, Jarrin et al. 2004). However, the genes discovered from this contig appear to be relatively distinct to already studied PKS, and so could produce new secondary metabolites. Additional analyses have to be performed with these genetic structures to study their potential. Interestingly, two glycosyltransferase related genes were present in the same operon. They are known to contribute to specific interactions between bioactive natural products (*e.g.*, antibiotics) and the biological target (*e.g.*, (Mendez and Salas 2001)). Thus these glycosyltransferases probably interact with the metabolites produced by these PKS for an undefined ecological function.

Based on the first sequencing effort done, the third replicate from the mercury enrichment 2 was the most efficiently metagenomically assembled (more than 80% of assembled reads, see figure 4). Based on this highly favorable ratio using only one million reads, a legitimate issue could be to know if we are close to an entire simplified soil metagenome assembly. Thus a singleton rarefaction curve was defined using 20 fractions of 50 000 reads. Based on this curve, we are still far away to sequence and assemble entirely even a highly simplified soil metagenome. However, contigs reconstruction is relatively efficient in comparison to controls for examples (see figure 4 and table 2), thus it is not necessarily to sequence entirely a simplified metagenome to reconstruct genomes. In addition, when one genome is largely dominant in DNA pool extracted from a microcosm, generated dataset could represent more an *in situ* culture than a soil metagenome, with some residual metagenomic contaminants impacting the singleton rarefaction curve and preventing the entire assembly of the dataset. If we consider this novel *in situ* culture approach, perspectives are considerable. It could be possible to “cultivate” several highly resistant microorganisms in their environment applying an induce stress, and then to modify conditions or just let time doing it office to observe genomic reorganizations (*e.g.*, gene duplication, mobile genetic elements role) by assembling again and again their genetic structure. To emphasize the interest of this approach, the Burkholderia draft genome will reconstructed after 4 months and 16 months of incubation with mercury enrichment 2. Analyze efforts will be made to observe hypothetical additional adaptation capacities of this taxa with time.

By compiling all generated datasets from Rothamsted Park Grass soil, it is possible to confront their functional potential to those observed in other environments (figure 3). Interestingly, while the microbial structure of natural communities was highly modified during the different incubations, all the datasets related to Rothamsted are grouped together and to soils from other countries. Thus, it is possible that a majority of microorganisms from soil, whatever they are highly or lowly represented in natural communities, possess globally the same functional pool. As a consequence, we need to enrich specific genomes based on the low percentage of unusual genes they possess (*e.g.*, mercury resistance genes, alcohol dehydrogenase genes) to be able to reconstruct them.

Draft genomes have now to be annotated on RAST (http://rast.nmpdr.org/) and integrated on the MG-RAST-CLOUD SEED database. The 36 datasets from Rothamsted and those from other environments will then be re-annotated with this SEED database enriched on Rothamsted draft genomes. The relative distribution of these genomes will be observed in various habitats, and in the natural Rothamsted soil. It will be then possible to know how low their distributions before the particular incubations that allow their reconstructions are. The global annotation of soil metagenomes, but also those from other environments could potentially be improved; even it is by the thousandth of a percentage. To divide and conquer the entire soil metagenome is possible. Because the task is considerable, we need more than ever the Terragenome international effort to access the soil metagenome gold mine. Finally, other complex communities (*e.g.*, oceans, sediments) could be mined using the same strategy.

Conclusion:

Complex microbial communities are currently largely untargeted during metagenomic surveys due to the presence of few predominant organisms that limit the access of the large majority of lowly represented taxa. This boundary is mainly due to the presence of a single evenness. Here, we proved that to design soil microbial evenness of interest in controlled environmental conditions prior sequencing efforts is a suitable approach to access other diversities and stimulated the reconstruction of naturally lowly represented genomes. As the number of condition that can be applied to complex environments is substantial, we will be limited only by our imagination constructing experimental designs to open their biological black boxes and mine genetic richness. Thus, opportunities to access lowly represented genomes from complex microbial communities are considerable. This enriched genomes assembly proposal aims to complement what can be find by microbial ecologists tracking natural evenness of interest from across the planet to reconstruct the genetic structure of the 10^{30} estimated microorganisms that pullulate at the surface of the third planet of our solar system.

References:

Allwood, A. C., M. R. Walter, et al. (2006). "Stromatolite reef from the Early Archaean era of Australia." Nature 441(7094): 714-8.

Anton, A., C. Grosse, et al. (1999). "CzcD is a heavy metal ion transporter involved in regulation of heavy metal resistance in Ralstonia sp. strain CH34." J Bacteriol 181(22): 6876-81.

Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res 32(Database issue): D138-41.

Bertin, P. N., A. Heinrich-Salmeron, et al. (2011). "Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics." ISME J 5(11): 1735-47.

Cane, D. E., C. T. Walsh, et al. (1998). "Harnessing the biosynthetic code: combinations, permutations, and mutations." Science 282(5386): 63-8.

Crawley, M. J., A. E. Johnston, et al. (2005). "Determinants of species richness in the Park Grass Experiment." Am Nat 165(2): 179-92.

Delmont, T. O., C. Malandain, et al. "Metagenomic mining for microbiologists." ISME J.

Delmont, T. O., P. Robe, et al. (2011). "Accessing the soil metagenome for studies of microbial diversity." Appl Environ Microbiol 77(4): 1315-24.

Dinsdale, E. A., R. A. Edwards, et al. (2008). "Functional metagenomic profiling of nine biomes." Nature 452(7187): 629-32.

Gans, J., M. Wolinsky, et al. (2005). "Computational improvements reveal great bacterial diversity and high metal toxicity in soil." Science 309(5739): 1387-90.

Gilbert, J. A., F. Meyer, et al. (2010). "The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010." Stand Genomic Sci 3(3): 249-53.

Ginolhac, A., C. Jarrin, et al. (2004). "Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones." Appl Environ Microbiol 70(9): 5522-7.

Griffiths, R. I., A. S. Whiteley, et al. (2000). "Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition." Appl Environ Microbiol 66(12): 5488-91.

Grzymski, J. J., A. E. Murray, et al. (2008). "Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility." Proc Natl Acad Sci U S A 105(45): 17516-21.

Hazen, T. C., E. A. Dubinsky, et al. (2010). "Deep-sea oil plume enriches indigenous oil-degrading bacteria." Science 330(6001): 204-8.

Hess, M., A. Sczyrba, et al. (2011). "Metagenomic discovery of biomass-degrading genes and genomes from cow rumen." Science 331(6016): 463-7.

Kostka, J. E., O. Prakash, et al. (2011). "Hydrocarbon-degrading bacteria and the bacterial community response in Gulf of Mexico beach sands impacted by the Deepwater Horizon oil spill." Appl Environ Microbiol.

Markowitz, V. M., N. N. Ivanova, et al. (2008). "IMG/M: a data management and analysis system for metagenomes." Nucleic Acids Res 36(Database issue): D534-8.

Mendez, C. and J. A. Salas (2001). "Altering the glycosylation pattern of bioactive compounds." Trends Biotechnol 19(11): 449-56.

Meyer, F., D. Paarmann, et al. (2008). "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes." BMC Bioinformatics 9: 386.

Narasingarao, P., S. Podell, et al. (2011). "De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities." ISME J.

Niu, B., L. Fu, et al. (2010). "Artificial and natural duplicates in pyrosequencing reads of metagenomic data." BMC Bioinformatics 11: 187.

Overbeek, R., T. Begley, et al. (2005). "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." Nucleic Acids Res 33(17): 5691-702.

Parks, D. H. and R. G. Beiko (2010). "Identifying biologically relevant differences between metagenomic communities." Bioinformatics 26(6): 715-21.

Pointing, S. B., Y. Chan, et al. (2009). "Highly specialized microbial diversity in hyper-arid polar desert." Proc Natl Acad Sci U S A 106(47): 19964-9.

Qin, J., R. Li, et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature 464(7285): 59-65.

Ranjard, L., E. Brothier, et al. (2000). "Sequencing bands of ribosomal intergenic spacer analysis fingerprints for characterization and microscale distribution of soil bacterium populations responding to mercury spiking." Appl Environ Microbiol 66(12): 5334-9.

Roesch, L. F., R. R. Fulthorpe, et al. (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." ISME J 1(4): 283-90.

Torsvik, V., L. Ovreas, et al. (2002). "Prokaryotic diversity--magnitude, dynamics, and controlling factors." Science 296(5570): 1064-6.

Tringe, S. G., C. von Mering, et al. (2005). "Comparative metagenomics of microbial communities." Science 308(5721): 554-7.

Tyson, G. W., J. Chapman, et al. (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature 428(6978): 37-43.

Venter, J. C., K. Remington, et al. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science 304(5667): 66-74.

Vogel, T. R., V. Y. Dombrovskiy, et al. (2009). "Has the implementation of EVAR for ruptured AAA improved outcomes?" Vasc Endovascular Surg 43(3): 252-7.

Whitman, W. B., D. C. Coleman, et al. (1998). "Prokaryotes: the unseen majority." Proc Natl Acad Sci U S A 95(12): 6578-83.

Chapter 4. Perspectives

1. How to avoid pitfalls in the metagenomic jungle?
2. Future possible microbial ecology directions:
   a. Digitizing microbial genetic structures before synthetizing new forms of life: from knowledge to imagination jumping evolution
   b. Synthetizing microbial life to optimize Martian terraformation labors: from the terrestrial evolution to in lab extraterrestrial adaptation experiments

In the chapters 2 and 3, efforts were done to respectively characterize and mine the Rothamsted soil metagenome. Using the relative experience I acquired studing this soil, I propose some perspectives in the field of metagenomic approaches.

In the section 1, an alternative approach was tested to study metagenomes by integrating methodological fluctuations to maximize standard deviations. As a main conclusion of the experiment, methodological fluctuations do not prevent inter-environmental metagenomic comparisons but can alter punctually results, so inducing faLse positives during the interpretation of datasets. To integrate these fluctuations aims to limit these errors. In addition, intra-environmental comparisons are difficult to perform using this approach, due to methodological fluctuations bigger than natural differences.

In the sections 2 and 3, future possible microbial ecology directions based on metagenomic discoveries and synthetic life possibilities are presented. I am involved in the digitalization of microbial genetic structures that aim to descript communities and discover new taxa and functions. However, the goal of this sequencing effort could be more to create an exhaustive digitalized functional library than to understand how communities live and interact across the planet. In fact, using this functional library, it will be possible in a near future to synthesize automatically microbial genomes optimized to perform particular processes. I have personally no particular expertise on synthetic life, but I am working on the construction of the functional library. As a consequence, I speculated about the future of metagenomics and synthetic biology (section 2) and how to use the generated knowledge for ambitious projects, like the Martian terraformation (section 3).

How to avoid pitfalls in the metagenomic jungle?

Tom O. Delmont, Pascal Simonet and Timothy M. Vogel

Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France

Abstract: A majority of microbial ecologists are attracted by the potential of metagenomic approaches to by-pass cultivation limits. Metagenomics provide positive perspectives but rely on the complete recovery of DNA and the use of appropriate and accurate bioinformatics tools to decrypt and compare microbial function and structure. Metagenomic approaches are currently imperfect and therefore limits and critical steps have to be defined and optimized. In particular, DNA extraction strategy and reads annotation process appear to modify considerably the distribution of functions and species in metagenomes. The problem is that we don't know which approach reflects for the best the *in situ* genetic diversity. In addition, these distribution variations react differently as a function of the environment, so complicating appreciably global comparisons. We present here the effects of these critical steps on metagenomic analyses in order to evaluate the risk of erroneous interpretations of the microbial function and structure in these ecosystems. Regarding the importance of metagenomic pitfalls and with the prospect of an alternative experimental design strategy, a debate is necessary and could adjust future surveys.

Key words: Metagenomic, distribution fluctuation, biases

Introducton:

Microorganisms have colonized almost all the surface of the planet since emerging three billion years ago (Altermann and Kazmierczak, 2003, Allwood et al., 2006) and are now essential actors of climate (e.g., Falkowski et al., 2001), animal survival (e.g., Turnbaugh et al., 2006) and agricultural advances. They represent also a considerable potential for the discovery of novel enzymes and metabolites although in some cases they pose a risk for human health.

They have been actively studied for centuries due to their complexity and importance for humanity. However, only few species have been isolated for further experiments due to cell culture limitations (Amann et al, 1995). Metagenomics aspires to study not only cultivable microorganisms but the entire biodiversity present by extracting DNA directly from an environment (sample).

Metagenomic approaches were first used to express genes from uncultured microorganisms by cloning metagenomic fragments into cultivable species (Handelsman et al, 1998). This effort stimulated the discovery of genes of interest that were difficult to access when following conventional approaches (e.g., Gillespie et al, 2002; Demaneche et al, 2008). In addition, extracted DNA was directly sequenced to study environmental microbial genomics (e.g., Venter et al, 2004; Tyson et al, 2004). Using metagenomic sequences, functional and taxonomical distributions were compared from metagenomes corresponding to different environments (e.g., Tringe et al, 2005; Dinsdale et al, 2008; Delmont et al, perspective).

With current sequencing technology, these distributions became statistically stable due to the billions of sequences generated per metagenome (e.g., Qin et al, 2010; Hess et al, 2011). However, in spite of the evident interest in sequencing environmental metagenomes, some crucial methodological steps strongly impact scientific results. In particular, Lysis stringency (tested in the Rothamsted soil, Delmont et al., in press) and reads annotation (reflected by the maximal E-value or the minimum percentage of identity selected; tested in soil, ocean and human feces microbial related datasets) modify considerably the distribution of function and species among metagenomes. In addition, these distributions vary as a function of the environment, thus complicating appreciably global comparisons.

One difficulty is the lack of knowledge about which approach provides the less biased view of a metagenome. So standards (e.g., DNA extraction protocol and E-value) cannot be defined and metagenomic approaches are limited and provide an altered vision of environmental microbial populations due to arbitrarily selected approaches. The impact of these steps for metagenomic analyses was described in order to stimulate the scientific community debate.

Material and methods:

6 X 3 available metagenomes corresponding to three distinct environments (ocean, soil and human feces) were selected based on their interest in term of global comparisons (different DNA extraction protocols for the soil, and different projects and sequencing technologies for ocean and human feces microbiomes) and the quantity of sequences. For all metagenomes generated using a pyrosequencing technology, artificial duplicates were deleted using the cd-hit-454 software.

For the soil related metagenomes, a trimming approach was tested: nucleotides for what the sequencing quality was lower than 25/40 (Roche criteria) were replaced by an "N". Then all sequences which possess more than 10% of "N" nucleotides were deleted. In addition to the suppression of artificial duplicates, 66.61% (±3.8) of the datasets was deleted prior to the annotation in these cases.

Metagenomes where then annotated using the MG RAST server (Meyer et al, 2008) by varying the maximal E-value (from 10^{-5} to 10^{-60}) and minimum percentage of identity (from 0% to 90%). Tables corresponding to these annotations were exported for further analyses. Principal component analyses and box plots were performed using the R software. STAMP software (Parks and Beiko, 2010) was used for statistical analyses. The Fisher's exact test was selected with a p-value threshold of 0.05 to define functional subsystems which vary significantly in distribution between two metagenomes (figure 1).

Results

Part I: Rothamsted Park Grass soil:

Dataset used here to illustrate potential pitfalls is derived in part from sequencing of the park Grass experiment in Rothamsted.

<u>Typical analysis</u>

Metagenomes were generated from the Rothamsted soil by sampling during different seasons, at different depths and with different protocols to extract DNA (Delmont et al, in press). Each metagenome possesses approximately one million of reads of 385 pb. This methodological effort emphasized an unexpected stability over time and important DNA extraction biases so confirming precedent results based on molecular tools (Delmont et al, 2011). An example of metagenomic distribution fluctuations is presented in the figure 1 (panel D). These fluctuations are observed by using the same annotation process (maximal E-value of 10^{-5} and no percentage of identity threshold) and 30.40% of detected functional subsystems were considered to possess a statistically different distribution between the two DNA extraction approaches. As examples, the distribution of the subsystem corresponding to CO_2 uptake (carboxysome) varies from 0.561% to 1.437%, cellulosome from 0.012% to

0.081%. In comparison, 0.12%, 6% and 7.45% of subsystems have a distribution varying significantly ($p<0.05$) when comparing the same metagenome before and after deleting artificial duplicates (panel A), before and after a particular trimming process (panel B), and between two metagenomes generated with the same DNA sample (technological reproducibility, panel C) respectively.

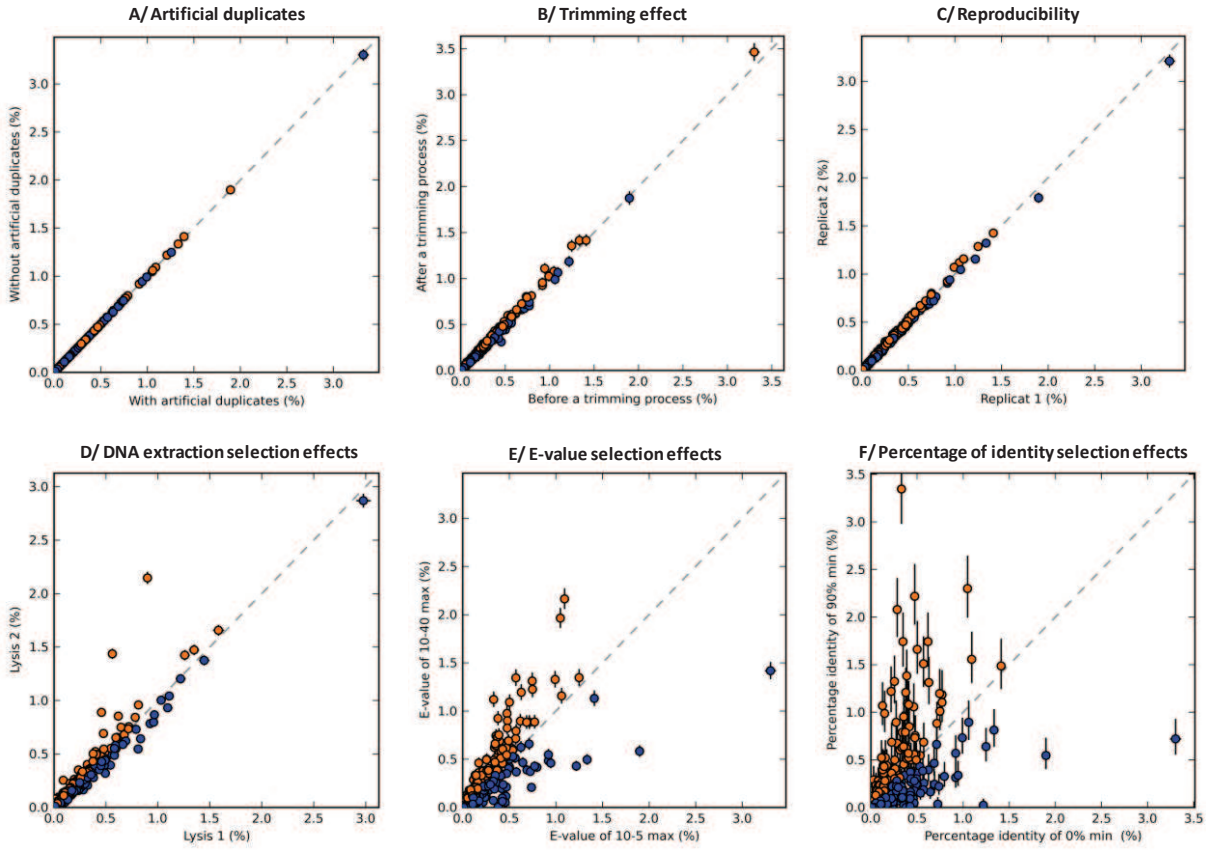


Figure 1: Relative functional distribution comparison of metagenomes in percentage. Metagenomes were annotated using the MG RAST software and tables were exported to the STAMP software for comparisons. The two axes represent the relative distribution in percentage of functional subsystems in two datasets. A/ The same metagenome before and after deleting artificial duplicates. B/ The same metagenome before and after a trimming process. C/ Two metagenomes generated from the same DNA sample. D/ Two metagenomes corresponding to the same soil sample but by applying two different cell lysis stringencies. E/ The same metagenome annotated by varying the E-value threshold. F/ The same metagenome annotated by varying the percentage of identity threshold.

Parameters were changed during the annotation process for the same Rothamsted soil metagenome. Two parameters modify considerably the distribution of this metagenome: E-value and percentage of identity.

E-value

E-value cut-offs aim to limit the inexactitude in comparing reads to the database of previously annotated genes. E-values are in a sense the statistical probability of finding the sequence by chance in a database. When comparing the functional distribution by selecting

a maximal E-value of 10^{-5} and of 10^{-40}, 54.21% of functions possess a distribution varying significantly (figure 1, panel E). Subsystem corresponding to bacterial cAMP signaling varies from 3.303% to 1.418%, Ton and Tol transport systems from 1.898% to 0.583%, CO_2 uptake (carboxysome) from 0.456% to 0.053%, and Sigma B stress response regulation from 0.169% to 0.022%. All these subsystems decrease when the stringency of annotation increases (E-value maximal of 10^{-40}). In contrast, the proportion of sequences related to bacterial RNA polymerase and tRNA aminoacylation for example increases when the maximal E-value decreases (from 0.332% to 1.120% and from 1.093% to 2.163% respectively).

Percent similarity

Percent similarity is an exact requirement of the similarity between the read and the hypothetical match in the database. When comparing the percentage of identity threshold (here between 0% and 90%) 29.18% of the subsystems vary significantly (figure 1, panel F). The proportion of sequences related to bacterial RNA polymerase (from 0.332% to 3.345%), serine glycoxylate cycle (from 1.048% to 2.299%), universal GTPases (from 0.475% to 2.218%), 16S rRNA (from 0.286% to 2.079%) and 23S rRNA (from 0.502% to 1.661%), bacterial proteasome (from 0.348% to 1.742%), respiratory complex I (from 0.619% to 1.742%), TCA cycle (from 0.571% to 1.510%), protein chaperones (from 0.389% to 1.382%), bacterial transcription factors (from 0.22% to 1.219%) and GroEL-GroES (from 0.126% to 1.068%) increases significantly when selecting a minimum percentage of identity of 90%. Subsystems related to bacterial cAMP (from 3.303% to 0.72%), ton and tol transport systems (from 1.898% to 0.546%), bacterial chemotaxis (from 1.335% to 0.813%), cobalt-zinc-cadmium resistance (from 1.249% to 0.61%), and multidrug resistance efflux pumps (from 0.645% to 0.163%) are less represented when using this threshold.

Six metagenomes corresponding to the same Rothamsted soil by varying the season, the depth and the DNA extraction approach where then selected and annotated by varying annotation process parameters. When the stringency of annotation increases (E-value or percentage of identity), the percentage of annotated sequences decreases considerably (see figure 2 and supplement figure). With a maximal E-value of 10^{-5}, 35.66% (±2.55) of sequences are annotated. When the threshold is 10^{-40}, only 1.28% (±0.43) of the metagenome is annotated. This variation of annotation impacts the functional distribution in the same direction for the six metagenomes (figure 2). For example the proportion of sequences related to protein metabolism increases when the annotation decreases. In contrast, subsystems corresponding to regulation and cell signaling and virulence are less represented when the stringency of annotation increases. These subsystems vary in a similar way when it is not the E-value but the percentage of identity which varies (see supplement figure).

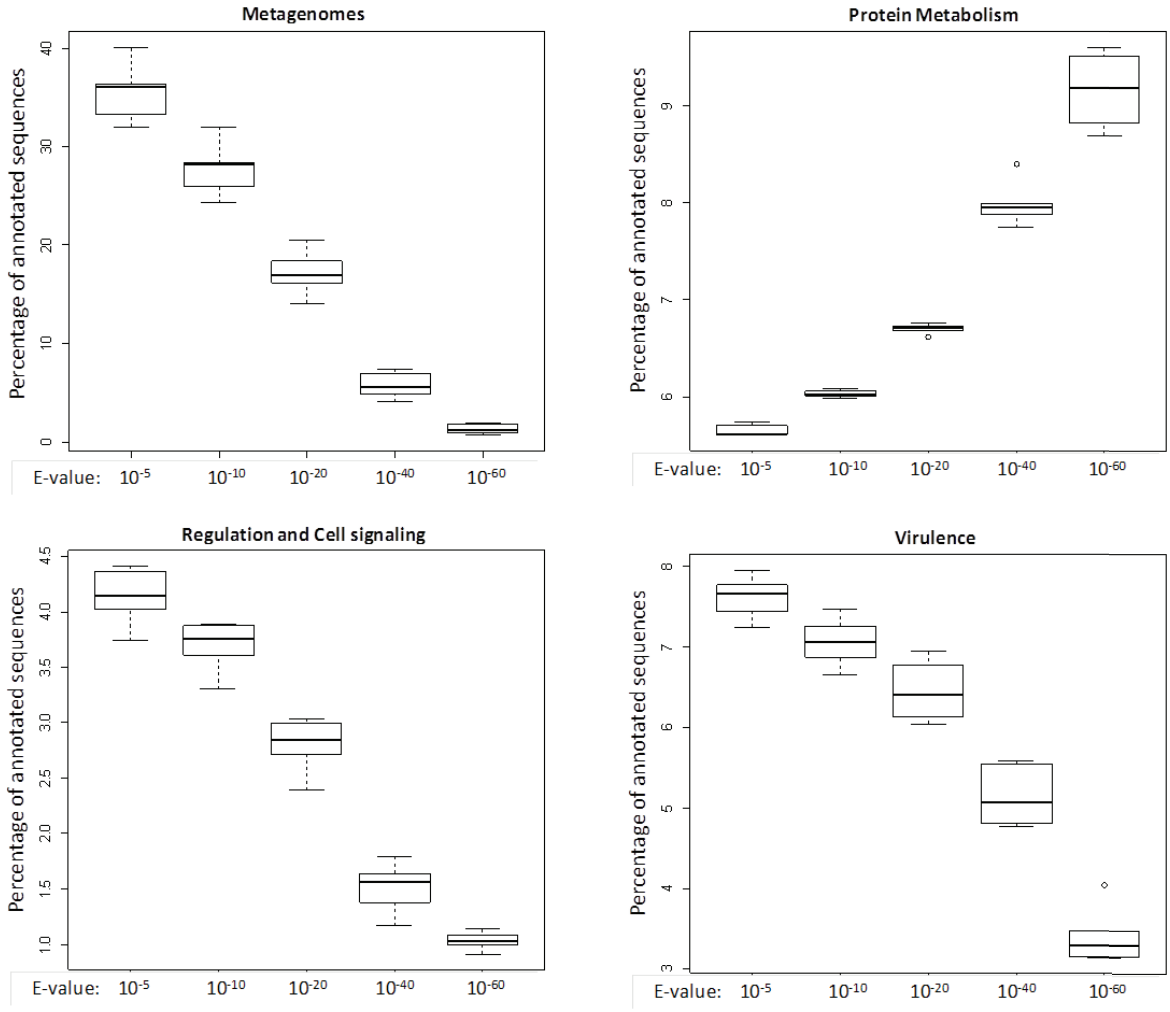


Figure 2: Proportions of annotated sequences in percentage (for all sequences, protein metabolism, regulation and cell signaling, and virulence) are compared by varying the E-value threshold. Box plots are based on the distribution fluctuations among six selected metagenomes corresponding to the Rothamsted Park Grass soil.

The functional distribution of the Rothamsted Park Grass soil was redefined by integrating these methodological fluctuations (figure 3). With six metagenomes and five different E-values (from 10^{-5} to 10^{-60}) 30 distributions were compiled for the same soil and analyzed. For some general functional subsystems the fluctuation is limited: the distribution of sequences related to stress response varies from 2.13% to 2.94%, fatty acids and lipids from 1.29% to 1.83% and nitrogen metabolism from 0.61% to 0.92%. However, the proportion of sequenced related to carbohydrates (from 13.16% to 20.75%), virulence (from 3.14% to 7.96%), regulation and cell signaling (from 0.91% to 4.41%), and motility and chemotaxis (from 0.70% to 3.10%) for example fluctuates noticeably as a function of the method used.

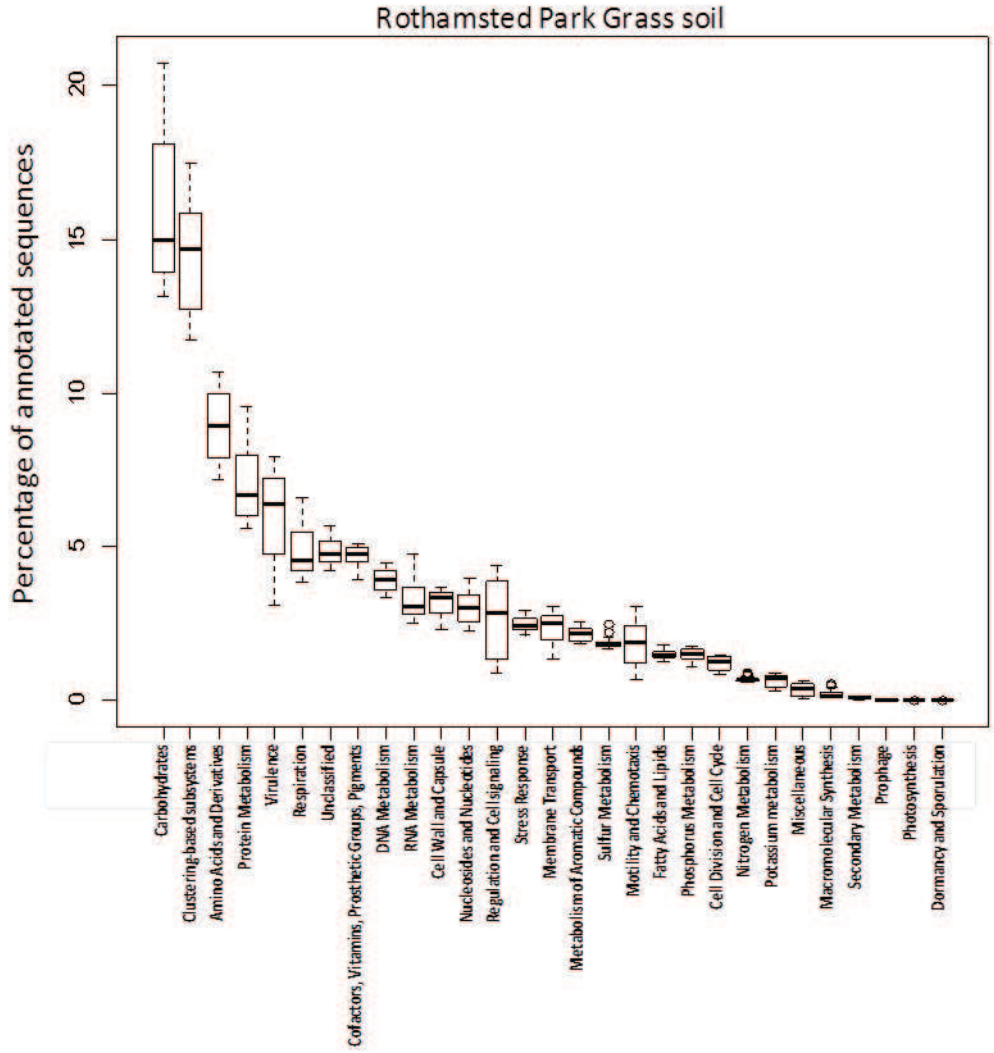


Figure 3: Relative distribution in percentage of general functional subsystems corresponding to the Rothamsted Park Grass soil metagenome by varying seasons, depths, DNA extraction approaches and E-value thresholds (from 10^{-5} to 10^{-60}). Box plots correspond to the fluctuation between 30 distributions.

Part II: Soil, ocean and human feces distribution fluctuations:

To study the impact of annotation processes in other habitats, some metagenomes related to ocean and human feces environments were selected and added to the soil metagenomes.

With a maximal E-value of 10^{-5}, sequences related to carbohydrate for example are more represented in human feces (15.22%, ±0.43) than in soil microbial populations (13.67%, ± 0.27) (figure 4). However, the distribution of this subsystem increases more quickly in the soil related metagenomes when the E-value threshold decreases and with an E-value of 10^{-40} the subsystem is more represented in average in soil (17.60%, ±0.51) than in human feces

(16.72%, ±1.05). Similar differences can be observed when comparing other subsystems (supplement data).

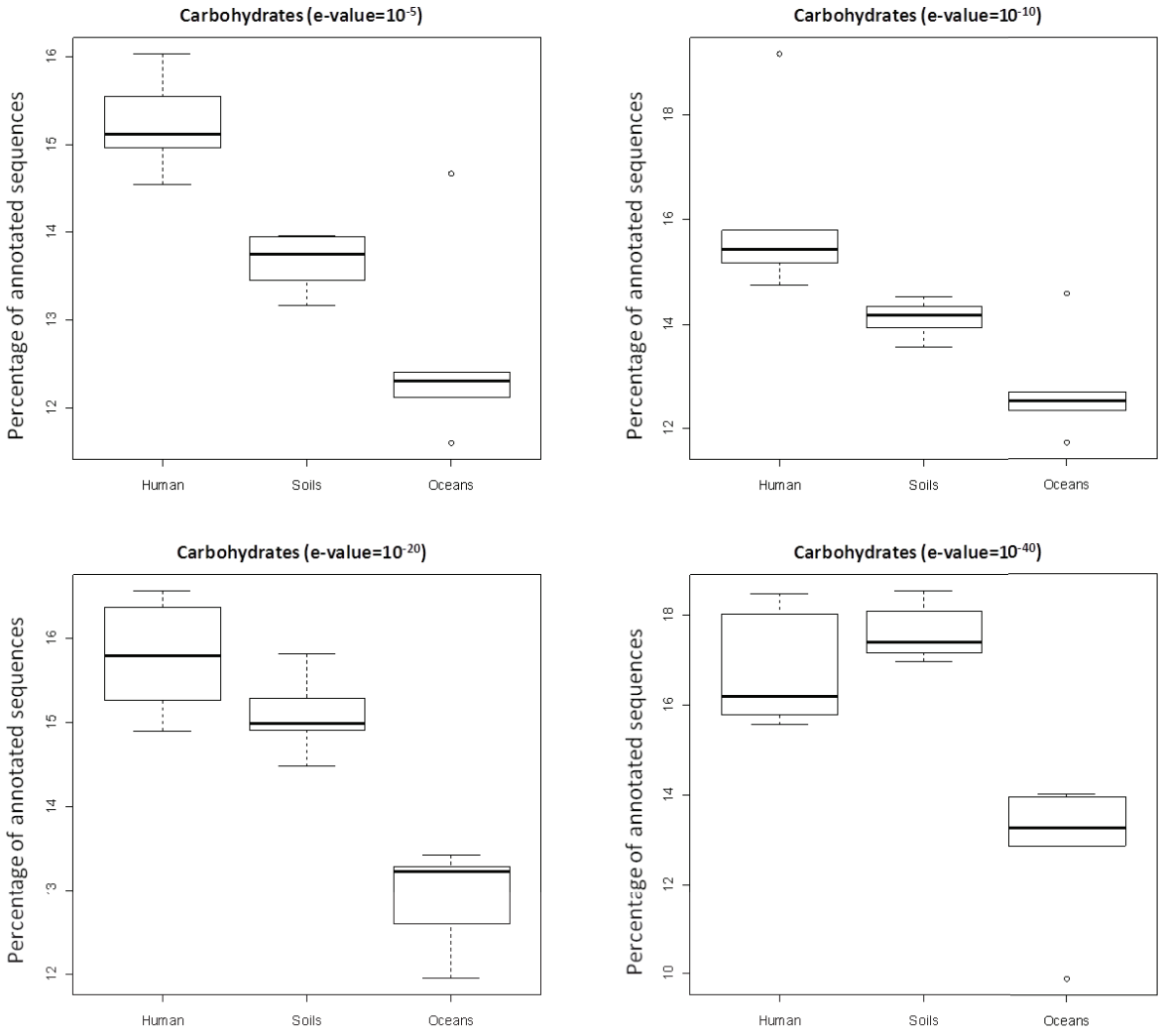


Figure 4: Relative distribution in percentage of the general subsystem corresponding to Carbohydrates in three distinct environments (soil, ocean and human feces) as a function of the E-value threshold). Box plots are based on the distribution fluctuations among six selected metagenomes per environment.

When comparing the distribution of the 819 detected functional subsystems among these 18 metagenomes by integrating the annotation fluctuations (E-value threshold of 10^{-5}, 10^{-10}, 10^{-20} and 10^{-40}), the 72 compared distributions are grouped as a function of the environment (figure 5, panel A). In fact, a majority of functions are more represented in one environment in spite of the methodological fluctuations (figure 5, panel B). Sequences related to tetracycline resistance (ribosome protection type) are more represented in human feces, cobalt-zinc-cadmium resistance in the Rothamsted soil and DMSP breakdown in oceans. However the proportion of sequences related to tRNA aminoacylation for example cannot be defined with a sufficient sensibility to conclude about distribution differences in these three environments.

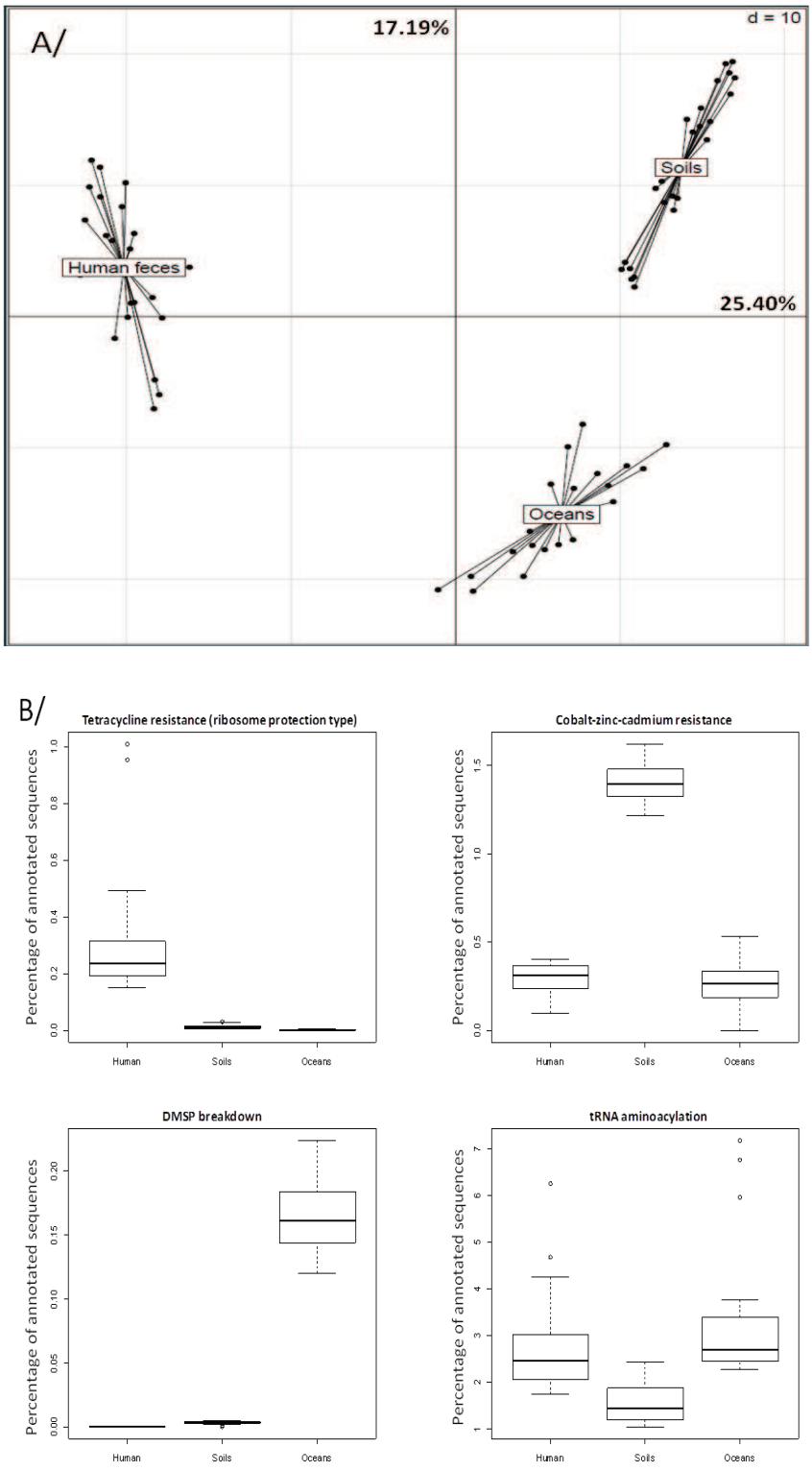


Figure 5: A/ Principal component analysis based on the distribution of 819 functions among 18 metagenomes corresponding to three distinct environments (soil, ocean and human feces) by integrating different E-value thresholds (from 10^{-5} to 10^{-40}). A total of 72 distributions are compared. B/ Relative distribution in percentage of four functional subsystems in three distinct environments (soil, ocean and human feces) represented by six metagenomes each and by integrating different E-value thresholds (from 10^{-5} to 10^{-40}). Box plots are based on the distribution fluctuations among the six metagenomes per environment annotated by applying four different stringencies.

Discussion:

Thanks to quick DNA sequencing revolution and free accessibility of user friendly annotation platforms, metagenomics is currently largely used by microbial ecologists without the requirement of any bioinformaticians. Thus, to extract DNA, generate metagenomes and analyze datasets is becoming a routine for several microbiologists and the results of these studies take a good place in major journals of the field.

Metagenomes are generally used to descript, compare and assemble metagenomes but are unfortunately not deprived of biases that influence results. The problematical aspect of these biases is that they are not integrated into metagenomic surveys and generally not accepted by microbial ecologists who trust blindly on the vast quantity of generated results. Thus the proportion of false positives (especially when comparing quantitatively datasets, what is generally done) is unknown and surprisingly largely unstudied while it could represent a major problem for microbial science. As a consequence, methodological experiences have to be performed to improve this recent domain of research.

Thus, this study aims to illustrate potential pitfalls in the interpretation of metagenomic data and to stimulate debates on metagenomic methodological alternatives. Based on the observations presented here, two highly distinct steps are strongly impacting the apparent structure of a metagenome (see figures 1 and 2). One step - the DNA extraction strategy - belongs to the responsibility of microbiologists; the other - the reads annotation process - is part of the bioinformaticians field.

These two steps represent the first and last part of a metagenomic study and appear to be crucial to descript and compare microbial communities. In addition, these methodological biases do not only impact the structure of a metagenome but are also different as a function of the microbial community studied (e.g., soil versus ocean versus human microbial communities, see figure 4) so complicating inter-environmental comparisons. However, the fluctuations observed when varying these two steps are not a limiting factor for global comparisons (see figure 5) and so do not represent an insoluble problem for the metagenomic field.

Two highly distinct strategies can emerge from these observations. One is the creation of a standardized approach to generate metagenomes from across the planet (e.g., utilization of a single and well standardized DNA extraction protocol with unchanged E-value cut-off). However, it could be scientifically difficult to define standards for specific methodological steps, especially when each possible approach provides a different vision of a black box (e.g., soil metagenome). In contrast, the second proposed strategy could be to integrate the maximum of methodological fluctuations into metagenomic surveys (Delmont et al., commentary in press) as an alternative to limit already defined biases. The major problems of this strategy are time, cost and sensitivity consuming, but could represent a relatively limited price to be more comfortable and confident with the generated datasets.

As a consequence, integrating or not DNA extraction and reads annotation fluctuations into metagenomic surveys should be at the minimum debated. To provide a first example of this alternative methodological strategy, the general functional distribution of the Rothamsted Park Grass soil metagenome was defined based on both DNA extraction biases and annotation fluctuations (figure 5). The standard deviation is considerable for some subsystems and relatively low for others. While limiting metagenomic sensitivity (e.g., >7% of fluctuation for the carbohydrate subsystem in the studied soil metagenome), this approach could limit false positives when comparing this soil microbial community to others. However, this approach increases probably the number of false negative results and so has to be discussed and if exploited, improved.

Conclusion:

While metagenomic approaches provide unique information about environmental functional potentials, it sensitivity is limited and standard deviations should be defined for each experiment. To perform replicates during metagenomic surveys could provide these standard deviations. However, to integrate both DNA extraction biases and reads annotation stringency (reflected by the E-value threshold) fluctuations to define the functional and taxonomical distribution of metagenomes before analyzing and comparing them is also a relatively easily feasible alternative. Metagenomic pitfalls presented here aims to create a debate; the discussion about how to integrate methodological fluctuations into metagenomic surveys is still open.

References:

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM. (2011) Accessing the soil metagenome for studies of microbial diversity. Appl Environ Microbiol 77:1315-24.

Delmont TO, Prestat E, Faubladier M, Bertels D, Robe P, Clark IM et al. Structure, Fluctuation and Magnitude of a Natural Prairie Soil Metagenome . Submited on the ISME journal.

Delmont TO, Malandain C, Prestat M, Larose C, Monier JM, Simonet P and Vogel TM. (in press) Metagenomic Mining for Microbiologists. ISME j.

Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth Get al. (2011) Metagenomic Discovery of Biomass-Degrading Genes and Genomes from Cow Rumen. Science 331:463-467.

Feinstein LM, Sul WJ, Blackwood CB (2009) Assessment of bias associated with incomplete extraction of microbial DNA from soil. 75:5428-33. Appl Environ Microbiol 75:5428-33.

Altermann W, Kazmierczak J (2003) Archean microfossils: a reappraisal of early life on Earth. Res Microbiol. 154:611-7.

Falkowski PG. 2001. Biogeochemical cycles. Encyclopedia Biodivers. 1:437-453.

Turnbaugh PJ et al. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 444-1027-1031.

Amann RI, Ludwig W, Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol Rev. 59:143-69.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular Biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5 :R245-R249.

Gillespie DE, Brady SF, Bettermann AD, Cianciotto NP, Liles MR, Rondon MR, Clardy J, Goodman RM, Handelsman J (2002) Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. Appl Environ Microbiol 68:4301-6.

Demaneche S, Sanguin H, Poté J, Navarro E, Bernillon D, Mavingui P, et al. (2008) Antibiotic-resistant soil bacteria in transgenic plant fields. Proc Natl Acad Sci USA 105: 3957-3962.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA et al (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66-74.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37-43.

Tringe SG, Mering CV, Kobayashi A, Salamov AA, Chen K, Chang HW et al. (2005) Comparative Metagenomics of Microbial Communities. Science, 308:554-557.

Dinsdale, EA, Edwards RA, Hall D et al. 2008. Functional metagenomic comparison profiling of nine biomes. Nature. 452:629-632.

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh Cet al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59-65.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.

Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. 26:715-21.

Head I. M., J. R. Saunders and R. W. Pickup. 1998. Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. Microb. Ecol. 35: 1-21.

Courtois S., A. Frostegård, P. Goransson, G. Depret, P. Jeannin, P. Simonet. 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. Environ. Microbiol. 3:431-439.

Martin-Laurent F., L. Philippot, S. Hallet, R. Chaussod, J. C. Germon, G. Saulas, and G. Catroux. 2001. DNA extraction from soils: Old bias for new microbial diversity analysis methods. Appl. Environ. Microbiol. 67: 2354-2359.

LaMontagne M. G., F. C. Michel, P. A. Holden, C. A. Reddy. 2002. Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis. J. Microbiol. Methods 49: 255-264

Carrig C., O. Rice, S. Kavanagh, G. Collins, V. O'Flaherty. 2007. DNA extraction method affects microbial community profiles from soils and sediment. Appl. Microbiol. Biotechnol. 77: 955-964.

Digitizing genetic structures prior to synthesizing new microorganisms: from knowledge to evolution v2.0

Tom O Delmont

Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France

Abstract: Microbial ecologists are now exploring genomes from across the planet, mining more than 3 billion years of evolution using efficient DNA sequencing technologies. They are actively building an exhaustive digital functional library by cumulating knowledge from all colonized environments. As sequencing costs are rapidly decreasing, the major obstacle is currently the accessibility of new microbial life styles. Novel sampling strategies and in lab experiments will help decipher the genomes of some of the 10^{30} Bacteria and Archaea estimated to live on Earth. However, in few decades, microbial discoveries will decrease, as the majority of functions, known and unknown, will already be present in the digital library. At this time, microbiologists will probably be actively mining this library to focus on the synthesis of new microorganisms for various human needs (*e.g.*, antibiotics, waste treatment). One major difficulty will be defining these needs and limit misapplications of synthetic biology.

Key words: metagenomic, synthetic biology, digital functional library.

Introduction:

Microorganisms emerged more than 3 billion years ago. They colonized various habitats (from deep oceans to arid deserts and Antarctic lakes) and the more complex forms of life (*e.g.*, humans). They are ubiquitous and govern life on this planet (1). To a certain point of view, animals are just entities providing a constant temperature and food supply to the billions of microorganisms living in their gut. In fact, while microorganisms can continue to exist (and probably did for a long period) without multi-cellular organisms, animals cannot stay alive a day without microorganisms. However, intelligence and science finally emerged from a constant evolution, and microbial ecologists can now study, mine and more recently, synthesize microorganisms. These new possibilities represent a shift, after billions of years of microbial self-determination, we are finally able to govern and manage microbial life in favor of humanity. Opportunities provided by the evolution v2.0 (synthetic biology) are considerable, the future of microbial science promising.

Digitizing microbial genetic structures

10^{30} Bacteria and Archaea are estimated to live on the third planet of our solar system (2). At this time, scientists do not know if the emergence and evolution of microbial forms of life is unique or a common process across the universe, but they have been able to cumulate considerable information about those present on Earth. Microorganisms were studied for centuries using direct observation tools and culture approaches (3). Even through a tiny part of the microbial diversity is currently cultivable (4), the study of their phenotype and genotype has improved our understanding of genetic functions and major mechanisms of life.

By using reference databases constructed from cultivable cells, important advances in sequencing technology capacities have allowed for the study of nucleic diversities (corresponding to different genetic structures) present in the environment that are called metagenomes (5-7). These technologies are now providing fast and cheap access to millions of relatively short sequences (30 to 800 base pairs) corresponding to the genetic code of environmental microbial communities. Several laboratories are currently focusing on the target of samples that correspond to not yet sequenced environments. Regardless of the high diversity of environmental metagenomes, only predominant microorganisms can be sequenced even when billions of reads are generated. However, accurate information can be extracted from the generated datasets (*e.g.*, (8, 9)). Surprisingly, the difficulty now lies in the acquisition of samples of interest (those representing an original evenness) rather than the deep description of their predominant communities. Two major strategies are emerging to access unusual microbial community evenness.

The direct one aims to target particular biological samples naturally present across the planet. As a major advantage, this strategy uses the billion years of microbial evolution that allowed the formation of microbial communities to become highly adapted to the physico-

chemical characteristics of particular places (*e.g.*, deep ocean, desert, arctic snow). In general, the lower the microbial diversity ~~is~~, the more unusual the generated dataset will be. As an example, acid mine drainage biofilms mainly colonized by a community of few species, provided access to a highly unusual metagenome and the reconstruction of predominant genomes from short reads (6, 10). In contrast, generated datasets appear to be globally similar across oceans or soils, two complex environments. In addition, the reconstruction of genetic structures from these complex environments is currently highly limited. However, various environments still have to be sequenced, and microbial ecologists are now attempting to manage this sequencing effort in the frame of the Earth microbiome project (11).

The indirect strategy is the artificial evenness design from complex microbial communities incubated under various controlled conditions. This method aims to divide a complex community (*e.g.*, from soil) prior the sequencing of extracted DNA to access particular sub-divisions of their entire metagenome and stimulate genomic reconstruction from lowly represented microorganisms {Delmont et al., under submission}.

These two strategies aim to study the genetic structure of the 10^{30} Bacteria and Archaea estimated to live on Earth as in depth as possible, whether they are predominant (first strategy) or lowly represented (second strategy), and will provide considerable information about functions present across the planet. Using in lab experiments, metagenomic approaches, additional culture efforts, improved single cell technologies (*e.g.*, screening step) and a considerable sequencing effort, an exhaustive digital functional library is going to be constructed and will be accessible to the international community for their mining (*e.g.*, for inter-environmental metagenomic comparisons (12)). However, metagenomic discoveries will quickly decrease as most of the functions, known or unknown, will already be present in the digital library. The interest in these sequencing strategies will decrease as a consequence.

Synthesizing microbial genetic structures

After sequencing, assembling and digitizing environmental nucleic diversity and due to considerable knowledge build from the genetic structures of microorganisms, microbiologists will logically focus on synthetic microbial life (13, 14). One simple definition of synthetic biology could be the design or re-design of self-replicating genetic structures. As a fundamental aspect, this domain of research proposes considerable opportunities for understanding how life emerged. On the other hand, this domain promises to optimize biological processes for the benefit of humanity. However, only few experiments are currently published and important challenges have still to be reached (*e.g.*, (15)). Synthetic biology is still in its infancy.

At this time, the most significant advancement published in the field is probably the reconstruction and self-replication of the small genome ($1.08.10^6$bp) of *Mycoplasma mycoides* using four assembly steps (16). As an important novelty, this experimentation was done entirely with chemically synthesized DNA cassettes of 1080pb including 80bp of overlap for assembly. 40 million dollars was necessary to synthesize the same small genome again, transplant it on yeast, and finally observe its self-replication capacities. While the progression appears to be expansive, difficult and relatively slow, perspectives resulting from this work are considerable.

While cassettes of 1kbp were designed in this work to reconstruct a pre-existing genome, it could be possible to imagine designing size variable modules that represent particular functions or groups of functions. Some of these modules will possess domestic genes and will be necessary for the design of any genetic structure. Others will correspond to additional functions, not crucial for microbial life in a given environment, but of interest for performing or stimulating particular processes, like energy, food, drugs, oil or oxygen production. These modules will be constructed based on the digital functional library generated from the 10^{30} Bacteria and Archaea cells implemented by man-made functions. Some companies are already providing some of these modules (*e.g.*, http://biobricks.org/).

For success in genomic design based on module~~s~~ selection, each has to possess high integrative capacities (*e.g.*, universal overlaps between modules) and to be independent from each other as much as possible. All modules selected to design a new genetic structure have to cohabit without negative interactions that prevent transcription or protein coding processes, for example. To by-pass this difficulty, modules of particular genes could be constructed for the purpose of becoming orthogonal (i.e. independent, (17)) with regard to others. Thus the digital library could be mined in this goal of providing thousands of modules, each possessing an independent and particular function, but capable of colliding each others. Parallel researches on microbial life mechanisms will help their cohabitation in self-replicating genetic structures.

An evolution v2.0 exploited for the benefit of humanity: the true challenge

As more and more modules will be available for synthetic life experiments, the number of possibilities will drastically increase. As an alternative to help microbiologists, artificial intelligences could be created to automatically generate microorganisms that respond efficiently to human needs. But what are these needs? How do we give priority to creating an efficient and cheap food supply to eradicate famine, rather than a perfume producing skin microbiome, when research funding is controlled by the privileged individuals in power? Unfortunately, the second initiative will yield greater profit and financial support, leading research opportunities away from larger-scale, basic human needs

Thus, instead of synthesizing microorganisms to produce drugs against major diseases in the third world, experiments will probably focus on the creation of a synthetic stomach microbiota that quickly brake down alcohols and fatty elements to fight against obesity. This is only one example of a possible misapplication of synthetic biology, but in case of success, targeted populations will quickly adapt their lifestyles to this apparent progress and just increase global consumption of alcohol and fatty elements. And the vicious circle will be initiated.

In some cases advances might potentially benefit all of us. For example, it could be possible in the future to synthesize new vaginal microbiota with particular capacities to limit most of the sexually transmitted diseases. With some optimism, one inoculation every month could be sufficient to eradicate the HIV/AIDS pandemic but samples have to be freely available in the third world, an idea that is currently utopian

In fact, the functional module market will probably become considerable due to it essential role to chemically build microorganisms of particular interest. Thus, some crucial modules and novel man-made functions will be produced on a small-scale to increase their cost independently of their potential to save lives. In the worst scenario, companies leading synthetic biology will be more powerful than countries and will decide major directions of evolution v2.0 without any latitude given to scientists.

Fortunately, humanity is not perfectly protected from a social revolution (*e.g.*, the Venus project, http://www.thevenusproject.com/). In case of profound improvements of the global system, microbiologists will have more of liberties to modify in depth the lives of billions of peoples across the planet, by providing food, energy, and access to drugs for all. The future will decide.

Acknowledgments:

I would like to express gratitude to Rachel L. Boate and Timothy M. Vogel who edited the text and stimulated in a positive way the comprehension of this perspective.

References:

1. Kowalchuk GA, Jones SE, & Blackall LL (2008) Microbes orchestrate life on Earth. *ISME J* 2(8):795-796.
2. Whitman WB, Coleman DC, & Wiebe WJ (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* 95(12):6578-6583.
3. Porter JR (1976) Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. *Bacteriological reviews* 40(2):260-269.
4. Amann RI, Ludwig W, & Schleifer KH (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews* 59(1):143-169.
5. Venter JC*, et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66-74.
6. Tyson GW*, et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428(6978):37-43.
7. Tringe SG*, et al.* (2005) Comparative metagenomics of microbial communities. *Science* 308(5721):554-557.
8. Qin J*, et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59-65.
9. Hess M*, et al.* (2011) Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 331(6016):463-467.
10. Bertin PN*, et al.* (2011) Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics. *ISME J* 5(11):1735-1747.
11. Gilbert JA*, et al.* (2010) The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. *Stand Genomic Sci* 3(3):249-253.
12. Delmont TO*, et al.* (Metagenomic mining for microbiologists. *ISME J*.
13. Endy D (2005) Foundations for engineering biology. *Nature* 438(7067):449-453.
14. Andrianantoandro E, Basu S, Karig DK, & Weiss R (2006) Synthetic biology: new engineering rules for an emerging discipline. *Mol Syst Biol* 2:2006 0028.
15. Porcar M*, et al.* (2011) The ten grand challenges of synthetic life. *Systems and synthetic biology* 5(1-2):1-9.
16. Gibson DG*, et al.* (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329(5987):52-56.
17. Wang B, Kitney RI, Joly N, & Buck M (2011) Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nature communications* 2:508.

Synthetizing microbial life to optimize Martian terraformation labors: from the terrestrial evolution to in lab extraterrestrial adaptation experiments

Tom O Delmont

Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France

Abstract: New experimental approaches providing perspectives about the synthesis of microorganisms to optimize Martian colonization labors are proposed. Genetic structures of microorganisms are now intensively digitalized using metagenomic approaches. The life style of microbial communities living in extreme environments (e.g. unusual temperature, radiation, pH, pressure) can already be confronted each other using these datasets. Global comparisons provide unique information about the strategies and originalities of highly adapted microbial communities, whatever they are recalcitrant or not to cellular cultures. Functions providing adaptation capacities to extreme conditions from different places of our planet could then be selected at the operon, gene or active site level, synthesized and grouped to design unique self replicating genomes. By applying next sequencing technologies to extreme environments, global metagenomic comparisons and robotized synthetic bacterial cell technologies, this genes selection at the planet level aims to create metagenomically engineered microorganisms. Resistance efficiency of the selected synthesized forms of life could then be optimized using a managed evolution. While probably maladjusted for our planet, these microorganisms could potentially be able to resist Martian surface conditions, or at least the sub surface. In a second time, additional modified microorganisms could be disseminated to alter Martian global climate for planetary ecosynthesis efforts (e.g. photosynthesis processes to inverse the ratio CO_2/O_2). This perspective aims to stress the importance of recent scientific advances on the frame of novel Martian terraformation possibilities.

"Earth is the cradle of humanity, but one cannot remain in the cradle forever."

Tsiolkovsky, Konstantin (1857-1935).

"I always thought the most significant thing that we ever found on the whole...Moon was that little bacteria who came back and lived and nobody ever said anything about it."

Apollo 12 Commander Pete Conrad, in 1991.

Life emerged more than 3.45 billion years ago in our planet [1], and then evolved to adapt and colonize almost all it surface, from hyper-arid deserts (e.g., [2]) to deep oceans (e.g., [3]) and Arctic snows (e.g., [4]). However, life is actually undetected outside in the universe and scientists currently don't know physico-chemical parameters necessary for their emergence. Thus, even if some extraterrestrial environments possess conditions relatively similar to extreme places on Earth, they are possibly inadequate for the emergence and evolution of life due to a lack of favorable habitat. As a consequence, some evolved forms of life from Earth could potentially be able to survive in specific extraterrestrial environments.

In particular, before any confirmation about life on Mars and even if some debates are present (e.g., [5]) due to both atmospheric water detection [6] and methane production [7], this planet is generally considered as a sterile environment. Extreme conditions at it surface explain easily this apparently absence of life (e.g., [8-10]). However, Mars is the most similar extraterrestrial environment of our planet in the solar system, and the fact that various extreme environments are colonized by extremophile microorganisms on Earth provides consistent hopes about future Martian colonization successes.

Microbial communities living in extreme environments possess particular characteristics to resist to hard conditions. These adaptations are possible due to the presence of genes coding particular proteins (e.g., [11]), or an unusual distribution of genes (e.g., gene duplication) providing an advantage for microorganisms (e.g., [12]). Due to the considerable adaptations of life on Earth, microbiologists are studying extremophile species capacities to resist several Martian conditions (e.g., [13-16]) and some experiments conclude about a possible development of these organisms over the red planet (e.g., [17-19]) but in general only for short periods (e.g., [20]). Some scientists even consider that specific microorganisms could be selected as prime candidates for Martian terraforming efforts (e.g., the cyanobacterium *Chroococcidiopsis*; [21]).

In spite of the evident interest of these approaches, to chemically create in controlled conditions evolved microorganisms could be a more suitable strategy to succeed in Martian or other extraterrestrial environment colonization. In fact, an important actual scientific frontier is that most microorganisms, including extremophiles, are recalcitrant to any cultivation approaches [22], and so cannot be studied individually. This constraint probably limits considerably the discovery of functions present on Earth and crucial to resist Martian conditions. In addition, each studied species correspond to a specific environment, far from Martian conditions.

To by-pass the culture limits, microbial communities' genetic structures are now extracted directly from the environment and then deeply sequenced to generate a pool of nucleic sequences corresponding to different organisms and commonly named a metagenome (e.g., [23-25]). Metagenomes from an always growing number of environments are available to the international community (e.g. http://metagenomics.nmpdr.org/; http://camera.calit2.net/; http://img.jgi.doe.gov/cgi-bin/m/main.cgi), and can then be

compared to emphasize specific adaptations [26]. These datasets are being used to create an exhaustive digital functional library, whatever functions are known or unknown.

Interestingly, metagenomic approaches could also be used to compare microorganisms' functional distributions among the principal extreme environments of our planet (e.g., extreme temperature, pressure, radiation, low pH). This effort can potentially emphasize ecosystems functional peculiarities, and so should provide new hypotheses about how these microbial communities survive in relatively unusual conditions. In addition, fosmid libraries can be created to study the functions of parts of genomes corresponding to uncultured extremophile microorganisms to discover the role of particular genes. As an example, cold-adapted enzymes were discovered from a clone library constructed using an Antarctic soil metagenome [27].

Of course, a majority of extreme environments have now to be sequenced, but some of them are already highly studied (e.g. acid mine drainage biofilms, deep oceans, Antarctic aquatic environments) and information can already be extracted from these metagenomes. As an example, sequences related to threalose biosynthesis appear to be unusually represented in specific acid mine drainage biofilms {Delmont and Vogel, under submission} were the pH is near 0 [24]. This function is known to serve as a source of energy and carbon [28], but also to protect proteins and cellular membranes from various stresses like desiccation, dehydratation, heat, cold and oxidation [29]. Based on metagenomic datasets, threalose biosynthesis probably possesses an important role in resistance in this extreme environment and could considerably help microorganisms in Martian conditions.

Sequences related to genes coding cold shock proteins ((CspA family of proteins) are as expected more represented in Arctic snows {Delmont and Vogel, under submission} but emphasize the role of these proteins in polar environments survival [30]. A strong cold adaptation is necessary for Martian adaptation, and therefore genes related to this stress have to be highly studied and probably optimized using managed evolutions.

To select all genes potentially capable of helping microorganisms survive but also develop over the red planet is a considerable task and requires the creation of an international consortium regrouping microbiologists specialized in extremophile environments and metagenomic approaches, but also bioinformaticians and geneticists. In addition, future sequencing projects (like the Earth microbiome project which aims to sequence more than 200 000 DNA samples from across the planet and including extreme environments; [31]) will condition the discovery of genes of interest for a Martian colonization.

After digitizing and selecting the key genes particular microorganisms exploit to resist to various extreme environments (as represented in the figure), this unique functional pool can be chemically grouped in genomes to create several synthetic microorganisms potentially capable of resisting to extreme environments like Martian surface or sub-surface. In fact, in parallel to functional digitizing advances, recent biotechnologies provide unique possibilities

and perspectives about synthetic life [32] and can be applied to the creation of microorganisms adapted to Martian habitats. Different chemically engineered cells possessing both resistance genes and functions involved in terraformation processes (e.g. photosynthesis processes to modify the ratio CO2/O2) could then be dispersed in particular places of the Martian surface to begin the first colonization effort of the red planet.

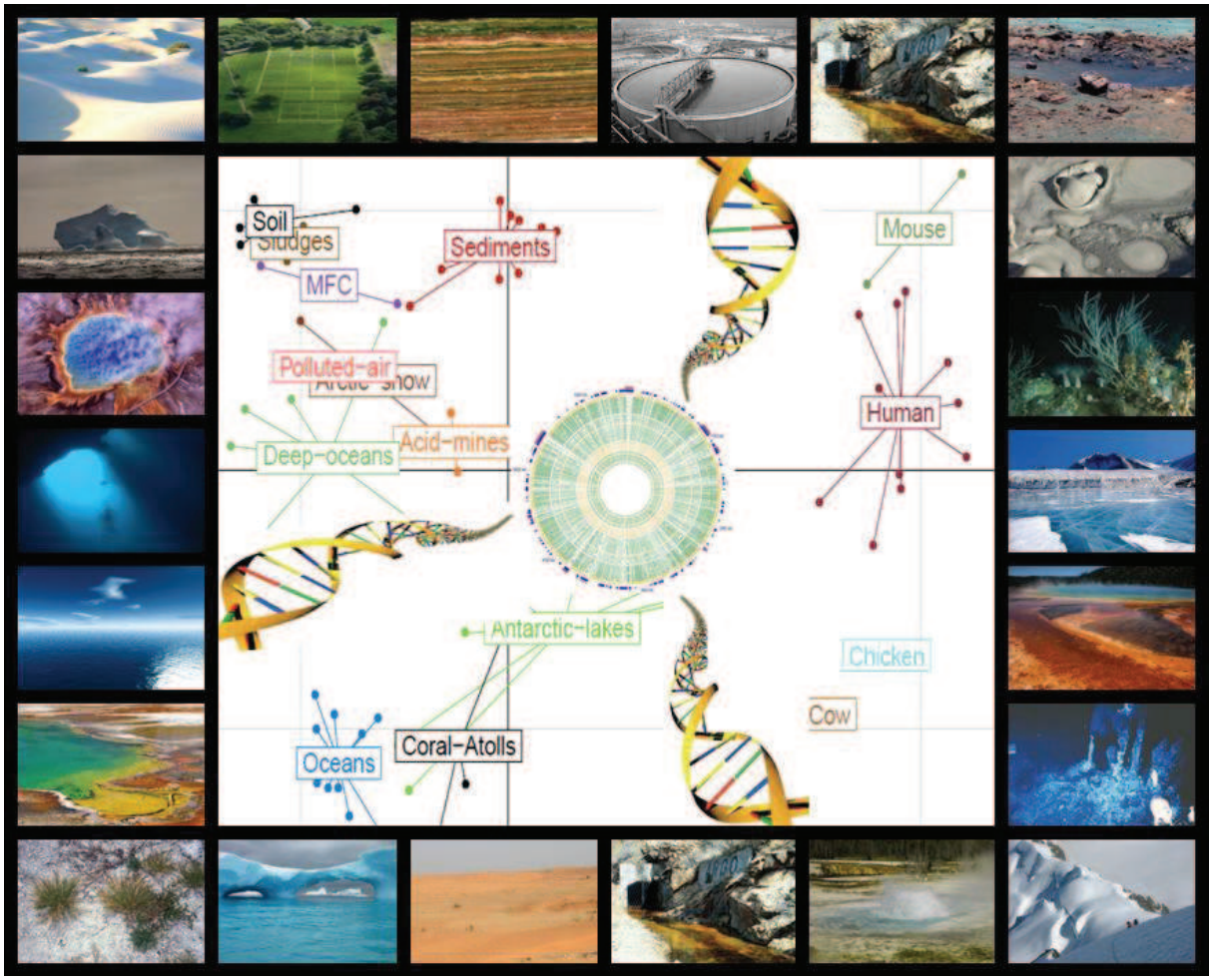


Figure: Schemas representing the global strategy of selecting and digitizing genetic structures of interest from across the planet using inter-environmental metagenomic comparisons and grouping them into synthetic genomes.

The principal difficulty will not be the selection of the genes, but more their cohabitation in genomes. Modules have to be constructed. Due to considerable construction possibilities (e.g., GC content, genes order, promoter efficiency variations), automated robots should be designed to build millions or billions of genetic structures, incorporate them in individual cells and test their resistances under various in laboratory controlled Martian conditions. Optimum engineered microorganisms could then be selected for the next stage. In fact, because some Martian conditions are more extreme than in any place on Earth surface, a

managed genetic evolution effort will probably be necessarily to increase the resistance efficiency of these selected synthetic microorganisms, by decreasing slowly the temperature generations after generations for example. Finally, these synthetic and evolved microorganisms should be cultured in quantity and disseminate at different depth of the red planet surface in collaboration with future Martian missions.

As additional perspectives, eukaryotic species could also be modified to help Terraforming efforts. In fact, lichens can potentially photosynthesize during weeks over Mars if liquid water is available [33], and to modify parts of it genomes could improve this capacity under extreme conditions. In addition, some kilograms of soil could also be enriched by in laboratory evolved microorganisms and deposed at the surface or the sub-surface of Mars. Soil is probably the most biodiverse environment on Earth with between 10^4 [34] and 10^7 species per gram [35]. Thus, when considering the vast microorganism's adaptation capacities (e.g., lateral gene transfers), some microorganisms could adapt themselves to Martian conditions, or provide to the synthesized microorganisms crucial genes that were not defined before.

As an ethic problem, this colonization effort, in case of success, could remove the proofs of the existence of an ancient or worse contaminate possibly actual forms of Martian life, so preventing extraordinary biological discoveries with the emergence of an astroarcheological science. But to warm this actually cold planet could possibly reactivate ancient microorganisms [36] living when it global climate was more similar to Earth [37, 38]. So a terraforming effort could also be the unique way to study ancient form of life in the red planet.

The aspiration of this perspective is only to provide alternatives about Martian Terraforming efforts. Problems associated to this possible future have to be discussed exhaustively, and not only between researchers, before any *in situ* experimental initiations. It is probable that the near future will emerge both extraterrestrial development perspectives and new philosophical concepts about the place of humanity on the known universe (are we really alone?), with a possible surfacing of tensions between scientists and the principal religions due to quick growing model contradictions. But this near future opens also incredible opportunities spreading life out of the planet.

In fact, if we are really alone on this universe, our responsibility to optimize life and intelligence survival across time is considerable, and microbial ecologists could play a crucial role in this effort by creating communities adapted to extraterrestrial environments. At this time, and because we have only access to extreme habitats in the solar system, the task is difficult, but possible. We just have to act in consequence, and provide solutions. This perspective proposes only one of these possible solutions, but aims mainly to emphasize the importance of recent scientific advances on the frame of Martian terraformation possibilities.

References:

1. Allwood, A.C., et al., *Stromatolite reef from the Early Archaean era of Australia.* Nature, 2006. **441**(7094): p. 714-8.
2. Pointing, S.B., et al., *Highly specialized microbial diversity in hyper-arid polar desert.* Proc Natl Acad Sci U S A, 2009. **106**(47): p. 19964-9.
3. Grzymski, J.J., et al., *Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility.* Proc Natl Acad Sci U S A, 2008. **105**(45): p. 17516-21.
4. Larose, C., et al., *Microbial sequences retrieved from environmental samples from seasonal arctic snow and meltwater from Svalbard, Norway.* Extremophiles, 2010. **14**(2): p. 205-12.
5. Kerr, R.A., *Planetary science. Liquid water found on Mars, but it's still a hard road for life.* Science, 2010. **330**(6004): p. 571.
6. Barker, E.S., et al., *Mars: Detection of Atmospheric Water Vapor during the Southern Hemisphere Spring and Summer Season.* Science, 1970. **170**(3964): p. 1308-10.
7. Formisano, V., et al., *Detection of methane in the atmosphere of Mars.* Science, 2004. **306**(5702): p. 1758-61.
8. Ulrich, R., et al., *Dynamic temperature fields under Mars landing sites and implications for supporting microbial life.* Astrobiology, 2010. **10**(6): p. 643-50.
9. Diaz, B. and D. Schulze-Makuch, *Microbial survival rates of Escherichia coli and Deinococcus radiodurans under low temperature, low pressure, and UV-Irradiation conditions, and their relevance to possible Martian life.* Astrobiology, 2006. **6**(2): p. 332-47.
10. Patel, M.R., et al., *Annual solar UV exposure and biological effective dose rates on the Martian surface.* Adv Space Res, 2004. **33**(8): p. 1247-52.
11. Weinberg, M.V., et al., *Cold shock of a hyperthermophilic archaeon: Pyrococcus furiosus exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoproteins.* J Bacteriol, 2005. **187**(1): p. 336-48.
12. Bratlie, M.S., et al., *Gene duplications in prokaryotes can be associated with environmental adaptation.* BMC Genomics, 2010. **11**: p. 588.
13. Dartnell, L.R., et al., *Low-temperature ionizing radiation resistance of Deinococcus radiodurans and Antarctic Dry Valley bacteria.* Astrobiology, 2010. **10**(7): p. 717-32.
14. Nicholson, W.L., et al., *Exploring the low-pressure growth limit: evolution of Bacillus subtilis in the laboratory to enhanced growth at 5 kilopascals.* Appl Environ Microbiol, 2010. **76**(22): p. 7559-65.
15. Wierzchos, J., et al., *Microbial colonization of Ca-sulfate crusts in the hyperarid core of the Atacama Desert: implications for the search for life on Mars.* Geobiology, 2011. **9**(1): p. 44-60.
16. Fajardo-Cavazos, P., A.C. Schuerger, and W.L. Nicholson, *Exposure of DNA and Bacillus subtilis spores to simulated martian environments: use of quantitative PCR (qPCR) to measure inactivation rates of DNA to function as a template molecule.* Astrobiology, 2010. **10**(4): p. 403-11.
17. Fendrihan, S., et al., *Investigating the effects of simulated martian ultraviolet radiation on Halococcus dombrowskii and other extremely halophilic archaebacteria.* Astrobiology, 2009. **9**(1): p. 104-12.
18. Kral, T.A., C.R. Bekkum, and C.P. McKay, *Growth of methanogens on a Mars soil simulant.* Orig Life Evol Biosph, 2004. **34**(6): p. 615-26.
19. Ronto, G., et al., *Solar UV irradiation conditions on the surface of Mars.* Photochem Photobiol, 2003. **77**(1): p. 34-40.
20. Kendrick, M.G. and T.A. Kral, *Survival of methanogens during desiccation: implications for life on Mars.* Astrobiology, 2006. **6**(4): p. 546-51.

21. Friedmann, E.I. and R. Ocampo-Friedmann, *A primitive cyanobacterium as pioneer microorganism for terraforming Mars.* Adv Space Res, 1995. **15**(3): p. 243-6.
22. Schloss, P.D. and J. Handelsman, *Biotechnological prospects from metagenomics.* Curr Opin Biotechnol, 2003. **14**(3): p. 303-10.
23. Venter, J.C., et al., *Environmental genome shotgun sequencing of the Sargasso Sea.* Science, 2004. **304**(5667): p. 66-74.
24. Tyson, G.W., et al., *Community structure and metabolism through reconstruction of microbial genomes from the environment.* Nature, 2004. **428**(6978): p. 37-43.
25. Tringe, S.G., et al., *Comparative metagenomics of microbial communities.* Science, 2005. **308**(5721): p. 554-7.
26. Delmont, T.O., et al., *Metagenomic mining for microbiologists.* ISME J.
27. Berlemont, R., et al., *Exploring the Antarctic soil metagenome as a source of novel cold-adapted enzymes and genetic mobile elements.* Rev Argent Microbiol, 2011. **43**(2): p. 94-103.
28. Elbein, A.D., *The metabolism of alpha,alpha-trehalose.* Adv Carbohydr Chem Biochem, 1974. **30**: p. 227-56.
29. Elbein, A.D., et al., *New insights on trehalose: a multifunctional molecule.* Glycobiology, 2003. **13**(4): p. 17R-27R.
30. Jung, Y.H., et al., *Overexpression of cold shock protein A of Psychromonas arctica KOPRI 22215 confers cold-resistance.* Protein J, 2010. **29**(2): p. 136-42.
31. Gilbert, J.A., et al., *The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010.* Stand Genomic Sci, 2010. **3**(3): p. 249-53.
32. Gibson, D.G., et al., *Creation of a bacterial cell controlled by a chemically synthesized genome.* Science, 2010. **329**(5987): p. 52-6.
33. de Vera, J.P., et al., *Survival potential and photosynthetic activity of lichens under Mars-like conditions: a laboratory study.* Astrobiology, 2010. **10**(2): p. 215-27.
34. Torsvik, V., L. Ovreas, and T.F. Thingstad, *Prokaryotic diversity--magnitude, dynamics, and controlling factors.* Science, 2002. **296**(5570): p. 1064-6.
35. Gans, J., M. Wolinsky, and J. Dunbar, *Computational improvements reveal great bacterial diversity and high metal toxicity in soil.* Science, 2005. **309**(5739): p. 1387-90.
36. Landis, G.A., *Martian water: are there extant halobacteria on Mars?* Astrobiology, 2001. **1**(2): p. 161-4.
37. Clifford, S., et al., *Introduction to special section: early Mars.* J Geophys Res, 1998. **103**(E13): p. 31405.
38. Solomon, S.C., et al., *New perspectives on ancient Mars.* Science, 2005. **307**(5713): p. 1214-20.

IV. Conclusion:

1. **A considerable opportunity for microbial ecologists: Metagenomic mining for microbiologists**
2. **A responsibility for metagenomic leaders: Decrypting microbial communities and performing global comparisons in the 'omic era: replicates vs flexicates**
3. **Debriefing**
    a. **Scientific conclusions**
    b. **Personnal conclusions**
    c. **Aknowledgments**

Metagenomic approaches represent a considerable opportunity for microbiologists to mine generated datasets for specific scientific questions. The section 1 emphasizes this aspect and to democratize this field of research. However, this domain of research is also full of pitfalls (see chapter 4, section 1) and it is difficult to provide data that can be used widely. Section 2 proposes the integration of flexicates instead of replicates for future metagenomic surveys. The main objective of this approach is to be more confident with datasets to describe microbial communities and perform global comparisons. Finally, section 3 presents the main scientific and personal conclusions of these three years of research, and acknowledgments.

The ISME Journal (2011), 1–7

www.nature.com/ismej

PERSPECTIVE

Metagenomic mining for microbiologists

Tom O Delmont[1], Cedric Malandain[2], Emmanuel Prestat[1], Catherine Larose[1], Jean-Michel Monier[1], Pascal Simonet[1] and Timothy M Vogel[1]
[1]*Environmental Microbial Genomics, Laboratoire Ampère, Ecole Centrale de Lyon, Université de Lyon, Ecully, France and* [2]*ENOVEO, 11 chemin de Boutary, Caluire et Cuire, France*

Microbial ecologists can now start digging into the accumulating mountains of metagenomic data to uncover the occurrence of functional genes and their correlations to microbial community members. Limitations and biases in DNA extraction and sequencing technologies impact sequence distributions, and therefore, have to be considered. However, when comparing metagenomes from widely differing environments, these fluctuations have a relatively minor role in microbial community discrimination. As a consequence, any functional gene or species distribution pattern can be compared among metagenomes originating from various environments and projects. In particular, global comparisons would help to define ecosystem specificities, such as involvement and response to climate change (for example, carbon and nitrogen cycle), human health risks (eg, presence of pathogen species, toxin genes and viruses) and biodegradation capacities. Although not all scientists have easy access to high-throughput sequencing technologies, they do have access to the sequences that have been deposited in databases, and therefore, can begin to intensively mine these metagenomic data to generate hypotheses that can be validated experimentally. Information about metabolic functions and microbial species compositions can already be compared among metagenomes from different ecosystems. These comparisons add to our understanding about microbial adaptation and the role of specific microbes in different ecosystems. Concurrent with the rapid growth of sequencing technologies, we have entered a new age of microbial ecology, which will enable researchers to experimentally confirm putative relationships between microbial functions and community structures.
The ISME Journal advance online publication, 19 May 2011; doi:10.1038/ismej.2011.61
Subject Category: integrated genomics and post-genomics approaches in microbial ecology
Keywords: metagenomics; community function; global ecosystems; community structure; environmental microbiology

Introduction

The explosion of metagenomic projects in an increasing variety of terrestrial and marine ecosystems (Tyson *et al.*, 2004; García Martín *et al.*, 2006; Kurokawa *et al.*, 2007; Nealson and Venter, 2007; Vogel *et al.*, 2009) and the availability of new high-throughput sequencing technologies are facilitating our understanding of the 'black box' of environmental microbial communities. This black box contains a wealth of novel genes that can aid in drug discovery and in a better understanding of processes for climate change, agronomy and pollution degradation. Importantly, this goldmine of biological information is becoming increasingly publically accessible through various databases and annotation platforms (http://metagenomics.anl.gov/; http://camera.calit2.net/; http://img.jgi.doe.gov/cgi-bin/m/main.cgi) and mining these data can aid in both providing answers to and helping to test and create new hypotheses for microbial ecologists. However, the demand for competent bioinformaticians and statistically valid data treatment methods often exceeds supply, thus leaving many microbial ecologists removed from this rush of metagenomic data. Many of the potential insights will come from comparing metagenomic data between ecosystems (Tringe *et al.*, 2005; Dinsdale *et al.*, 2008; Willner *et al.*, 2009). However, currently accessible data are underexploited despite their ecological relevance. Although this lack of data use and the perceived requirement for trained bioinformaticians could raise the question of the value of investing heavily in metagenomics projects (Baveye, 2009), we believe that benefit will come in the form of inter-ecosystem comparisons of microbial functions of interest, intra-ecosystem variations in microbial function, identification of novel genes and correlations between functions (and species) that will shed light on microbial interactions and adaptation.

Global metagenomic comparisons can be used to probe for answers to (or tickle the curiosity about) different aspects of microbial ecology by the

Correspondence: TM Vogel, Environmental Microbial Genomics, Laboratorie Ampère, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, Ecully 69134, France.
E-mail: timothy.vogel@ec-lyon.fr
Received 1 December 2010; revised 21 March 2011; accepted 21 March 2011

application of new user-friendly bioinformatics and statistical tools for understanding the strength of observed differences. As an example, we compared the diversity and distribution of 77 metagenomes (most being publically available) corresponding to various projects and environments by using the MG-RAST public platform (Meyer *et al.*, 2008). The relative numbers of sequence reads that were annotated in the metabolic subsystems provided were analyzed by principal component analysis (PCA). In addition, STAMP was used to evaluate the statistical significance of observed differences (Parks and Beiko, 2010). The metagenomes from these different ecosystems (oceans, coral atolls, deep oceans, Antarctic aquatic environments, Arctic snows, soils, hypersaline sediments, sludges, microbial fuel cell biofilms, acid mine biofilms, polluted air and animal microbial populations) are clearly separated (Figure 1a). Significant variations between research labs, sample types, DNA extraction and sequencing techniques for a given ecosystem do not seem to inhibit cross ecosystem comparisons. As an example, for the metagenomes from three ecosystems, ocean, soil and human microbiome, DNA was extracted by different researchers using different methods and different sequencing technologies (three sequential pyrosequencing technologies and the Sanger technology) were used, yet, these metagenomes are still grouped as a function of their original environment (Figure 1a). Thus, although methodological fluctuations exist between laboratories (Leek *et al.*, 2010), these results show a limited ('batch') effect of methodology (for example, sequencing technology) in high-throughput data when comparing globally different environments. In addition, studies focused on 16S rRNA gene sequencing also clustered animal microbial populations separately from the marine and terrestrial ecosystems (Ley *et al.*, 2008). The exploration of other gene classes responsible for these differences would provide insight into the overall functioning of these ecosystems.

A considerable limit of MG-RAST and other annotation platforms is the use of 'annotated' sequences, that is, those that have been classified as belonging (with a fixed probability of similarity, in our case here we used an *E*-value limit of 10^{-5}) to some established functional subsystems (for example, carbohydrate metabolism) or other databases, and the exclusion of non-annotated sequences, which might provide both novel functions and important differences between ecosystems. These non-annotated sequences provide a tremendous resource for future functional experiments and protein modeling. Some novel and potentially ecologically important functional groups are not being identified because of the dependence of current platforms on the already sequenced (and hopefully well annotated) genomes of mostly cultivated microorganisms. An added caveat to the discovery and exploitation of non-annotated

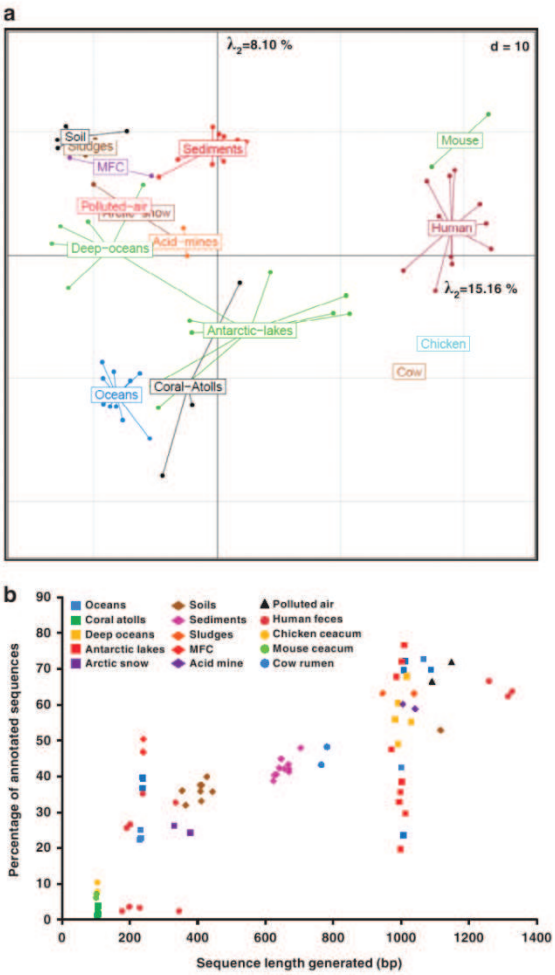


Figure 1 (**a**) PCA based on the relative distribution of annotated sequences (*E*-value $< 10^{-5}$) categorized in 838 different functional subsystems detected in the 77 metagenomes. Distributions were normalized as a function of the number of annotated sequences for each metagenome. The percentages of the illustrated two major axes correspond to the fraction of the total variance that they represent (see insert showing all of the axes and their percentage of the overall variance). (**b**) Relationship between average sequence length and the percentage of annotated functions (*E*-value $< 10^{-5}$) for the metagenomes used here. The different average sequence sizes are due in part to variations in sequencing technology. In addition, ocean and Antarctic metagenomes have annotations varying considerably for the same average sequence length. This fluctuation is due in part to the presence of sequences related to eukaryotic and virus sequences for oceans and Antarctic aquatic environments.

sequences (including possible gene assembly) is the dependence of sequence length on the percentage of annotated sequences (Figure 1b). In panel 1b, a clear correlation between the percentage of annotated sequences and sequence length is shown; however, there seems to be a confounding effect, which is the proportion of Eukarya- or virus-related sequences in

the metagenomic data set. For example, the percentage of annotated sequences in some ocean metagenomes (sequence length of about 1000 nt) is negatively correlated to the proportion of eukaryotic sequences ($R^2 > 0.86$, when using 59 metagenomes from the global ocean survey). Due to important annotation fluctuations independent of microbial community structure (Figure 1b), both functional and taxonomical distributions were normalized as a function of the number of annotated sequences as was carried out for Figure 1a and not the entire data set.

Although comparing ecosystems based on their relative sequence ('reads') frequencies in different metabolic subsystems can provide insight into functional differences, specific functions (or species) can be and should be individually examined in order to answer specific questions or to test hypotheses. Comparing large numbers of metagenomes can highlight unusual functional and phylogenetic distributions either between or within ecosystems. We provide a few examples of this approach to emphasize its significance (Figure 2). Oceans possess the highest relative number of metagenomic sequences related to dimethylsulfoniopropionate (DMSP) breakdown (Figure 2). DMSP occurs in considerable amounts in marine algae, for which this molecule and its breakdown products probably serve as an antioxidant system (Sunda *et al.*, 2002). But more importantly, its degradation can release dimethyl sulfide molecules (DMS) into the atmosphere, where they might improve cloud formation and limit solar radiation at the planet surface (Charlson *et al.*, 1987). This functional subsystem is distinctly more abundant in the ocean ecosystem than in any of the 14 other environments. In ocean-related metagenomes and within this subsystem ('DMSP'), sequences corresponding to DmdA (DMSP demethylase) and to DmdB2 (DMSP breakdown hydrolase) were found. On the other hand, inorganic sulfur assimilation-associated sequences are not particularly higher in the oceans than in other ecosystems and are on the same order as that for DMSP degradation in the ocean (about 0.1% of annotated sequences). Inorganic sulfur assimilation is more highly represented in the two activated sludge metagenomes, corresponding to

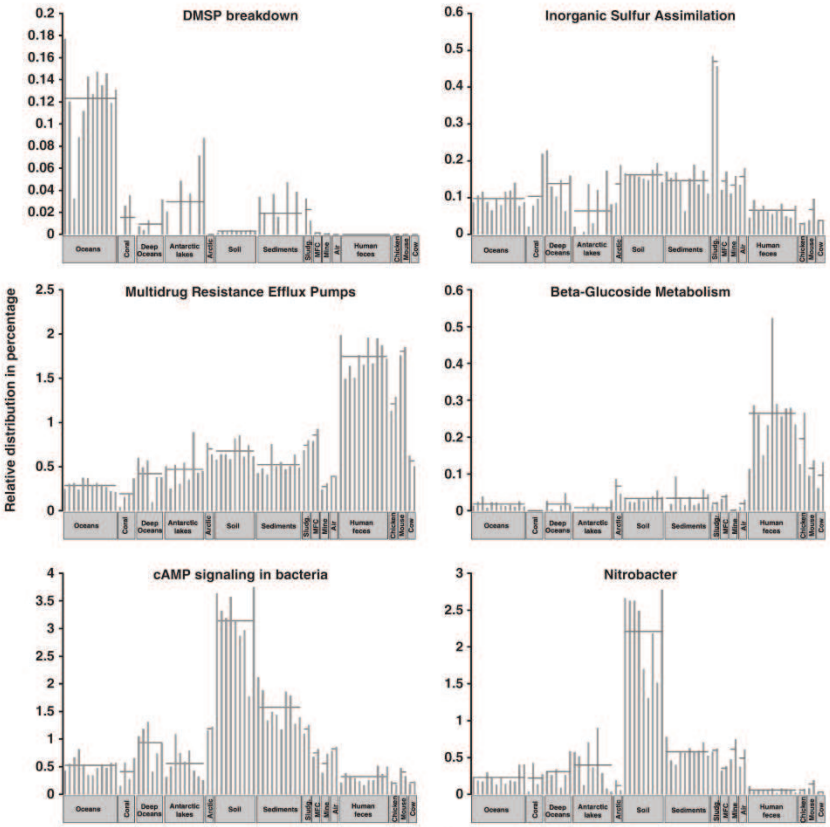


Figure 2 Comparison of the relative distribution in percentage (based on the annotated sequences (*E*-value $< 10^{-5}$)) of five functional classes and one genus (SEED annotation) among the 77 metagenomes deposited in MG-RAST. The horizontal line corresponds to the average of the relative distribution for each of the 15 environments.

~0.5% of the annotated sequences (Figure 2). The sulfur in sewage sludge can form gaseous SO_2 and cause associated acid-rain problems, if incorporated in sludge co-combustion processes. Therefore, biological mechanisms involved in sulfur cycling have immediate impacts on environmental processes.

Another example focuses on multidrug resistance efflux pumps involved in antibiotic resistance that have been extensively studied in pathogens (Li and Nikaido, 2004). However, these pumps are present in all living organisms and are not restricted to antibiotic compounds (Martinez *et al.*, 2009). They are also capable of extruding heavy metals, solvents and antiseptics (Pumbwe *et al.*, 2007). They are thought to be largely present in soil and in association with plants (Konstantinidis and Tiedje, 2004). However, we found that they appear to be more represented in human feces and chicken and mouse cecum (Figure 2), where they could have an important role in pathogen antibiotic resistance. These pumps are present in all the environments studied, confirming their multifunction role, but are relatively limited in oceans, deep oceans, polluted airs and acid mine drainage biofilms, where they are apparently less dominant. Although correlations can be calculated between functions that are relatively dominant in the same ecosystem (for example, beta-glucosides for animal-associated microbial communities), these correlations do not necessarily have any mechanistic value, but could be simply co-correlated to other phenomena. To provide more clear relationships, the presence of these two targeted subsystems on the same sequence (whether on a read or a contig) is required. The cAMP signaling is another example where we find more sequences related to bacterial cAMP signaling in soil than elsewhere (Figure 2).

The cAMP is an important secondary messenger in all three domains of life. Interestingly, as a cAMP subversion mechanism, some bacterial pathogens inject adenylate cyclase protein toxins into plants (Agarwal and Bishai, 2009). Thus, soil microbial communities appear to possess a considerable potential for deceiving plant signaling mechanisms, if cAMP is involved in pathogenicity; however, its role in other metabolic functions cannot be disregarded.

Metagenomic sequence data can also be used to evaluate the microbial community structure. In metagenomes without targeted gene amplification, the number of housekeeping genes present that can be used to assess which species are present is rather limited. On the other hand, all annotated sequences could be assigned to a closest related species and used to define the community structure with the caveat that only known species will be defined. An example is the apparent distribution of the genus *Nitrobacter* (known to oxidize nitrite to nitrate; Schmidt, 1978) in different ecosystems and its relative dominance in soil (Figure 2).

Another approach for comparing ecosystem metagenomes could be the combination of results from annotation platforms and the number of sequences determined by using BLAST for specific genes (Altschul *et al.*, 1990). This BLAST approach can be applied by tagging metagenomes and developing 'in house' annotation systems that researchers can create for specific questions. For example, a sample from the Gulf of Mexico is among the ocean metagenomes, therefore, these metagenomes could be screened for their relative petroleum hydrocarbon degrading abilities. Using BLAST, pooled and tagged metagenomes were screened for sequences similar to those associated with the degradation of oil compounds. For example, we looked at genes that code for the AlkB and AlkM enzymes, which are capable of degrading aliphatic oil compounds. Other genes including those associated with cytochrome P450 (CYP153 family), which has also been implicated in aliphatic hydrocarbon degradation, were included in our metagenomic screening. The relative distribution of these genes was used to predict variations in hydrocarbon degradation potential among ecosystems. These distributions were normalized as a function of the number of annotated sequences on annotation platforms as carried out for the MG-RAST subsystems. We simultaneously compared metagenomes from some of these ecosystems using both functional subsystems associated with hydrocarbon degradation and the specific hydrocarbon degradation genes cited above using a PCA approach. Antarctic aquatic environments, human feces and hypersaline sediments were similar in their general lack of sequences (other than the presence of those associated with anaerobic aromatic compound metabolism) associated with hydrocarbon degradation (Figure 3). The presence of anaerobic aromatic compound metabolism sequences was, however, observed in all ecosystems, but this type of metabolism is not limited to petroleum hydrocarbons. The location of the different functional genes can be projected on the same PCA plot in order to provide a visual clue as to which functional genes are associated with which ecosystems (Figure 3). For example, the majority of deeper ocean samples (from at least 500 m depth) have relatively more aliphatic oil-degrading genes than the surface ocean and soil, which contain more aromatic oil-degrading genes. The important fluctuations observed for deep oceans are due to the limited number of functions compared in this PCA. In addition, based on these selected hydrocarbon degradation functions, oceans and soil are relatively similar.

This rapidly growing metagenomic sequence data from different environments can also help researchers target microbial communities that might have roles in a range of important functions. Although gene presence *per se* is not indicative of enzymatic activity, and the ecosystems compared here are not equal in amount of sequence data, understanding the relative proportions of these

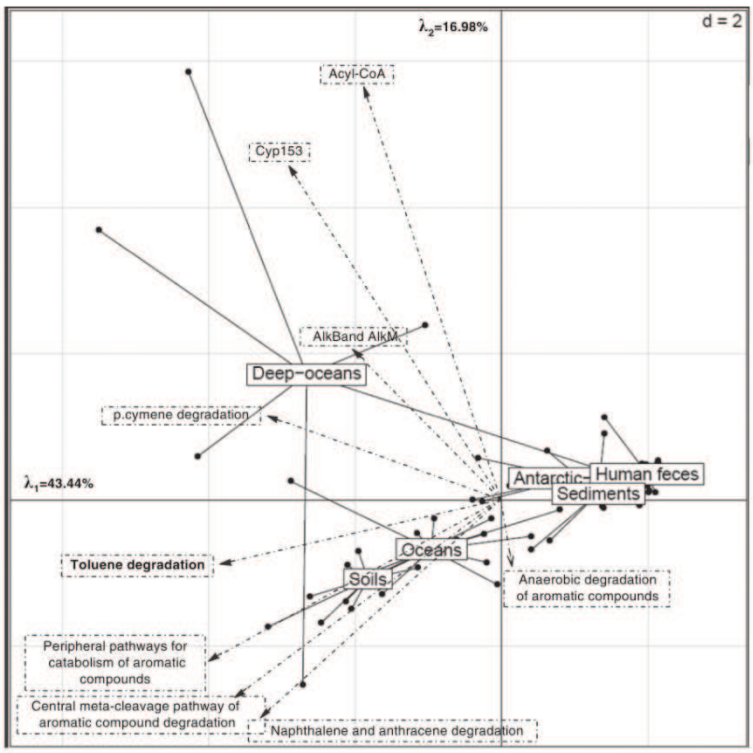


Figure 3 PCA of six selected ecosystems based on their number of sequences associated with petroleum hydrocarbon degradation functions (*E*-value $< 10^{-5}$). The functional classes as provided by MG-RAST and the local blasts are plotted on the same PCA as the samples in order to observe relationships between function and environment.

genes in specific ecosystems might provide better insight into their relative importance. Perhaps metatranscriptomic and metaproteomic approaches will help to understand the temporal nature of the specific activities and functions that are expressed. After such primary analysis, clone libraries could be constructed (and are being constructed in some cases) and probed for the sequences of interest, after which the genetic environment of these functional genes could be elucidated.

We have shown only a very limited analysis of existing metagenomic data here in order to illustrate existing resources available to microbial ecologists today. These resources are being constantly replenished by increasing data sets and sequenced ecosystems. The distribution of every defined function and species can already be evaluated at different taxonomical levels in hundreds of metagenomes using both annotation platforms and local BLAST for specific questions. Differences in functional gene families and specific functions (or target sequences) between metagenomes from different environments can aid our understanding of how microbial communities function. The beauty of this influx of metagenomic data is that so much remains to be discovered. As an example, estimations established that between 10^4 (Torsvik *et al.*, 2002) and 10^7 (Gans *et al.*, 2005) different species can be present in 1 g of soil. In addition, we have emphasized the considerable difficulties in accessing soil genetic richness that limit the number of detected species when using only one DNA extraction approach (Delmont *et al.*, 2011). As a consequence, currently available metagenomes related to soil and other ecosystems represent only a fraction of their existing genetic potential.

In the future, continuous advances in sequencing technologies (and sequenced genomes) will generate not only more, but also longer sequences, thus increasing significantly metagenomic sensitivity and possibly the percentage of annotated sequences (Figure 1b). The re-annotation of metagenomes when additional reference genomes become available will also stimulate and improve annotations, if those sequenced genomes are accurately annotated themselves. Increased number and length of metagenomic sequences will also lead to genome assembly and possible improved cultivation techniques. In addition to this revolution, the continuous increase in metagenome sequencing projects (for example, TARA, Earth Microbiome Project, Terragenome and Microbial Earth project; Gilbert *et al.* 2011)

and new global metagenomic comparison tools are aiding researchers enter a new age of microbial ecology. However, experiments (including metatranscriptomic and metaproteomic analyses) are becoming essential to confirm the biological roles of annotated functions (and microorganisms) *in situ* and to increase our knowledge concerning the vast quantity of non-annotated sequences.

Acknowledgements

TOD was supported by the Rhône-Alpes Region and part of this work was supported by the French National Research Agency (Agence National de Recherche) ANR Genomique programme: METASOIL project.

References

Agarwal N, Bishai WR. (2009). cAMP signaling in *Mycobacterium tuberculosis*. *Indian J Exp Biol* **47**: 393–400.
Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
Baveye PC. (2009). To sequence or not to sequence the whole-soil metagenome? *Nat Rev Microbiol* **756**: 757.
Charlson R, Lovelock J, Andreae M, Warren S. (1987). Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**: 655–661.
Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P *et al*. (2011). Accessing the soil metagenome for studies of microbial diversity. *Appl Environ Microbiol* **77**: 1315–1324.
Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM *et al*. (2008). Functional metagenomic comparison profiling of nine biomes. *Nature* **452**: 629–632.
Gans J, Wolinsky M, Dunbar J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**: 1387–1390.
García Martín H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC *et al*. (2006). Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol* **24**: 1263–1269.
Gilbert J, O'Dor R, Vogel TM. (2011). Survey data are still vital to science. *Nature* **469**: 162.
Konstantinidis KT, Tiedje JM. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *P Natl Acad Sci USA* **101**: 3160–3165.
Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A *et al*. (2007). Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* **14**: 169–181.
Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE *et al*. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**: 733–739.
Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nat Rev Microbiol* **6**: 776–788.
Li XZ, Nikaido H. (2004). Efflux-mediated drug resistance in bacteria. *Drugs* **64**: 159–204.
Martinez JL, Sánchez MB, Martínez-Solano L, Hernandez A, Garmendia L, Fajardo A *et al*. (2009). Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems. *FEMS Microbiol Rev* **33**: 430–449.
Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al*. (2008). The Metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.
Nealson KH, Venter JC. (2007). Metagenomics and the global ocean survey: what's in it for us, and why should we care. *ISME J* **1**: 185–187.
Parks DH, Beiko RG. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* **26**: 715–721.
Pumbwe L, Skilbeck CA, Wexler HM. (2007). Induction of multiple antibiotic resistance in *Bacteroides fragilis* by benzene and benzene-derived active compounds of commonly used analgesics, antiseptics and cleaning agents. *J Antimicrob Chemoth* **60**: 1288–1297.
Schmidt EL. (1978). Nitrifying microorganisms and their methodology. In: Schlessinger D (ed.). *Microbiology—1978*. American Society for Microbiology: Washington, DC, pp 288–291.
Sunda W, Kieber DJ, Kiene RP, Huntsman S. (2002). An antioxidant function for DMSP and DMS in marine algae. *Nature* **418**: 317–320.
Torsvik V, Ovreas L, Thingstad TF. (2002). Prokaryotic diversity-magnitude, dynamics, and controlling factors. *Science* **296**: 1064–1066.
Tringe SG, Mering CV, Kobayashi A, Salamov AA, Chen K, Chang HW *et al*. (2005). Comparative metagenomics of microbial communities. *Science* **308**: 554–557.
Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al*. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.
Vogel TM, Simonet P, Jansson JK, Hirsch PR, Tiedje JM, van Elsas JD *et al*. (2009). TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat Rev Microbiol* **7**: 2.
Willner D, Thurber RV, Rohwer F. (2009). Metagenomic signatures of 86 microbial and viral metagenomes. *Env Microbiology* **11**: 1752–1766.

Appendix

Oceans
4441573.3, 4441574.3, 4441576.3, 4441577.3, 4441591.3, 4443688.3, 4443697.3, 4443713.3, 4443714.3, 4443716.3, 4443725.3, 4443729.3.

Coral attols
4440279.3, 4440037.3, 4440039.3, 4440041.3.

Deep oceans
4441619.3, 4441656.4, 4441620.3, 4442503.3, 4441663.3, 4442500.4.

Antarctic aquatic environments
4443683.3, 4443680.3, 4443682.3, 4443684.3, 4443679.3, 4443686.3, 4443685.3, 4443687.3, 4443681.3.

Arctic snows
4443128.3, 4443127.3.

Soils
4441091.3, 4446153.3, http://metasoil.univ-lyon1.fr/ for metagenomes corresponding to Rothamsted Park Grass soil experiment.

Sediments
4440964.3, 4440963.3, 4440965.3, 4440966.3, 4440967.3, 4440969.3, 4440970.3, 4440968.3, 4440971.3, 4440972.3.

Phosphorus removing sludges
4441092.3, 4441093.3.
Microbial fuel cells
4447261.3, 4447259.3.

Acid Mine Drainage Biofilms
4441137.3, 4441138.3.

Singapore indoor polluted airs
4447940.3, 4447941.3.

Human feces
4440825.3, 4440460.5, 4440614.3, 4440611.3, 4440613.3, 4440616.3, 4440595.4, 4440452.7, 4440939.3, 4440942.3, 4440943.3.

Chicken Cecum
4440283.3, 4440284.3.

Mouse cecum
4440463.3, 4440464.3.

Cow rumen
4441679.3, 4441680.3.

These accession numbers correspond to metagenomes available on MG-RASTv2 server (http://metagenomics.anl.gov/v2/).

Decrypting microbial communities and performing global comparisons in the ‘omic era: replicates and flexicates.

Tom O. Delmont, Pascal Simonet, and Timothy M. Vogel

Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France

Abstract: Metagenomic is opening new doors in microbial ecology and makes it possible the study of the unseen microbial majority using direct sequencing approaches. Unfortunately, these new tools are not unbiased. One problem is that we cannot know which method (from the DNA extraction to the sequences annotation) provides the best picture of microbial communities in term of evenness. As a consequence, classical metagenomic approaches make quantitative analyses risky and most scientists assume the relative distributions measured reflect the true distributions. Thus, regarding the impact of methodological parameters on the structure of generated metagenomes, replicates are probably insufficient and methodological fluctuations should also be integrated. This effort would help define a larger variance of all taxa and functions detected in metagenomes, so creating a global picture of any system or environment. Applying multiple approaches to access environmental genetic diversities should stimulate the discovery of new species and functions, so helping understanding how cosmopolitan microbes are. This commentary aims to present both replicates (one method) and flexicates (different methods) when decrypting microbial communities and performing global comparisons in the ‘omic era.

Key words: metagenomics, replicates, flexicate, flexicates, earth microbiome project

Running title: Replicates vs flexicates

Subject Categories:

- Microbial population and community ecology
- Integrated genomics and post-genomics approaches in microbial ecology
- Microbial ecology and functional diversity of natural habitats
- Microbial ecosystem impacts

Environmental metagenomic approaches emerged with the cloning of DNA fragments extracted directly from the environment into cultivable microorganisms to stimulate the discovery of new enzymatic activities (Handelsman et al., 1998). This new field of research has evolved and is now largely focused on massive sequencing of DNA directly extracted from various environments. The number of available metagenomic datasets is quickly growing and represents sequences from a wide range of microbial communities (see available datasets on MG-RAST and IMG/M public annotation platforms for examples). Interestingly, while a majority of studies are done by analyzing datasets from a single ecosystem, considerable information can also be extracted by performing inter-environmental metagenomic comparisons (e.g., Tringe et al., 2005; Delmont et al., 2011, a). A critical question is the proportion of the total metagenome that is detected and the effect of incomplete metagenome datasets on qualitative and quantitative analyses.

Optimism concerning the proportion of the total metagenome sequenced probably varies inversely to the biodiversity of the environment or ecosystem targeted. Other factors are of course sample selection, preparation, and DNA extraction as well as the amount of sequencing done. Naturally, most scientists involved feel the more sequencing of a specific ecosystem, environment, biome or sample, the higher the proportion of the actual metagenome sequenced. One of the difficulties is this target described as the metagenome. Some think of the metagenome as the complete sequencing of all biological members of the targeted system, but even this could be one genome per species (however species definition would be debated) or all individuals present in the sample/system including therefore the genetic variations of the same species. Today, we are nowhere near either of these possibilities for many environments. So quantitative and qualitative comparisons assume that the sequenced (and in many cases annotated) fractions of the total metagenome provide accurate proportional information of both phylogenetic and metabolic characteristics. This may be accurate. How can one judge the value of missing data? Testing the missing data by deeper sequencing or drastically increase the number of sequences from different ecosystems might help if there is no systematic bias that would only be repeated. Another approach would be to try sequencing multiple samples with different biases. Missing from this discussion is any mathematical model or description of the rarefaction curves as a function of sampling, DNA extraction, and sequencing biases. Our work published in this issue (page XXX, Delmont et al., 2011, b) applied the multiple sequencing approach to the same soil. Others try to produce one deeply sequenced run per environment. Plans for sequencing a large number of ecosystems or biomes, (the Earth Microbiome Project) have also been developed (Gilbert et al., 2010; http://www.earthmicrobiome.org/). The principal challenge is to connect structure, function and environmental characteristics at both a local and a global level without losing the biological information during the transformation of *in situ* microbial ecology to *in silico* analyses. Current global metagenomic analyses suggest that biases have little effect on comparisons between significantly different ecosystems, such as oceans, soils, and human digestive tracts (Delmont et al., 2011, a), but this might be due to the large differences in the structure and function of the microbial communities. These differences might shrink when comparing two different soils to the extent that methodological biases become the overriding factor. One way to measure the acceptable degree of similarity at which methodological biases do not hinder global comparisons is to apply different methods to each of the biomes under study.

Another approach is to apply the same method everywhere and consider that its associated biases apply to all samples. This would provide data from more biomes than the approach sequencing multiple samples from the same biome. This approach would lead to comparing of thousands of datasets corresponding to various environments using one standardized protocol. Nevertheless, concerns that sampling and DNA extraction biases appear to vary as a function of environment, season, depth, temperature, organic content, etc would not be addressed. Today, considerable debate about which methods are the "best" suggests that the choice would probably be more political than scientific. How can anyone select one method when each method provides a different image (dataset) of the metagenomic structure? Even the one with the most DNA extracted cannot be defend as long as not all of the DNA is extracted. We know the extraction step has the potential to influence conclusions when using 'omics approach.

Even if sequencing is considered relatively accurate, taxonomical and functional differences between samples based on each "metagenomic" datasets should be considered as a mix of differences spread unevenly around the true distributions. We do not know what the true distributions are. Two data presentational approaches are used: qualitative and quantitative. For the qualitative approach, the recognition of the presence of a taxonomic unit or a metabolic function is used to describe the environmental microbiology of the site from where the sample was taken. The quantitative approach is to provide the relative proportion of the taxonomic unit or metabolic function relative to others in the same sample or to other samples or environments. Described this way, both the absence and the relative proportion seem conceptually fragile. The presence of a taxonomical unit or function seems clearer (ignoring for the moment the bioinformatics problem of phylogenetic or metabolic assignment). Thus, a number of different samples from the same site with their different biases could be used to construct a stochastic image of the true taxonomic and functional distributions.

Environments can possibly be sampled by varying time and space, and by varying DNA extraction and sequencing and annotation methodologies. Thus, the non-biological factors that influence our perception of the structure of a metagenome need to be varied, too. With the same objective of stimulating the relevance of metagenomic experiments, Jim Prosser emphasized in the form of an editorial in 2010, the frequent lack of replicates amongst a part of scientific articles submitted to illustrious journals and the importance of this effort to perform accurate science. Assuming that considerable biases can be hidden behind highly reproducible replicates (an example is presented in the figure 1), replicates might be insufficient, and the application of different approaches to access the genetic diversity present in the environment might be required when performing metagenomic surveys.

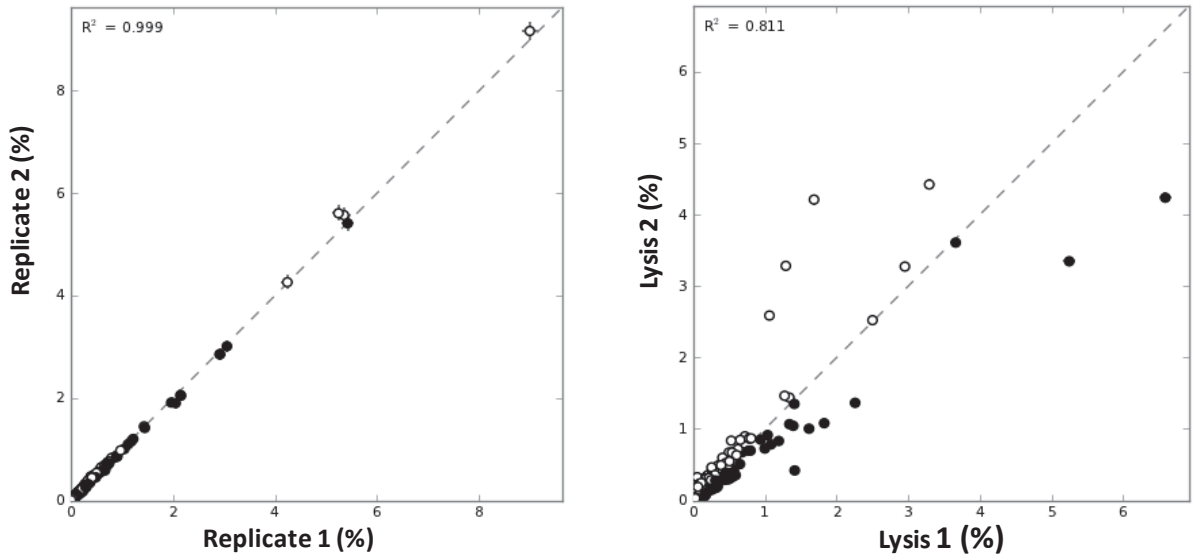


Figure 1: The two graphs represent the relative distribution in percentage of genera (using MG RAST, SEED annotation (E-value < 10^{-5}) ans STAMP) between two datasets. Replicates 1 and 2 correspond to DNA pools extracted from two distinct control microcosms after 4 months of incubation. Lysis 1 and 2 correspond to the same soil sample and two distinct DNA extraction protocols.

Of course, this strategy would be relevant only when varying methodology for steps that increase the access of genomic sequences rarely or poorly extracted by one method and when it is impossible to know which possible approach provides the truest image of the microbial community. These different methods used to study a single sample could be defined as “flexicates” and be applied using a wide range of strategies depending on environmental specificities. In addition, replicates can and should be performed for each tested method to integrate reproducibility variations into methodological fluctuations. After sampling, managing and sequencing DNA pools, datasets could then be compiled to represent a global picture of each studied environment. The functional and taxonomical variances so obtained should represent both natural and methodological fluctuations. The main goals of this approach are to help microbial ecologists decrypting communities, understanding how cosmopolitan microbes are and performing global metagenomic comparisons with more confidence in datasets (an example of already possible comparison between human feces and oceans by partially integrating methodological fluctuations is presented in the figure 2, panel A).

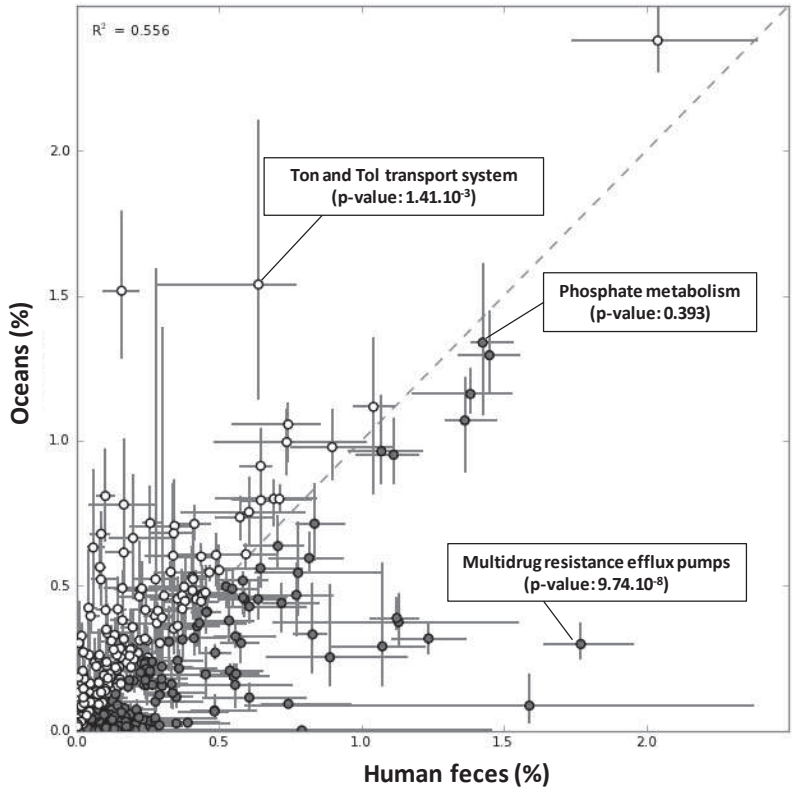


Figure 2: Relative distribution of microbial functions from the Ocean (12 datastes) and Human feces (11 datasets) (MG-RAST annotation, functional level 3, E-value < 10^{-5}) by partially integrating spatial, temporal and methodological fluctuations (metagenomes information is presented in Delmont et al., 2011, a) using the STAMP v2.0. Standard deviations are presented for each functional subsystem and for the two environments.

If this strategy is used, then a metagenomic definition of ecosystem boundaries at the microorganism level could be when inter-environmental distribution differences are globally stronger than intra-environmental fluctuations, whether they are natural or methodological. We can hope that technological advances might limit the methodological aspect of future metagenomic surveys, but until then better safe than sorry.

References

Delmont TO, Malandain C, Prestat M, Larose C, Monier JM, Simonet P and Vogel TM. (in press) Metagenomic Mining for Microbiologists. ISME j doi:10.1038/ismej.2011.61. a

Delmont TO, Prestat E, Faubladier M, Bertels D, Robe P, Clark IM et al. Structure, Fluctuation and Magnitude of a Natural GrassLand Soil Metagenome . Submited on the ISME journal. b

Gilbert JA, Meyer F, Jansson J, Gordon J, Pace N, Tiedje J et al. (2010) The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory. Stand Genomic Sci 3:249-53.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. (1998) Molecular Biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. 5 :R245-R249.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA. (2008) The Metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 9:386.

Parks DH and Beiko RG. (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26:715-21.

Prosser J. (2010) Replicate or lie. Environ Microbiol 12:1806-10.

Tringe SG, Mering CV, Kobayashi A, Salamov AA, Chen K, Chang HW et al. (2005) Comparative Metagenomics of Microbial Communities. Science, 308:554-557.

Debriefing

a. Scientific conclusions

b. Personnal conclusions

c. Aknowledgements

Scientific conclusions

Introduction:

During these three years working on microbial ecology, my goal was the study of the Rothamsted Park Grass soil microbial communities by microbial metagenomic approaches. My assets were a two million fosmid library and 90 pyrosequencing runs that we generated with DNA extracted from indigenous Rothamsted microorganisms. While the different chapters presented in this thesis are interconnected, this work represents different facets of soil metagenomic approaches that aim to by-pass the considerable difficulties in studying these communities. In addition, the introduction and bibliography are not focused on a single environment and present global metagenomic comparisons, a relative added value in metagenomics. This is a unique approach to appreciate soil microbial community peculiarities and stimulated my understanding of the particular life style of these microorganisms.

Study of the natural structure of the Rothamsted soil microbial communities:

The principal aim of my thesis was to study the Rothamsted soil microorganisms in their natural environment (no modified conditions) to emphasize their structure and functional potentials. If metagenomic approaches were ideal tools without any introduced biases, the effort to describe these communities would have been rapid and efficient by applying a unique method and the deep sequencing effort provided by the Metasoil project.

Unfortunately, biases are present in almost all steps of metagenomic studies, and we decided to integrate these limits to provide a global picture of these communities. Thus, a study was performed to present what we can do in term of strategy to access the nucleic diversity present in soil. Among other interests, this work emphasized important DNA extraction biases when studying DNA samples from soil and provided an original approach to stimulate considerably the detection of species in this environment. Since this study was applied in term of sampling and DNA extraction strategy for the Rothamsted soil metagenome sequencing efforts, it was a crucial step of my thesis and probably the foundation of all my doctoral work.

Thus, thirteen distinct Titanium pyrosequencing runs were done using eleven different methods by varying seasons, depths and DNA extraction protocols. This sequencing effort provided a standard deviation of both functional and taxonomical distributions of the natural Rothamsted soil metagenome by integrating natural and methodological fluctuations. Then, these thirteen metagenomes were grouped with other available soil metagenomes (from Italia, Puerto Rico and North America) and compared to metagenomes corresponding to other environments (oceans and human feces). This comparison emphasized soil microbial community peculiarities and helped understand their role in a global context. Finally, metagenomic assembly efforts were performed and the observations presented helped estimate the amplitude of the soil metagenome, and subsequently, the sequencing effort necessary to assemble genomes from the metagenome.

Study of modified structures of the Rothamsted soil microbial communities:

It became apparent that to sequence this soil metagenome by varying only space, time and methodology would not be sufficient to access the lowly represented species (sometimes called the

"rare biosphere" by the international community). Thus, the second step of my work was to change the natural distribution of these communities using various strategies prior the extraction and massive sequencing. Obviously, to limit the predominance of few genera (*e.g.*, Bradyrhizobium) and to stimulate the sequencing of other genera was a considerable challenge.

The principal strategies used were i) to transfer the Rothamsted soil into other sterilized soils (9 metagenomes in tag generated), ii) to fractionate a DNA sample (500 micrograms of purified DNA, what is considerable) as a function of the GC content (in collaboration with William Holben; 6 metagenomes generated, results only partially presented), iii) to apply various conditions to these communities in microcosms (24 metagenomes generated), and finally iiii) to study DNA extracted from soil samples dried and stored since 1876 in the Rothamsted experimental station (6 metagenomes in tag and 6 in compete plates generated, results only partially presented).

The second difficulty was the reduced yield of DNA extraction in some cases after modifying the structure of the microbial community. Unfortunately, it was a general rule that the lower the DNA extraction yield, the more efficient the method to access hidden microorganisms. In these cases, a considerable number of extractions were done from the same soil sample to satisfy the need for sufficient DNA to sequence (generally 10 micrograms of purified DNA). However, the strategies to access different metagenomes from the same location provided (with more or less of success) considerable information about the Rothamsted soil microbial community adaptation capacities and uncovered the presence of unexpected species and functions. In addition, assembly performances were improved in some specific samples (*e.g.*, with heavy metal and mercury enrichment microcosms) with the creation of large contigs (up to 400 kb). Of course, a considerable sequencing effort is still necessary to sequence this soil metagenome in is totality.

Perspectives:

In the article entitled "How to avoid pitfalls in the metagenomic jungle", we proposed a strategy that integrates both DNA extraction approaches and annotation fluctuations when studying a metagenome and especially when comparing datasets corresponding to a same or to different environments. In addition, using our experience from the Rothamsted soil sequencing effort and the global metagenomic comparisons, we have a proposal for the sampling and sequencing strategy of the emergent Earth microbiome project that aims to sequence and compare a considerable number of samples from across the world.

Personal conclusion:

When accepting to be part of the Metasoil/Terragenome project in July 2008, I had no idea about the human adventure this thesis would represent for me. It brought me to England (Rothamsted experimental Station), Sweden (Bageco conference, Upsalla), California (to visit LBNL and JGI institutes), Seattle (ISME conference), Chicago (Argonne metagenomic workshops 2010 and 2011), Germany (Soil metagenomic conference, Braunschweig), China (Earth microbiome project conference, Shenzhen) and finally Switzerland (FEMS conference, Geneva).

All this travel provided the occasion to meet microbial ecologists, bioinformaticians and geneticists. I feel part of something now. So thank you Penny Hirsch, Ian Clark, Janet Jansson, Marc Bailey, Folker Meyer, Jim Prosser, Stefan Green, Jack Gilbert, Kornelia Smalla, Nikos Kyrpides, George Kowalchuk, and Jed Fuhrman for your friendly discussions.

Finally, the biggest trip was probably done in the office of my supervisor, Timothy M. Vogel when we spoke among others things about science based on the results I generated regularly. They represented my best hours in the lab. I especially want to thank him for the liberty he gave to me in spite of the relative importance of the Metasoil project for the microbial genomics group. Tim, I hope you have no regrets now…

After these three years working on microbial communities living in the tiny plot 3d from the Park grass field of the Rothamsted experimental station and in spite of the evident geographical limit of this project, I have no scientific and personal regrets. As a result of this experience, I believe that we can work our entire life studying what is present in one gram of soil and discover molecules that can improve our quality of life.

However, I want now to change my scientific horizon to stimulate my curiosity. I don't want to represent Mister "Soil DNA extraction biases" for microbial ecologists. Thus, I hope other scientists will discover the pleasure of soil microbial community study difficulties and I will take my own path. Enjoy…

The majority of people I know believe that a doctorate provides considerable opportunities. However, considering the size of the known universe and our actual limited capacity to travel in space, I prefer to say that my prospects are highly limited both temporarily and spatially. But I will do my best to stimulate science in microbial ecology with the currently accessible environments available for study. I am, of course, extremely jealous of microbial ecologists who will explore life on other planets in the future. To sequence an alien gut or create a Universe Microbiome Project would be fun, wouldn't it?

Aknowledgments :

Remerciements :

Parce que je ne serais tout simplement pas là sans eux, je tiens tout d'abord à remercier mes parents, Jean-Claude et Dominique. Je les remercie du simple fait qu'ils m'ont mis au monde bien sûr, mais aussi parce qu'ils m'ont soutenu psychologiquement et financièrement durant les années d'études qui m'ont été nécessaires avant d'entamer cette thèse. L'éducation qu'ils m'ont transmis à aussi joué un rôle majeur dans mon intérêt ainsi que ma vision personnelle de la science en générale, la biologie en particulier. Me spécialiser dans la microbiologie à été plus un concours de circonstances qu'un objectif en soit, mais je ne regrette en aucun cas de faire ce que je fais. Je voudrais ensuite remercier ma famille en général et m'excuser de ne pas avoir fait suffisamment acte de présence durant ces trois années. Promis, je vais essayer de me rattraper.

Je voudrais maintenant remercier mon directeur de thèse, Timothy M. Voge. Merci pour ces trois années d'études et de travail ponctuées de voyages et rencontres. Ces années m'ont fait oublier la notion de temps libre qui m'était si chère précédemment, mais m'ont enseigné bien plus. Je suis arrivé dans un domaine de recherche sans même connaître les notions les plus importantes (je n'écoutais pas beaucoup en cours). Pourtant j'ai pu prendre suffisamment de vitesse grâce à ce que tu m'a transmis pour m'agripper et rentrer tant bien que mal dans un wagon, puis avancer tranquillement jusqu'à l'avant du train. Ce train symbolise l'avancé de la science bien sûr, et y jouer un rôle, même mineur m'a procuré une joie profonde.

Durant ces trois années, j'ai appris l'anglais (plus ou moins), ai voyagé dans pas mal de contrées, et n'ai même plus peur de l'avion. Je suis même capable de ne pas perdre mon sang froid face à une horde de scientifiques prêts à déceler la moindre faiblesse pour m'écorcher vif (bon là j'exagère un peu). Tu m'as permis de rencontrer un nombre considérable de personnes à travers le monde partageant la même envie de faire avancer la science. En fait, après réflexion, certains partagent probablement plus une envie irrésistible d'augmenter leur facteur H (directement corrélé aux nombres d'articles publiés et de leur utilité pour la communauté scientifique) mais personne n'a dit que la science était parfaite. Après tout ce n'est que l'Homme qui la façonne, et il lui faut un moteur performant, faute de financements suffisants. Personnellement, il me tarde d'augmenter mon facteur H…. Pour tout ça, merci a toi Tim, et bonne continuation.

En parlant de financements, je voudrais remercier la région Rhône-Alpes pour avoir financé mon salaire durant ces trois années. Je n'aurais pas pu faire grand-chose sans eux. De plus, je remercie l'Agence de Nationale pour la Recherche pour le financement qu'elle à fournit dans le cadre du projet ANR Metasoil. Ce projet a permis à l'équipe de génomique microbienne environnementale du laboratoire Ampère de devenir un leader incontestable de l'étude metagenomique du sol, et m'a permis de fournir un certain nombre de données sur la structure génétique des communautés microbiennes présentes. J'en profite pour remercier l'Ecole Centrale de Lyon, qui héberge notre équipe de recherche. Son cadre de vie à certainement été un plus pour moi.

Je voudrais particulièrement remercier Pascal Simonet, qui m'a pris sous son aile en Master1 et qui m'a permis de faire cette thèse avec Tim que je ne connaissais que très peu a l'époque (et oui j'ai été pistonné!). Merci d'avoir eu confiance en moi dès le début. Merci à Libragen de m'avoir permis de rester quelques semaines dans son antre malgré mon manque évident de professionnalisme (merci

donc Renaud Nalin et Patrick Robe). Denis LePaslier et Eric Pelletier du Genoscope m'ont été d'une grande aide durant ce projet. Je les remercie donc pour l'intérêt qu'ils ont porté a mon approche. Merci aussi a Julie Poulain que je n'ai pas eu la chance de rencontrer.

Je voudrais tout particulièrement remercier mes stagiaires, Florentin, Sandra, Davide et Sophie. Merci pour votre aide dans mes travaux, mais aussi pour le plaisir que vous m'avez donné. Vous avez été les meilleurs, et vous me manquez.

Tout aussi logiquement, je souhaite remercier toute les personnes qui ont partagé ma vie professionnelle, et pour certaines ma vie personnelle, durant ces trois années à Ampère. Merci donc Isabelle, Sandrine, Sébastien (probablement le seul à être plus hyperactif que moi dans ce labo), Jean-Michel, Saliou, Laure, Josiane, Marie-Christine, Aude, Silvia, François Buret, François le français aussi, Catherine, Fred, Manu (le maître du piton des neiges), Michael, Aurélie et Maude (les ex inséparable), Jun, Joseph (oui j'aurais dû cacher nos ordinateurs cette fameuse nuit à Genève, encore désolé), Cédric, Sandra, Céline (oui, Enoveo à laissé un vide dans le labo), Sam (courage, tu es le prochain…), Laurine (désolé pour ces trois années a mettre mes pieds sur ton bureau), Mayssa et Jérémie (le père heureux), Margaux (je ne t'oublie pas), Albin (ou Alban je sais plus trop, mais de toute façon je l'ai jamais vraiment beaucoup aimé), et tout ceux que j'ai oublié (on écrit toujours les remerciements en dernier et dans le stress de la fin de thèse pour ma défense…).

Je voudrais aussi remercier Michel Jago, Paul-Edouard Mias, et Rachel Boate pour leur effort d'édition d'une partie de ce manuscrit. Vous comptez tous les trois beaucoup pour moi.

Bien sur, je finirais par remercier les membres de mon Jury pour avoir accepté de prendre de leur précieux temps pour se déplacer en France et se plonger dans ma thèse. Vos critiques ne peuvent que stimuler ma vision de la microbiologie environnementale. Merci a vous donc, Penny, Robert, Gorge, et Jed. Ne me ménagez pas.

Merci.

Annexes:

1. **Metagenomic Comparison of Soil Microbial Community Description by Direct and Indirect DNA Extraction Approaches**
2. **No apparent effect of long term cold storage on a soil metagenome**
3. **Soil metagenomic exploration of the rare biosphere**
4. **Metagenomic exploration of antibiotic resistance in soil**
5. **Ongoing experiments:**
   a. **Generation of high and low GC content soil metagenomes provides access to distinct genetic diversities**
   b. **Metagenomes extracted from dry soil samples archived for decades provide access to highly unusual nucleic diversities**
6. **Scripts for bioinformatics analyses**

The different chapters presented in this manuscript of thesis are inter-connected and aim to tell a comprehensive story to the reader. Thus I decided to not include some of my works to limit confusion and help emphasize major conclusions. However, these aspects of my PhD cannot just be removed and so were placed in the annexes, even if there is no particular links between the different sections.

Section 1 presents a note that compare two metagenome corresponding to two distinct approaches to access DNA from soil. The major conclusion of this work was that the indirect approach (by extracting cells prior DNA) provides access to a similar functional and taxonomical diversity in spite of a low yield.

Section 2 presents an ongoing note that compare four metagenomes corresponding to a fresh soil and the same soil stored at -20°C during one year. The major conclusion of this work was that communities are stable under this storage condition, and that samples can be stored prior extracting and sequencing DNA for soil metagenomic surveys.

Section 3 presents a book chapter that present the Metasoil project and some results extracted from the work presented in the chapter 2, section 3.

Section 4 presents a perspective to explore antibiotic resistance genes in the soil environment and wrote by Jean-Michel Monier. I collaborated in this perspective by emphasizing the interest of global metagenomic comparisons to understand the role of each detected resistance in the different sequenced environments.

Finally, two ongoing projects are partially presented in the section 5, some bioinformatics scripts in the section 6

Journal of Microbiological Methods 86 (2011) 397–400

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Note

Metagenomic comparison of direct and indirect soil DNA extraction approaches

Tom O. Delmont [a], Patrick Robe [b], Ian Clark [c], Pascal Simonet [a], Timothy M. Vogel [a,*]

[a] Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 ECULLY, France
[b] LibraGen, 3 rue des Satellites, 31400 Toulouse, France
[c] Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK

ARTICLE INFO

Article history:
Received 21 March 2011
Received in revised form 15 June 2011
Accepted 18 June 2011
Available online 25 June 2011

Keywords:
Soil metagenomics
DNA extraction
Nycodenz

ABSTRACT

Full pyrosequencing runs of both direct-extracted (high yield, low DNA size) and indirect-extracted DNA (low yield, high DNA size) from the same prairie soil show that the sequence distribution of the majority of the metabolic functions and species detected were statistically similar. Although some microbial functions differed at the 95% confidence interval in bootstrap analyses, the overall functional diversity was the same.

Soil metagenomic approaches require access to high quality DNA in order to construct clone libraries and DNA sequences in sufficient quantity (or representativity) to begin to understand soil microbial ecology (Vogel et al., 2009). Soil DNA extraction is a key step for these metagenomic approaches (Bertrand et al., 2005; Frostegård et al., 1999; Lakay et al., 2007; Delmont et al., 2011) and can be separated in two general strategies. The first strategy, which is more commonly used, is direct DNA extraction and consists of cell lysis directly within a soil sample (e.g. in 1 g of soil) (Ogram et al., 1987; Van Elsas et al., 1997). With the second strategy, indirect DNA extraction, cells are first removed from a soil (e.g. 60 g of soil) and subsequently lysed (Berry et al., 2003; Jacobsen and Rasmussen, 1992). This method separates bacterial and archaeal cells from eukarya cells to some extent by using a density gradient (e.g., Nycodenz density gradient: (Bertrand et al., 2005; Courtois et al., 2001; Lefevre et al., 2008)). Of course, this approach is not the best strategy when eukaryotic sequences are of interest or for studying interactions between eukarya and bacteria or archaea, but can be helpful when eukarya are to be excluded or when high DNA fragments are required (e.g. to construct fosmids and cosmids clones (around 40 kb inserts)). A critical methodological aspect is the DNA yield especially when large quantities are needed for high throughput sequencing or cloning.

Previous studies concluded that in spite of a lower purity, the DNA yield in terms of mass of DNA per mass of soil is greater with direct than indirect extraction (Leff et al., 1995; Steffan et al., 1988) – up to 100-fold greater (Courtois et al., 2001; Roh et al., 2006). A critical question is whether the reduced DNA yield of the indirect extraction strategy results in a significant loss of functional diversity. Due to our current inability to sequence an entire soil metagenome (roughly 10^{15} bp), only the relative genome proportions in the extracted DNA pool can be compared in order to assess the accessibility of the soil microbial genetic richness. The two approaches were compared by analyzing pyrosequencing data from each method, including different sampling strategies due to the important differences in terms of the soil quantity required. Both functional and taxonomical distributions were compared to examine differences in apparent community diversity based on pyrosequencing of DNA resulting from either a direct DNA extraction approach with less soil or an indirect DNA extraction approach with more soil.

Samples were collected from the untreated control plot (3 d) of Park Grass, Rothamsted (England) in March 2009. The Park Grass soil is an internationally-recognized resource and is targeted as a reference for soil metagenomics (Vogel et al., 2009). It is classified as Chromic Luvisol and is a silty clay loam (pH 5.2 measured in H_2O). Soil samples (the top 21 centimeters) were collected during the day with soil cores, fractioned vertically and then homogenized manually by thorough mixing. For both direct and indirect DNA extraction, we used the FastPrep® lysing matrix (MP biomedical). While this approach might not detect all genera present in a soil (Delmont et al., 2011), this protocol is relatively stringent and is thought to lyse the majority of the cells (Howeler et al., 2003; Lakay et al., 2007).

The direct DNA extraction from the 0 to 21 cm core was done as follows: the soil was cored with a 2.5 cm diameter core from 0 to 21 cm depth. The core was subsampled every 3 cm (7 subsamples). Each subsample was mixed manually before DNA extraction. These subsamples were used for direct DNA extraction. Direct DNA

* Corresponding author at: Environmental Microbial Genomics, Laboratoire AMPERE, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 ECULLY, France. Tel.: +33 4 72 18 65 14; fax: +33 4 78 43 37 17.
E-mail address: tvogel@ec-lyon.fr (T.M. Vogel).
URL: http://www.GenomEnviron.org (T.M. Vogel).

0167-7012/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2011.06.013

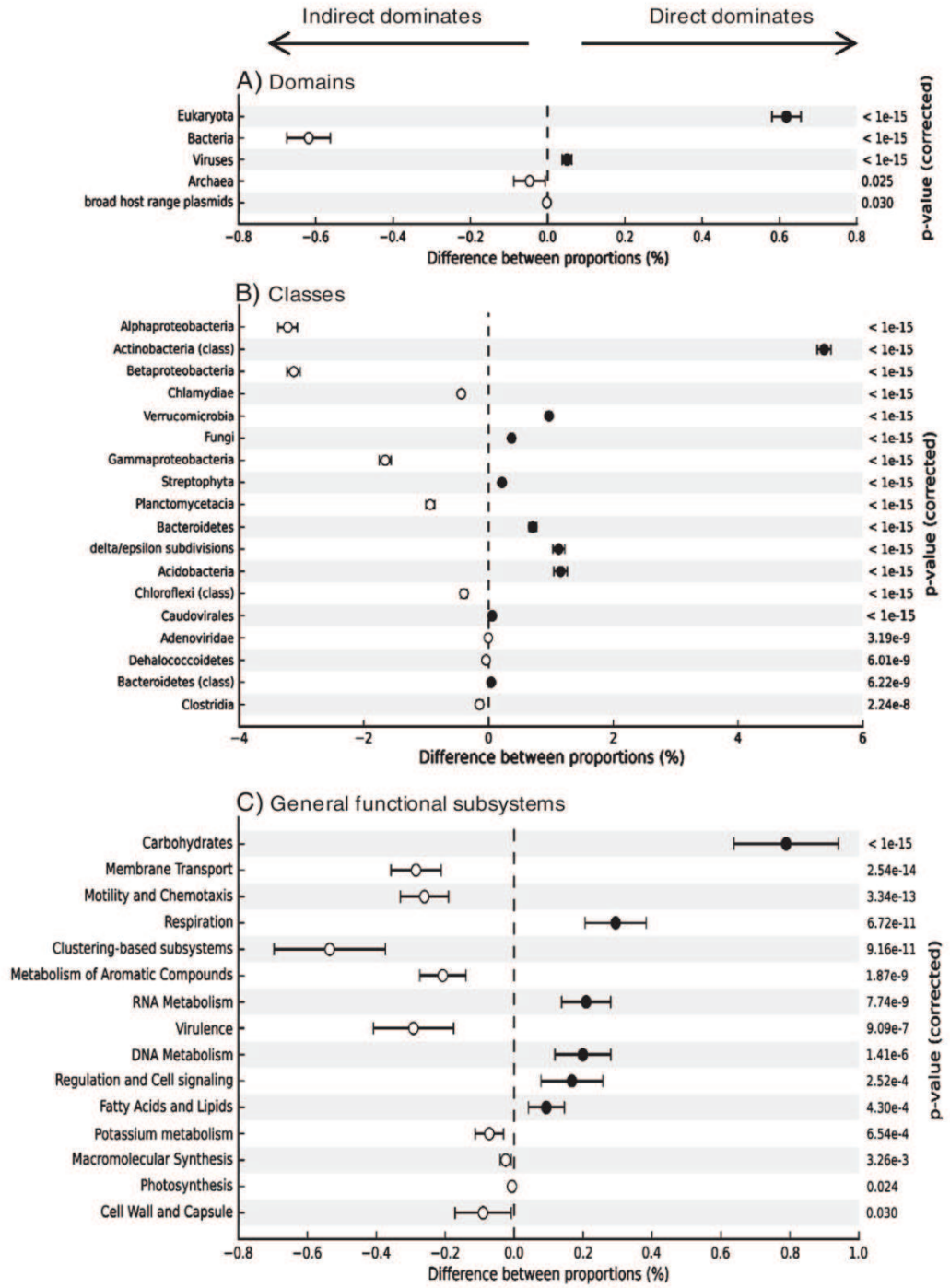


Fig. 1. Differences between relative proportions in percentage of pyrosequences generated with the direct and indirect approaches. Sequences were clustered into domains (panel A), classes (SEED annotation) (panel B) and general functional subsystems (panel C) using the MG-RAST annotation platform and STAMP statistical software. The *p*-value was calculated using the Fisher's exact test. The *p*-value threshold was 0.05 for panels A and C and 5.10^{-5} for the B.

extraction via bead-beating was done as described above with 0.5 g of soil from each subsample. At least 3 μg of DNA was extracted from each subsample. Equal amounts of DNA from each of these 7 DNA extracts were pooled and sent for one pyrosequencing run. The indirect DNA extraction was done as follows: The vertical profile from 0 to 21 cm depth of two cores was subsampled twice (from 0 to 9 cm

and from 10 to 21 cm). The two top fractions were homogenized together to produce one "top" soil and the two bottom fractions were also homogenized to produce one "bottom" sample.This step was necessary to provide a sufficient quantity of soil to cumulate DNA. These two soil subsamples "top" and "bottom" underwent (separately) indirect extraction (with a minimum of 60 g of soil each) with Nycodenz gradient gels (Courtois et al., 2001). The cellular purification was subsequently used for DNA extraction. DNA extraction of the Nycodenz cell fractions was performed as described above. More than ten micrograms of DNA were extracted for each of the two fractions. Equal amounts of DNA from the two subsamples were pooled and sent for one pyrosequencing run.The direct extraction technique produced more than 33 times the DNA per gram of soil that the indirect technique produced. DNA concentrations were determined using the Qubit dsDNA BR Assay (Invitrogen).

The two Titanium (454 Roche) pyrosequencing runs generated more than one million sequences each and were uploaded in the MG-RAST online server (http://metagenomics.nmpdr.org/) for the sequences annotation (Meyer et al., 2008). Only hits with an E-value $< 10^{-5}$ were considered to be significant for further analyses. The distribution of these functions and species was exported to the STAMP software (Parks and Beiko, 2010) for statistical analyses.

The number of Eukarya sequences decreased by more than 40% when using indirect DNA extraction (0.81% of total) as compared to direct DNA extraction (1.43% of total). As a consequence the percentage of sequences related to *Archaea* (from 1.39% to 1.44%) and *Bacteria* (from 97.03% to 97.65%) increase with the indirect approach while this was not the case for the distribution of virus sequences (from 0.14% to 0.09%) (Fig. 1, panel A). However, the number of functional subsystems and species identified at an E value of under 10^{-5} were similar for the two pools of DNA (795 and 796 functional subsystems, 977 and 998 sequenced species for the direct and indirect approach respectively). Thus, superficially they contained the same diversity. Although the overall sequence diversity seems similar between the two pools of DNA, the relative importance of different species and functions were statistically different in some cases. The number of sequences assigned to different bacterial taxonomic classes varied for some of the more dominant classes (Fig. 1, panel B). For example, the *Proteobacteria*: *Alpha-*, *Beta-* and *Gammaproteobacteria* were several percent more common in the DNA from the indirect extraction than in the DNA from the direct extraction. On the other hand, *Actinobacteria*, *Acidobacteria* and fungi were more common in the DNA pool extracted directly from the soil. About 38% of the 76 bacterial classes detected have significantly different proportion of sequences (p-value < 0.05 with the bootstrap analysis) between DNA from the direct and indirect extraction. However, only 30.50% of the 1059 species detected in at least one of the two data sets were significantly different, which is probably due in part to the lower number of sequences associated with each species. Thus, while there are statistically relevant differences, the percentage variations are not large (generally 3% or less). In addition, the number of unique species identified in either of the DNA pools was around 18%.

The two pools of DNA (direct and indirect extraction) had significant differences in their relative proportions within the general function distribution (MG-RAST subsystem hierarchy 1). Over half (53%) of the 28 general function classes had significantly different sequence assignment proportions (cutoff at $E = 10^{-5}$) (Fig. 1, panel C). Specifically, carbohydrates and respiration were more represented in the direct DNA pool than the indirect DNA pool. On the other hand, virulence, membrane transport, motility and chemotaxis, and metabolism of aromatic compounds were more represented in the indirect DNA pool. However, like for the taxonomic comparison, when comparing the lower functional level, only a minority (19.79%) of the 809 functions annotated were statistically different between the two sets (as compared to only 7% for direct DNA extraction of soil cores 20 m apart). Again, this could be due in part to lower number of sequences per function. In addition, the differences between the two pools of DNA in terms of function were smaller than that for the taxonomic assignments. This suggests that different species possess an important part of common genes, so limiting functional distribution differences.

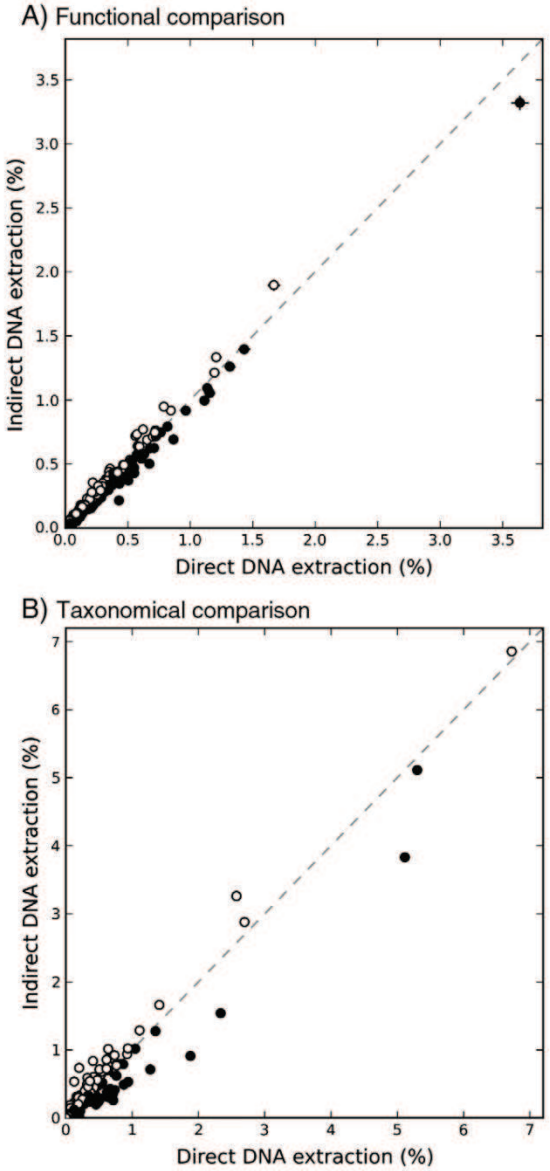


Fig. 2. Direct comparison of MG-RAST output using the STAMP statistical comparison. Panel A shows the relative proportion of the sequences assigned to each functional subsystem in the two DNA pools (indirect vs direct extraction). Panel B compares the relative proportion of sequences assigned to each identified species. In both graphs, the black and white dots represent the percentages in the different pools and help visualize those dominant in the indirect DNA pool (open circles) and those dominant in the direct DNA poll (closed circles).

The two DNA extraction approaches, called direct and indirect methods, were compared for their potential to provide soil metagenomic DNA. The indirect method requires specific material, sufficient quantity of soil, and is time-consuming in comparison to a direct cell lysis in soil. However, the indirect approach appears to possess some advantages compared to the direct DNA extraction approach (e.g., reduced proportion of eukaryotic sequences and increased DNA length). However, in spite of the two different soil sampling strategies due to important differences in the quantity of soil (and DNA) required, the distribution of the annotated pyrosequences between the 809 functions and 1059 species are generally similar (Fig. 2). These comparisons illustrate the differences and similarities between the two DNA extraction approaches performed on the Rothamsted soil. In light of the proposed number of species in a gram of soil (e.g., 10^5 or 100 times more than that annotated) and the relative small fraction of the entire soil metagenome sequenced (4×10^8 bp sequenced/10^{15} bp in gram of soil or 1 out of 2.5×10^6 bp), the potential differences in identification of rare microorganisms based on the two strategies for extracting microbial DNA from soil cannot be easily evaluated, although a bootstrap approach was applied by STAMP software to statistically evaluate relevant differences. The overall DNA diversities were quite similar in the two DNA extraction pools although the indirect method might have accessed a slightly greater soil genetic diversity in spite of the low yield. While the overall genetic diversity will not be clearly demonstrated until significantly more sequencing is accomplished (calculations suggest at least a thousand pyrosequencing runs), the indirect DNA extraction does not appear to be particularly more biased than the direct approach, and thus, is a useful approach for in-depth microbial community sequencing and cloning for the creation of fosmid libraries.

Acknowledgements

We want to thank the French National Research Agency (ANR) for financing Metasoil (Project ANR-08-GENM-025) and the Rhone-Alpes Region for financing Metagene and T.O.D., and Denis Le Paslier and Eric Pelletier of the French national sequencing center (Genoscope) for the sequencing.

References

Berry, A.E., Chiocchini, C., Selby, T., Sosio, M., Wellington, E.M., 2003. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. FEMS Microbiol. Lett. 223, 15–20.
Bertrand, H., Poly, F., Van, V.T., Lombard, N., Nalin, R., Vogel, T.M., et al., 2005. High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. J. Microbiol. Methods 62, 1–11.
Courtois, S., Frostegård, A., Goransson, P., Depret, G., Jeannin, P., Simonet, P., 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. Environ. Microbiol. 3, 431–439.
Delmont, T.O., Robe, P., Cecillon, S., Clark, I.M., Constancias, F., Simonet, P., et al., 2011. Accessing the soil metagenome for studies of microbial diversity. Appl. Environ. Microbiol. 77, 13–24.
Frostegård, A., Courtois, S., Ramisse, V., Clerc, S., Bernillon, D., Le Gall, F., et al., 1999. Quantification of bias related to the extraction of DNA directly from soil. Appl. Environ. Microbiol. 65, 5409–5420.
Howeler, M., Ghiorse, W.C., Walker, L.P., 2003. A quantitative analysis of DNA extraction and purification from compost. J. Microbiol. Methods 54, 37–45.
Jacobsen, C.S., Rasmussen, O.F., 1992. Development and application of a new method to extract bacterial DNA from soil based on separation of bacteria from soil with cation–exchange resine. Appl. Environ. Microbiol. 58, 2458–2462.
Lakay, F.M., Botha, A., Prior, B.A., 2007. Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils. J. Appl. Microbiol. 102, 265–273.
Lefevre, F., Robe, P., Jarrin, C., Ginolhac, A., Zago, C., Auriol, D., et al., 2008. Drugs from hidden bugs: their discovery via untapped resources. Res. Microbiol. 159, 153–161.
Leff, L.G., Dana, J.R., McArthur, J.V., Shimkets, L.J., 1995. Comparison of methods of DNA extraction from stream sediments. Appl. Environ. Microbiol. 61, 1141–1143.
Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., et al., 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinforma. 9, 386.
Ogram, A., Sayler, G.S., Barbay, T., 1987. The extraction and purification of microbial DNA from sediments. J. Microbiol. Methods 7, 57–66.
Parks, D.H., Beiko, R.G., 2010. Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26, 715–721.
Roh, C., Villatte, F., Kim, B.G., Schmid, R.D., 2006. Comparative study of methods for extraction and purification of environmental DNA from soil and sludge samples. Appl. Biochem. Biotechnol. 134, 97–112.
Steffan, R.J., Goksoyr, J., Bej, A.K., Atlas, R., 1988. Recovery of DNA from soils and sediments. Appl. Environ. Microbiol. 54, 2908–2915.
Van Elsas, J.D., Mantynen, V., Wolters, A.C., 1997. Soil DNA extraction and assessment of the fate of *Mycobacterium cholorophenolicum* strain PC-1 in different soils by 16S ribosomal gene sequence based most probable number PCR and immunofluorescence. Biol. Fertil. Soils 24, 188–195.
Vogel, T.M., Simonet, P., Jansson, J.K., Hirsch, P.R., Tiedje, J.M., Van Elsas, J.D., et al., 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. Nat. Rev. Microbiol. 7, 2.

No apparent effect of long term cold storage on a soil metagenome structure

Tom O Delmont and Timothy M Vogel

Environmental Microbial Genomics Group, Laboratoire AMPERE, UMR CNRS 5005, Ecole Centrale de Lyon, 36 avenue Guy de Collongue, 69134 Ecully cedex, France.

An a note

Abstract: The one year cold storage (-20°C) effect of the Rothamsted Park Grass soil was studied using a metagenomic approach. No significant distribution differences were found at both functional ad taxonomical levels, so emphasizing the apparent accuracy of this method to stabilize a complex community structure prior performing metagenomic surveys.

Soil microbial communities are known as one of the more complex on Earth based on the actual knowledge of microbial ecologists. Their diversity estimation varies between 10^4 (Torsvik, Ovreas et al. 2002; Roesch, Fulthorpe et al. 2007)) and 10^7 species per gram (Gans, Wolinsky et al. 2005) as a function of the method used and is possibly highly underestimated due to DNA extraction limits (Delmont, Robe et al.). Consequently, soil nucleic diversity represents a gold mine of information and is actively studied to discover activities of interest, improve agriculture and biodegradation processes, but also for a wide range of fundamental researches.

Interestingly, recent advances make it possible to deeply study these communities using high throughput sequencing technologies (Kahvejian, Quackenbush et al. 2008; Shendure and Ji 2008). However, while the quick growing size of the generated metagenomes due to the emergence of new sequencing technologies improves considerably results sensitivity, several methodological steps impact the structure of these datasets. Thus, to define these biases is crucial to perform accurate metagenomic studies and surveys and to refine the picture of communities *in situ*. As examples, lysis stringency appears to impact strongly the structure of soil metagenomes {Delmont et al., in press} while the use of an indirect approach (extraction of cells) prior DNA extraction does not (Delmont, Robe et al. 2011).

One step that can potentially impact the structure of a soil metagenome is storage interlude. In fact, while it is generally recommended to use fresh samples, to store soil for further analyses is often necessary. As an example, a series of microcosms conditions were perform in our lab. Soil samples were stored at -20°C after four months of incubation, and DNA directly extracted from a part of these samples and sequenced using the pyrosequencing Titanium technology. Two months were necessary to generate, annotate and study metagenomes. Then some conditions were selected for a deeper sequencing effort using Illumina HiSeq technology due to assembly interests. Thus, DNA was extracted again from samples stored for months to provide the quantity necessary for this second sequencing stage. Of course, to extract a sufficient quantity of DNA from all samples is a more suitable approach, but is time and cost consuming.

Different conditions were experimented to stabilize the structure of communities close to its initial phase during storage period. However, while the activity of microorganisms can be restricted by limiting water content or temperature, it is harsh to prevent totally any community evolution. In fact, several studies emphasized different aspects of soil storage impact on communities structure and activity (e.g., (Lauber, Zhou et al. 2010) (Pesaro, Nicollier et al. 2004) (Sessitsch, Gyamfi et al. 2002), more or less important as a function of the condition and the tool used. Globally, frozen samples appears to be more efficient than air-frying to stabilize communities (e.g., (Wallenius, Rita et al.)) and thus is the more common method in soil microbial ecology.

With the quick emergence of metagenomics in microbial ecology and the sensitivity of statistical tests that can be use to compare samples, it is now critical to observe the impact

of soil storage not only on biomass, fingerprints and enzyme activities but also on the structure of generated metagenomes. In this aim, we studied the impact of cold storage (-20°C during one year) on the relative distribution of taxonomic and functional subsystems of the Rothamsted soil metagenome, the reference soil of the Terragenome consortium (Vogel, Simonet et al. 2010).

Soil samples corresponding to the 20 first centimeters of the Park Grass experiment (plot 3d) were transported on ice from England to France in July 2010. These samples were pooled and sieved at 2mm. 50 grams of sieved soil were transferred in a falcon (volume of 50ml) and two distinct DNA extractions where done using the MP BIO 1O1 bead beating lysis E (see Delmont et al, in press for more details). DNA pools so obtained were directly sequenced using the Titanium pyrosequencing technology and the falcon stored at -20°C during one year. Then two other DNA extractions were done using the same protocol, and DNA pools were sequenced with the same technology.

The four datasets were cleaned using the Roche software and artificial duplicates were deleted using cd-hit-454 with default parameters. 1.2 million of sequences (±7.41%) were generated for each metagenome.

Then metagenomes were annotated on the MG-RAST-CLOUD server (Meyer, Paarmann et al. 2008) with an e-value cut-off of 10^{-5} and functional and taxonomical tables exported to STAMP (Parks and Beiko 2010) for statistical analyses. The distribution of subsystems was normalized as a function of the number of annotated sequences in each dataset.

A total of 2367 genera (M5NR annotation) and 824 functional subsystems (level 3) were detected in these datasets. Based on the taxonomical annotation, the cold storage appears to impact the structure of this soil metagenome (figure 1).

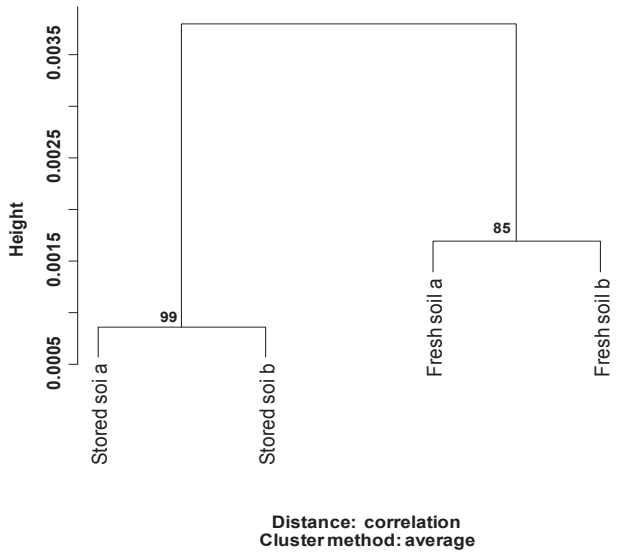


Figure 1: Dendogram based on the relative distribution of 2367 genera among four datasets.

However, when using a statistical test (t-test with equal variance or Welchs' t-test) and the Bonferroni correction, no statistical differences were detected between metagenomes corresponding to fresh and cold stored soils in spite of some variations in functional and taxonomical distributions (figure 2). Thus a period of one year of cold storage at -20°C of a soil sample appears to stabilize sufficiently the structure of communities to prevent unwanted distribution differences.

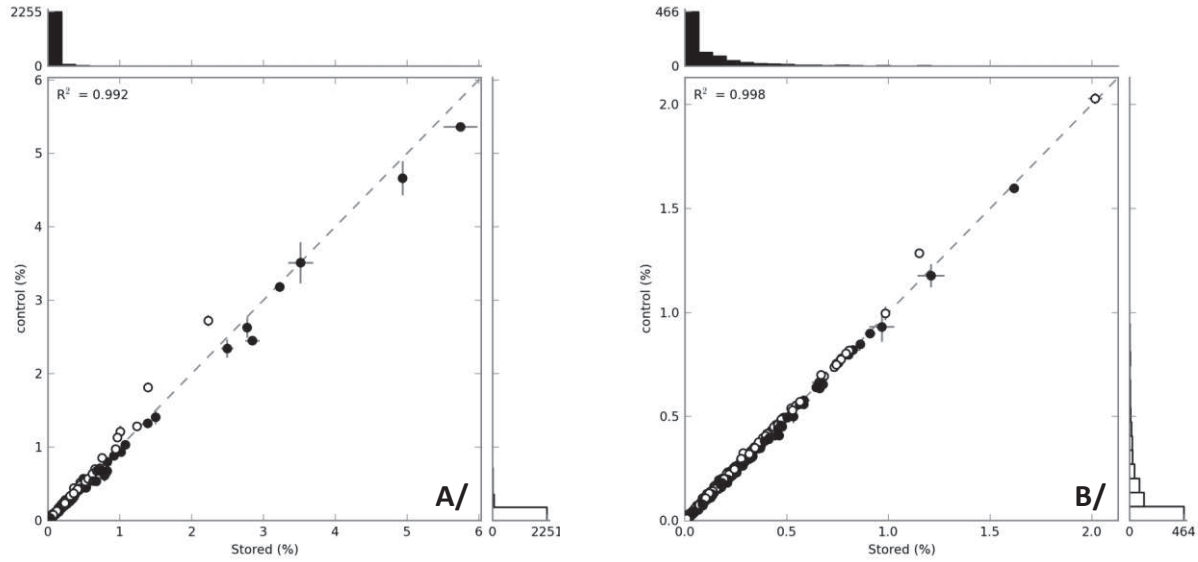


Figure 2: Relative distribution of taxonomical (genera level, panel A) and functional subsystems (functional level 3, panel B) between groups of metagenomes corresponding to biological duplicates and generated with DNA extracted from fresh and stored soil. Line sand crosses observed in some subsystems represent variations between replicates.

Of course, additional experiments have to be done with other soils and environments or longer times to increase our knowledge about cold storage on metagenomic structure. However, based on these results soil samples can possibly be stored during relatively long periods without consistent structural modifications of the generate metagenomes. This temporal flexibility could help scientists defining complex experimental designs that require the storage of samples for further analyses. In particular, it could be possible to apply more efficient or accurate DNA extraction protocols discovered later or to perform additional sequencing runs if necessary (e.g., when runs failed or a new technology emerge).

Finally, experiments based on metatranscriptomic and metaproteomic analyses have to be done to study the effect of cold soil storage on these molecules and proteins so studied.

Acknowledgments:

We want to thanks Penny Hirsch and Ian Clark for sanding the soil, and Denis LePaslier and Eric Pelletier for sequencing the runs. We also want to thanks the Rhone-Alp region and Metasoil ANR for the funding they provided for this study.

References:

Delmont, T. O., P. Robe, et al. "Accessing the soil metagenome for studies of microbial diversity." Appl Environ Microbiol **77**(4): 1315-1324.

Delmont, T. O., P. Robe, et al. (2011). "Metagenomic comparison of direct and indirect soil DNA extraction approaches." J Microbiol Methods **86**(3): 397-400.

Gans, J., M. Wolinsky, et al. (2005). "Computational improvements reveal great bacterial diversity and high metal toxicity in soil." Science **309**(5739): 1387-1390.

Kahvejian, A., J. Quackenbush, et al. (2008). "What would you do if you could sequence everything?" Nat Biotechnol **26**(10): 1125-1133.

Lauber, C. L., N. Zhou, et al. (2010). "Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples." FEMS Microbiol Lett **307**(1): 80-86.

Meyer, F., D. Paarmann, et al. (2008). "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes." BMC Bioinformatics **9**: 386.

Parks, D. H. and R. G. Beiko (2010). "Identifying biologically relevant differences between metagenomic communities." Bioinformatics **26**(6): 715-721.

Pesaro, M., G. Nicollier, et al. (2004). "Impact of soil drying-rewetting stress on microbial communities and activities and on degradation of two crop protection products." Appl Environ Microbiol **70**(5): 2577-2587.

Roesch, L. F., R. R. Fulthorpe, et al. (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." ISME J **1**(4): 283-290.

Sessitsch, A., S. Gyamfi, et al. (2002). "RNA isolation from soil for bacterial community and functional analysis: evaluation of different extraction and soil conservation protocols." J Microbiol Methods **51**(2): 171-179.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.

Torsvik, V., L. Ovreas, et al. (2002). "Prokaryotic diversity--magnitude, dynamics, and controlling factors." Science **296**(5570): 1064-1066.

Vogel, T. M., P. Simonet, et al. (2010). TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Microbiol. **7:** 252.

Wallenius, K., H. Rita, et al. "Sample storage for soil enzyme activity and bacterial community profiles." J Microbiol Methods **81**(1): 48-55.

Chapter 33

Soil Metagenomic Exploration of the Rare Biosphere

TOM O. DELMONT, LAURE FRANQUEVILLE, SAMUEL JACQUIOD, PASCAL SIMONET, AND TIMOTHY M. VOGEL

33.1 INTRODUCTION

Microorganisms can be considered the fundamental driving force of the biosphere and have dominated life on Earth for over 3 billion years. They have colonized all ecological niches, from caves [Pasic et al., 2010] to the stratosphere [Wainwright et al., 2003] and from deep ocean thermal vents [Huber et al., 2007] to deserts [Pointing et al., 2009] and polar snow [Larose et al., 2010; see also Vol. II]. Other forms of life are completely dependent upon these microscopic organisms. For example, humans cannot survive without the rich microbial flora inhabiting their own bodies [Turnbaugh et al., 2006]. When examining the importance of different ecosystems, soil stands out as the habitat on Earth that harbors by far the largest microbial diversity [Torsvik et al., 1990, 2002; Whitman et al., 1998; van Elsas et al., 2006; see also Chapter 9, Vol. II]. The genetic resources in a single gram of soil possess 3 million times more sequences than in the human genome; and only three grams of soil contain more bacteria than the Earth does humans.

Soil microbial communities are known to drive major geochemical cycles [Falkowski et al., 2001], to support healthy plant growth [Ortíz-Castro et al., 2009], and to degrade organic matter and pollutants [Singh et al., 2008]. However, little is known about the vulnerability of their key functions and how they respond to human-induced environmental perturbations, such as climate change and land use. Cultivation-based approaches, although limited in scope, have already shown that soil harbors diverse antibiosis-related functions [Adesina et al., 2007], pollutant degrading bacteria [Richard and Vogel, 1999], plant growth promoting bacteria [Leveau, 2007], bacteria resistant to heavy metals and antibiotics [Baker-Austin et al., 2006; Dcosta et al., 2007; Demaneche et al., 2008], and bacteria capable of surviving in extreme environments [Dib et al., 2008; Hery et al., 2003].

However, only a fraction (less than 0.5%) of the microbiota in soil has been cultivated using any of an array of techniques [Amann et al., 1995; Davis et al., 2005; Stalay and Konopka, 1995], thus limiting considerably our understanding of these microorganisms. In order to access this tremendous biodiversity, researchers have recently developed sophisticated molecular techniques that can explore and exploit a wide range of soil biodiversity. These techniques derived their pertinence from the direct analysis and exploration of microbial community DNA (metagenomics) [Courtois et al., 2003; Demaneche et al., 2008; Ginolhac et al., 2004]. In this context, soil metagenomics is the study and exploitation of the collective genome of all organisms present in a particular soil sample [Handelsman et al., 1998]. The critical technical issue today is the nonbias access to the entire soil microbial diversity.

To date, there have been only superficial, although ambitious, attempts to explore the soil metagenome. Due in part to incomplete genomic extraction techniques and incomplete sequencing, today scientists are only exploring a minute fraction of soil metagenomes. Potential errors in data interpretation due to the limited vision of these small community fractions might be driving hypotheses away from incorporating functional redundancies and the importance of the minor members of the soil community (rare biosphere). Recent advances in high-throughput cloning

Handbook of Molecular Microbial Ecology, Volume I: Metagenomics and Complementary Approaches, First Edition. Edited by Frans J. de Bruijn.

and new generation sequencing (see Chapter 18, Vol. I) might improve the prospect of completely sequencing the soil metagenome and thus obtaining a more complete picture of the soil microbial community function.

33.2 SOIL METAGENOMIC LIMITS

The soil microbial community is relatively diverse [Curtis et al., 2002, Robe et al., 2003] and appears to have the highest prokaryotic diversity of any environment [van Elsas et al., 2006; Roesch et al., 2007]. One gram of soil is reported to contain up to 10 billion microorganisms and thousands of different species [Knietch et al., 2003]. Thus, the potential access to the genomes of these microorganisms requires diverse approaches due in part to their differential location in the soil microstructure, their differences in membrane consistency, and their differences in relative numbers. In order to overcome some of these difficulties, different metagenomic associated methods have been developed. All of these methods themselves have some type of shortcoming; however, when pooled together, they will provide greater depth than any technique by itself. For example, DNA purification techniques have numerous biases that limit the quantity of the so-called metagenomic DNA [Delmont et al., 2011]. Second, the proportion of the genomes is nonuniform [Ranjard and Richaume, 2001]; thus, access to the less represented genomes is reduced statistically. In any case, these type of techniques must be applied in order to overcome cell culture limitations [Amann et al., 1995]; but because of their own limits, a complete soil metagenomic study remains an important challenge.

Classically, different approaches have been applied to estimate soil biodiversity richness. The typical method is by the amplification, cloning, and then sequencing of the ribosomal small unit RNA genes ("16S"; see Chapter 16, Vol. I). Rarefaction curves can be created based on genetic diversity and sequence proportions in order to estimate the total soil microbial diversity with (and unfortunately within) a limited sample. The 16S clone libraries constructed by Tringe et al. [2005] and Morales et al. [2009] were extrapolated to determine soil microbial diversities between 3000 and 4000 different species (based on 97% similarity of the 16S sequence for species identity) after sequencing 1700 and 5000 clones, respectively. Roesch et al. [2007] did 16S amplification and direct pyrosequencing (without the cloning step) of a part of this gene. With this approach and 55,000 16S sequences, they evaluated the diversity to be 6000 different species based on 97% similarity of the 16S sequence for species identity.

A completely different approach based on genome reassociation kinetics measures the rate of rehybridization of metagenomic DNA. The kinetic measurements are evaluated based on different hypothetical species distributions. For example, Torsvik et al. [1990]; see also Chapter 2, Vol. I estimated soil diversity to be 10,000 different species assuming that all species were equal in number. More recently, Gans et al. [2005] estimated the diversity to be 8×10^6 species in the tested soil by assuming a log-normal species number distribution.

One explanation for this wide range of predicted species numbers in soil can be the metagenomic DNA diversity represented in the 16S amplification mix used for their analyses. Most researchers amplify the 16S gene from a minor (and possibly nonrepresentative) fraction of the metagenomic DNA present in a gram of soil. Thus, the amplification mix contains only about 10 ngs (less than 1% of extracted DNA), which represent only 10^6 copies of prokaryote genomes, from about several micrograms of extracted DNA per gram of soil (quantity is dependent on the soil type and sample depth). To reduce this limitation, Roesch et al. (2007) used a total of 10 μgs of DNA (96 different 16S amplifications). However, the diversity estimate were not drastically different from the results obtained by Morales et al., [2009] with 5000 16S clones using one 16S variable region of the DNA extracted. These observations do emphasize the existence of a much more important limit. The sequence redundancy due to nonuniform species proportions leads to the dilution of the less represented ones during the first cycles of 16S amplification, thus limiting considerably the diversity obtained. This possible limitation seems more prevalent and likely than that due to the low initial amount of DNA in the PCR mix.

Another important factor that often leads to underestimated biodiversity results is the DNA extraction protocol. Most studies focus their efforts on assessing microbial diversity with only one extraction technique. The microbial diversity uncovered with different methods demonstrates the danger of such a narrow approach if the goal is to optimize access to biodiversity. The DNA extraction appears to be a crucial step in soil metagenomics because it defines in part the extent of the information available. To overcome this bias, a broad DNA recovery strategy is needed [Delmont et al., 2011; see Chapters 10 and 11, vol. II].

33.3 SOIL METAGENOMIC ADVANCES

Soil metagenomic studies have rapidly expanded since 1998 (**Fig. 33.1**) and an important number of tools and methods have been developed to recover and study prokaryotic DNA diversities [Delmont et al., 2011; Rajendhran and Gunasekaran, 2008; see Chapter 10, Vol. II]. The two major metagenomic advances were high-throughput cloning and new-generation sequencing approaches. These new technologies became precious

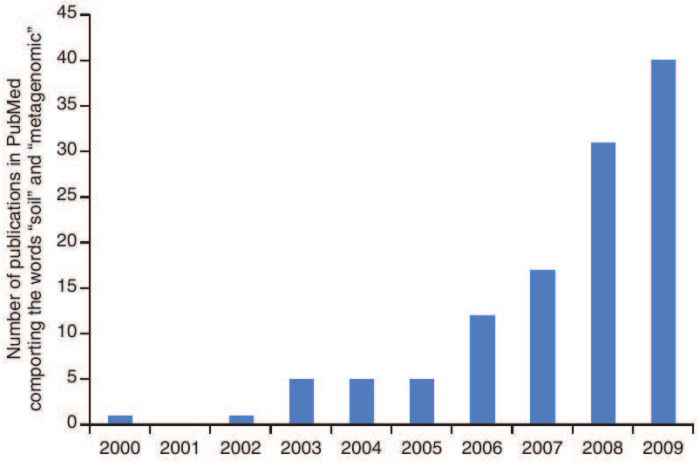


Figure 33.1 Number of publications in Pubmed with the words "soil" and "metagenomic" from 2000 to 2009.

tools for the discovery of new species, genes and functions of interest. While there is a tendency to explore the metagenomics of a soil through sequencing, the clone library approach often provides better validation of newly discovered genes.

Genomic studies stand at the vanguard of science as technological advances are providing access to the functioning of different biological systems. After the Human Genome Project (3 billion base pairs) and the Sargasso Sea marine sequencing effort (6 billion base pairs), soil exploration constitutes a new and ambitious challenge as it harbors about 10 trillion base pairs (if 10^4 species are present) to 1000 trillion base pairs of diversity (if 10^6 species are present) per gram of soil. Just as knowledge of the human genome promises to revolutionize medical science, the application of genomic technologies to microbial evolution and environmental biology promises to revolutionize microbiology. The soil microbial community represents a true goldmine for genes that encode novel biocatalysts involved in either biosynthetic or biodegradation processes, including the production of polyketide synthases [Ginolhac et al., 2004, 2005] and the degradation of human-made pollutants [Handelsman et al., 2002; Boubakri et al., 2006; Galvao et al., 2005; see also Sections 6 and 7, Vol. II]. In spite of new possibilities to study metagenomic DNA, only a minor fraction of soil metagenomes have been sequenced to date and only predominant species were discovered and studied. The majority of soil genetic richness and function still remains to be discovered.

33.4 AN INTERNATIONAL PARTNERSHIP TO OPEN THE BLACK BOX

As new sequencing technologies are increasing both in sequence length and number of sequences [Shendure and Ji, 2008], massive DNA sequencing is becoming a serious possibility. As a consequence, large national and international projects were launched to sequence environments such as the ocean and the human gut. Yet, soil nucleic explorations have been relatively limited in part due to the enormous soil heterogeneity, the vast biodiversity, and the difficulty to access this biodiversity locked up in the soil matrix. Soil microbiologists, microbial ecologists, geneticists, molecular biologists, and bioinformaticians have agreed to collaborate to provide the first complete metagenomic sequence(s) of a soil. This metagenome sequence data will constitute a "reference" to which other soils around the world could be compared. Thus, other metagenomic projects devoted to sequencing parts of different soil genomes through the world will be able to use this complete metagenome as a scaffold for annotation of "core" genes representing common soil microorganisms in other soils and as a basis for estimating differences in diversity, completeness, and richness between soils. The combination of metagenomics approaches and broad-scale sequencing will open a totally new era in soil microbiology with advances ranging from detection of climatic indicators and greenhouse gas production to drug discovery, as well as from correlating biodiversity and function to predicting the biosphere's resilience to human-induced perturbation. The soil system chosen for investigation, Park Grass, Rothamsted (UK), is a "charismatic and internationally recognized resource" [Silvertown et al., 2006]. This unique long-term ecological site includes ongoing experiments that have been running for over 150 years. The research center at Rothamsted provides a history of soil biology and chemistry as well as an archive of soil samples from detailed studies of different plot treatments (http://www.rothamsted.ac.uk/). Metadata are available concerning climate, soil use, and chemical inputs.

The success of soil metagenomics depends to a large extent on intelligent decision-making concerning sample selection, DNA extraction methods, cloning strategies, screening methods, technological advances in sequencing approaches, and data management and sharing. Recent progress in methods to capture the vast scale of genetic diversity within soil microbial communities will enable deep metagenomic sequencing [Morales et al., 2009; Delmont et al., 2011]. These advances include methods to dissect the community using DNA or cell extraction separation methods that can be optimized to detect abundant and rare members of the community. These promising new metagenomic approaches, which are currently being applied to the Rothamsted soil, rely on massive parallel high-throughput sequencing of DNA extracted from soil microorganisms. Furthermore, DNA microarrays loaded with soil-dedicated probes have been used in order to identify the diversity and relative abundance of phylogenetic and functional genes in different DNA fractions [Gebert et al., 2008; Huyghe et al., 2008; Sanguin et al., 2006; Ward et al., 2007]. Cloning of large DNA fragments from soil to create metagenomic libraries enables the examination of genetic pathways and contiguous sequences (e.g., operons). In addition, the new-generation sequencing methods, including enhanced 454 Titanium with up to 1000-bp reads, will complement the cloning-based approaches and avoid the associated cloning bias (DNA weight and cloning step). Data generated will provide insight into which soil organisms are actually involved in soil processes, especially when RNA is extracted. These approaches, coupled with novel bioinformatics methods (e.g., MG-RAST [Meyer et al., 2008; see Chapter 37, Vol. I]), provide an in-depth analysis of both generated sequence and array data. One bioinformatics limit stands out in the interpretation of this data; there are insufficient sequenced genomes or sequences that have been correctly annotated; therefore, around 50% of soil metagenome sequences cannot be used due to the lack of homologues in the databases.

33.5 THE METASOIL PROJECT

The Metasoil project, funded by the French National Research Agency, is initiating the Park Grass metagenome sequencing and cloning as well as the international consortium (Terragenome) [Vogel et al., 2009]. The first part of this project is a 100-Gbp soil metagenomic DNA sequencing (length up to 500 bp) with up to 100 titanium pyrosequencing runs. This sequencing effort corresponds to the equivalent of more than 20,000 prokaryotic genomes. Moreover, in parallel with the in-depth soil pyrosequencing, Metasoil includes the construction of a two million fosmid (40-kb inserts) library to serve the international scientific community for studying this soil metagenome and for searching for new genes of interest. The fosmid library and pyrosequencing runs will be done using various metagenomic DNA samples extracted from the same soil, but with a panel of different methods (soil, cells, and DNA separations) to access a maximum diversity [Delmont et al., 2011]. Due to DNA extraction biases, such as uneven species proportion and soil heterogeneity, a mixed DNA acquisition strategy is being evaluated in order to enhance complete soil metagenome exploration. This strategy is based on variable metagenomic DNA separation methods and aims to increase the soil DNA diversity recovery significantly. Moreover, the strategy should standardize the relative species proportion at the nucleic level as much as possible and, thus, increase access to the rare biosphere.

33.6 SOIL METAGENOMIC DNA RECOVERING STRATEGY

33.6.1 Soil Fractionation

The Park Grass untreated control plot (3d) selected for the project possesses a 249-m^2 surface (13.28 × 18.75 m). The sampling strategy consists of a randomized soil sampling in different areas of the plot [horizontal sampling] and at different depths (vertical sampling). Preliminary prokaryote diversity studies were carried out using RISA (ribosomal intergenic spacer analysis fingerprint after extracting DNA with MP Bead beating Fast Prep protocol [Griffiths et al., 2000]) and showed a stable diversity pattern horizontally. But significant differences have been discovered as a function of depth (data not shown). Further experiments were done based on different depth fraction (0–3 cm, 3–6 cm, and 18–21 cm) using a phylochip technology (**Fig. 33.2**; see Chapter 58, Vol. I).

Results show that some families are detected only in one soil fraction and that the proportion of the different families can fluctuate strongly with depth. Differences in soil depth appear to provide access to different metagenomic DNA fractions and can be helpful for increasing the diversity recovered and for limiting sequence redundancy.

33.6.2 Prokaryote Cells Extraction

After soil fractionation, we compared different methods for DNA recovery. In order to obtain metagenomic DNA, cells can be lysed within the soil sample [Orgam et al., 1987; van Elsas et al., 1997] or first removed from the matrix and then lysed [Berry et al., 2003]. This second method is commonly called indirect extraction and attempts to separate prokaryotic cells [the cell ring]

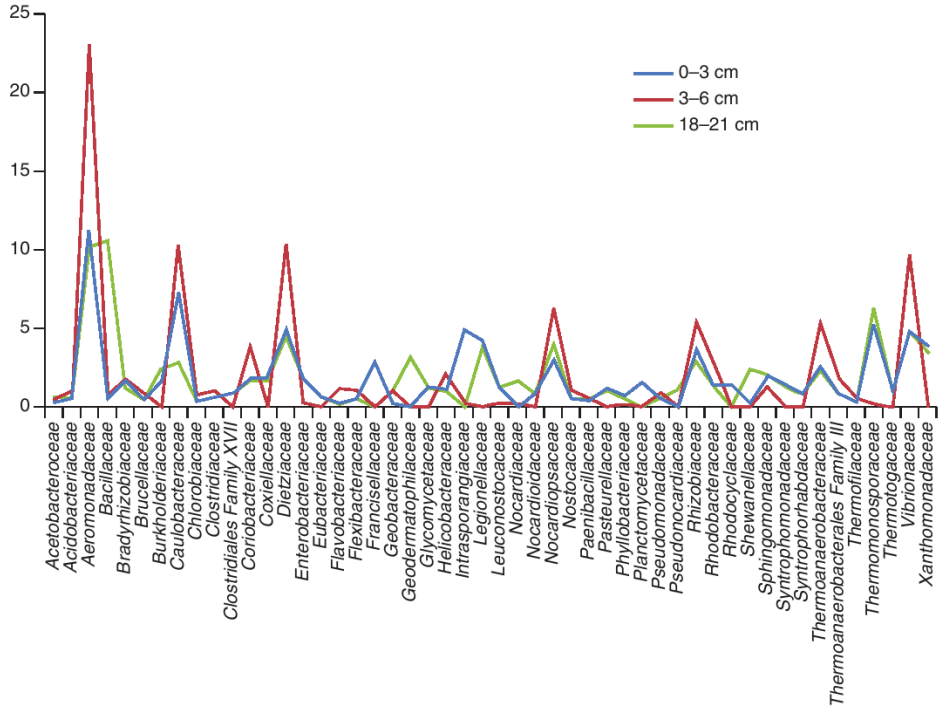


Figure 33.2 Relative proportion of families as a function of soil depth. The intensity corresponds to the microarray hybridization fluorescence signal.

from eukaryotic cells via a Nycodenz density gradient [Courtois et al., 2001; Lefevre et al., 2008]. To compare the two methods, DNA was extracted with the same protocol (MP Bead beating Fast Prep [Griffiths et al., 2000], but with direct (horizontal and vertical soil sampling) or indirect (with different centrifugation speeds during the cell ring formation) extraction approaches. The RISA profiles were then compared in a principal component analysis (PCA) (**Fig. 33.3**). Results demonstrate the differences between the direct and the indirect extraction methods in terms of metagenomic DNA. The RISA profiles are clearly different between the two methods, showing different peaks and intensities (data not shown). However, varying the centrifugation speed does not significantly change the DNA within the Nycodenz ring.

33.6.3 Prokaryote Cells Separation

Using the indirect extraction method, a prokaryote cell ring appears inside the density gradient. But interestingly, some cells stay above the ring and can be separated from the others based on their centrifugation velocity differences. The number of these cells above the Nycodenz ring was limited: Only a few nanograms of DNA could be extracted, but the relative species proportions vary strongly during the gradient (**Fig. 33.4**).

This physical cell separation increases considerably the number of detectable species, including those that would normally be undetected, due to their separation from the regular prokaryotic cell ring. Moreover, centrifugation speed can be modified because it changes relative species proportions in less dense fractions without affecting the cell ring.

33.6.4 Cells Lysis Stringency Fractionation

After separating soil and cells, extractable DNA has been further separated using cells lysis stringency. The process varies between protocols in terms of physical, chemical, and enzymatic lyses. Experiments were carried out directly from soil, or from the cell ring, in suspension or in agarose plugs. These approaches accessed different diversities with different species proportions represented in the extracted DNA (**Fig. 33.5**).

Moreover, because this step is essential in any DNA extraction protocol, it produces considerable biases in soil DNA recovering. Every DNA extraction protocol is biased to some degree; therefore, the choice of one unique approach is dependent on the researcher. An alternative would be to use several different lyses

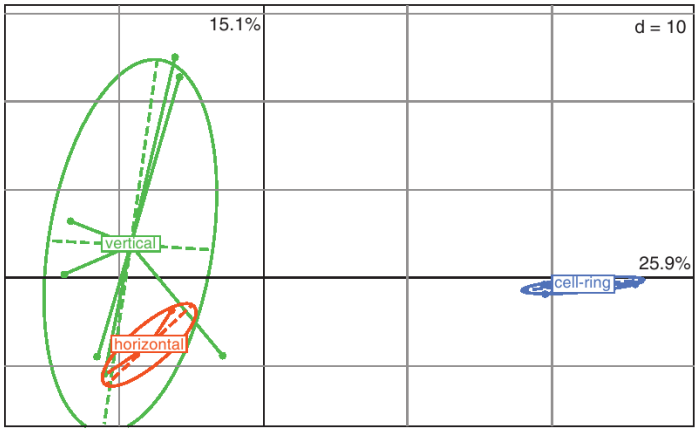


Figure 33.3 Principal component analysis [showing the first and second components] of the matrix data containing the RISA electropherograms corresponding to vertical sampling (seven fractions of three centimeters between 0 and 21 cms), horizontal sampling (fraction 3–6 cm for four distinct areas of the Park Grass 3d plot), and cell rings formed with different centrifugation speeds (1000*g*, 2000*g*, 5000*g*, and 9000*g*).

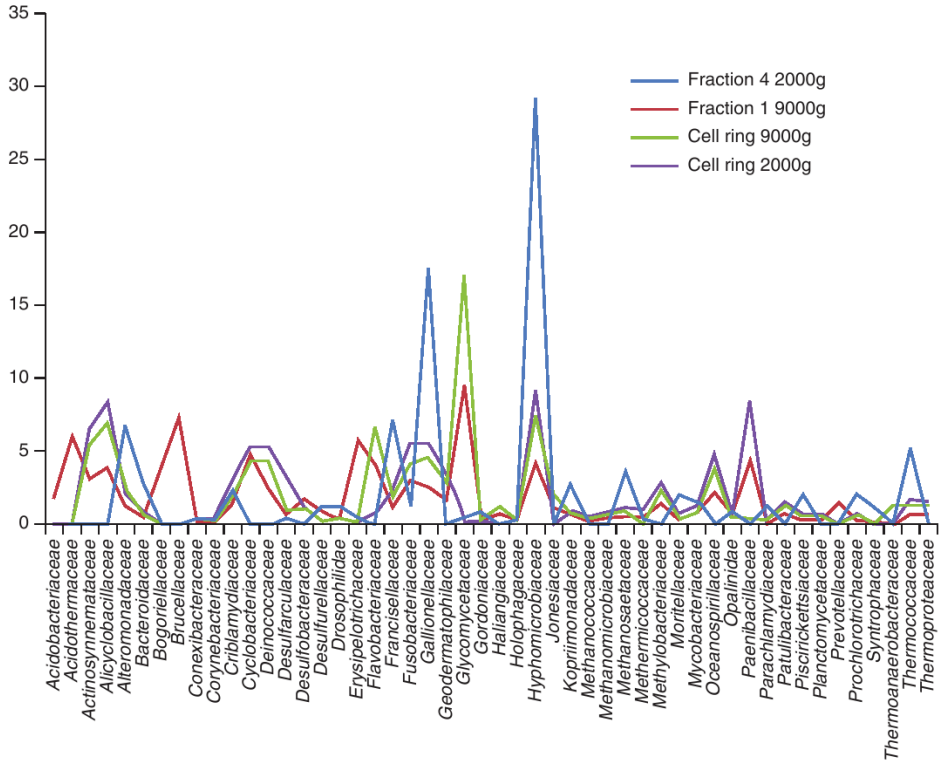


Figure 33.4 Relative proportion of families as a function of a density gradient. The intensity corresponds to the hybridization fluorescence signals.

methods before starting metagenomic studies [Delmont et al., 2011].

33.6.5 Molecular DNA Weight Separation

Finally, metagenomic DNA (agarose plug protocol) were separated as a function of DNA molecular weight in a pulsed field gel electrophoresis (PFGE). Phylochip analyses showed some prokaryote diversity differences when extracting two bands at different points of the DNA smear (50 and 250 kbp) (**Fig. 33.6**).

In particular, some families were detected only in one of these two bands. These results demonstrate that nucleic diversity separation occurred during the soil metagenomic DNA migration. The diversity differences appear to be

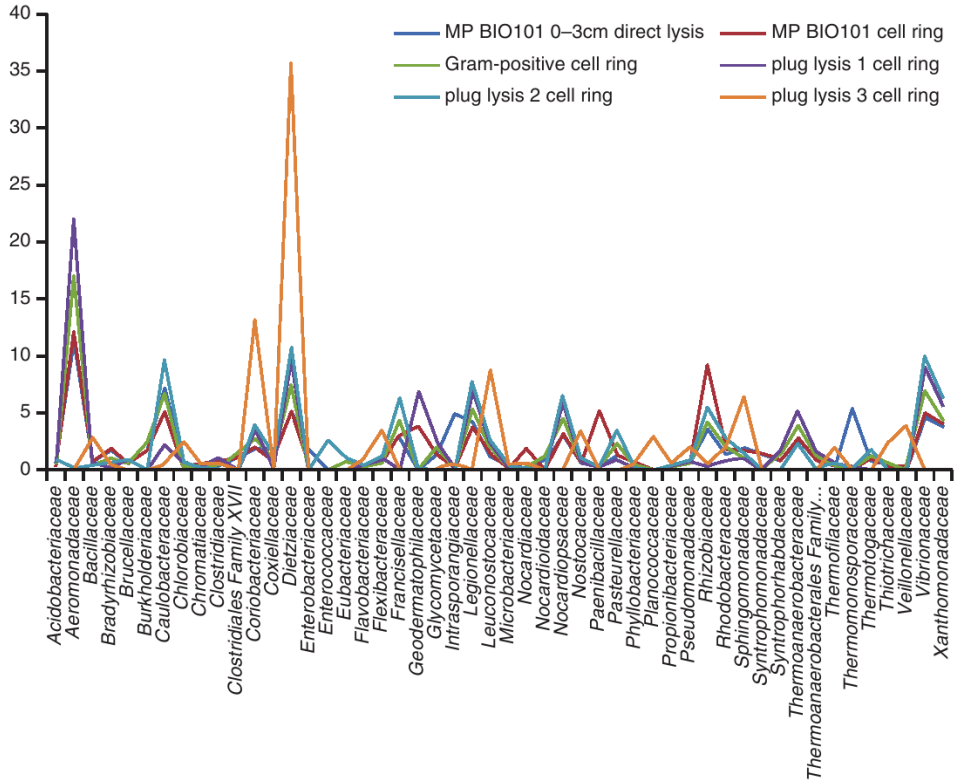


Figure 33.5 Relative proportion of families as a function of the different lyses. The intensity corresponds to the hybridization fluorescence signals.

less than for the three other DNA separation methods (soil, cells, and lysis stringency variables), but they suggest a bias involved in fosmid library constructions as DNA needs to have a molecular length between 25 and 40 kbp.

33.7 METAGENOME COMPARISONS

In spite of its ambitious objectives, Metasoil is just one of a multitude of environmental sequencing projects ongoing in laboratories from around the world [Dinsdale et al., 2008; Willner et al., 2009]. Thus, in addition to data production, there will be future requirements for analyzing large datasets. In the future, bioinformatics will be a critical tool to describe the distribution and function of the 10^{30} prokaryotes on earth. However, the specific problems, biases, and limits of these comparisons need to be well understood in order to accurately assess these data. Moreover, this understanding is crucial to choose the best normalization to compare the distribution of species and functions. A nonexhaustive list of 10 variables that might limit strong metagenomic data comparisons is suggested here:

1. The metagenome diversity can vary tremendously between environments. This diversity is known to be high in sediments and soils but limited in animals and very low in some extreme environments like acid mine biofilms.
2. The average genomes sizes change as a function of the environment.
3. The metagenome variability at a fine scale can be important in specific environments.
4. The metagenome variability over time can be high in some environments and low in stable ecosystems.
5. The metagenomic sequencing depth is unequal due to variable efforts in different projects.
6. The DNA extraction approaches used are often different, thus leading to sequencing different diversities.
7. The length of sequences varies as a function of the sequencing technology used and can influence the number of functions and species detected as well as the quality of the sequence match.
8. The phylogenetic distribution of the complete sequences used for sequence annotations is biased.

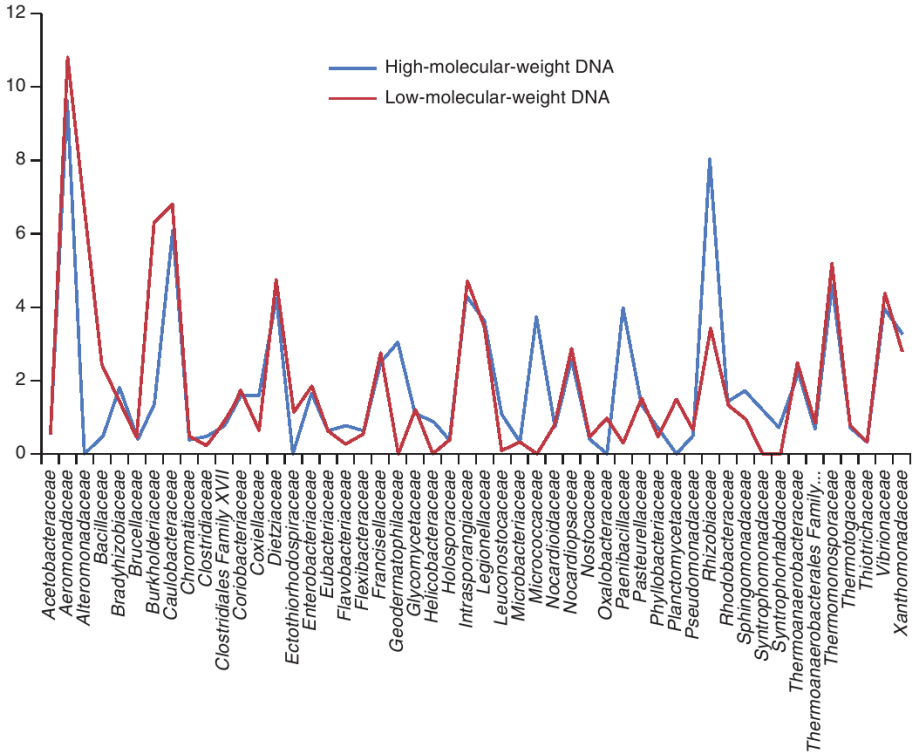


Figure 33.6 Relative proportion of families as a function of the DNA molecular weight. The intensity corresponds to the hybridization fluorescence signals.

Table 33.1 Five Characteristics Corresponding to Seven Metagenomes from Oceans, Soils, and Human Feces

Metagenome	Average of Sequences	Number of Sequences	Number of Base Pairs	Uploading Date	Functions Annotated (%)
Ocean1 (4441573.3)	1012.13	317180	321026307	Tue Oct 21 15:52:12 2008	72.18
Ocean2 (4441574.3)	1007.74	368835	371688861	Tue Oct 21 15:52:47 2008	69.71
Ocean3 (4443713.3)	238.87	217549	51966974	Wed Jun 3 16:11:04 2009	39.32
Soil1 (4445203.3)	533.8	609811	325515132	Mon Oct 26 11:57:26 2009	38.03
Soil2 (4445205.3)	538.93	552268	297632387	Mon Oct 26 12:06:43 2009	32.3
Human feces1 (4440825.3)	190.39	495865	94405318	Fri May 2 13:38:05 2008	25.64
Human feces2 (4440613.3)	335.01	302780	101434082	Thu Feb 14 10:45:23 2008	32.71

9. The databases evolved in time.
10. The percentage of annotated sequences can be highly varied as a function of the metagenomes.

Of course, these 10 variables do not all have the same influence on metagenomic comparisons, but it is difficult to know which are critical and which are not. Moreover, some of these variables are only due to the inconsistency of the metagenomes, and thus they are difficult to avoid. In order to determine if these variables are incompatible with metagenomic comparisons, seven metagenomes were analyzed (MG-RAST online server, http://metagenomics.nmpdr.org/) and the annotated function distribution was compared (E value $<10^{-5}$). As expected, the functional characteristics of the metagenomes are clearly different (**Table 33.1**).

Based on either the number of base pairs, the number of sequences, or the percentage of annotated sequences, three data normalizations can be compared based on metabolic system distributions. A PCA corresponding to the functional frequency of these metagenomes was produced as a function of these normalizations (**Fig. 33.7**). The normalization based on the total number of annotated

Figure 33.7 PCA corresponding to the functional distribution of seven metagenomes sequenced from oceans, soil, and human feces as a function of three normalization possibilities. (**A**) Normalization as a function of the number of base pairs sequenced; (**B**) normalization as a function of the number of sequences; (**C**) normalization as a function of the annotated sequences.

sequences appears to be the most consistent method of the three in order to reduce the impact of variation related to the different sequencing projects.

This successful comparison suggests that global comparisons of different ecosystems are possible and that the different biases are insignificant when compared to the large differences in metabolic systems between these ecosystems. This might not be the case when comparing within ecosystems.

33.8 CONCLUSION

In order to fully assess the biodiversity in one soil, deep sequencing is required. The possible sequencing depth is a function in part of the DNA extraction methods and how rare DNA can be extracted and sequenced. An initiative to sequence the Rothamsted soil was started with a French national project (Metasoil), which will provide 100 Gpb of pyrosequences and two million fosmid (40-kb inserts) clones. By pushing back the frontiers of one soil, other soil sequencing projects will be able to assess the likelihood of novel discoveries with their soils. If the rare biosphere exists in all ecosystems like in soil, then in the future these other environments need to be deeply sequenced to discover new active molecules and for global metagenomic comparisons.

For successful deep sequencing, DNA extraction methods need to be developed in the different environments. In fact, to date, only yield and purity optimization efforts are generally used by researchers [e.g., Lemarchand et al., 2005; Krasova-Wade and Neyra, 2007] to access metagenomic DNA. Other parameters like molecular weight DNA can be optimized for specific projects (e.g., fosmid library construction) [Bertrand et al., 2005]. Nevertheless, results presented in this chapter underline the incapacity of this kind of approach to access complete soil metagenomes. The international community needs to improve diversity recovery optimization strategies and particularly for deep sequencing metagenome projects, such as we have done for Metasoil. This DNA acquisition effort will decrease biases involved in metagenomic approaches and increase our understanding about the 10^{30} prokaryotes present on Earth and improve our rate of discovery of genes of interests.

In addition, it is important to enhance metatranscriptomic projects (see also Chapter 62–65, Vol. I) to study not only the genetic potential of an environment, but also the active genes. In this way, the importance of the different ecosystems (especially soil) for different services (nutrient cycling, climate change buffering, crop production and protection, drug discovery) can be accurately assessed.

Acknowledgments

Work described in this chapter that was conducted in the authors' laboratory has been supported by the French National Research Agency (ANR GMGE Metasoil), and TD was financed by the Rhone–Alpes region.

REFERENCES

Adesina MF, Lembke A, Costa R, Speksnijder A, Smalla K. 2007. Screening of bacterial isolates from various European soils for *in vitro* antagonistic activity towards Rhizoctonia solani and *Fusarium oxysporum*: Site-dependent composition and diversity revealed. *Soil Biol. Biochem*. **39**:2818–2828.

Amann RI, Ludwig W, Schleifer KH. 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev*. **59**:143–169.

Baker-Austin C, Wright MS., Stepanauskas R, McArthur JV. 2006. Co-selection of antibiotic and metal resistance. *Trends. Microbiol*. **14**:176–182.

Berry AE, Chiocchini C, Selby T, Sosio M, Wellington EM. 2003. Isolation of high molecular weight DNA from soil for cloning into BAC vectors. *FEMS Microbiol. Lett*. **223**:15–20.

Bertrand H, Poly F, Van VT, Lombard N, Nalin R, Vogel TM, Simonet P. 2005. High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *J. Microbiol. Methods* **62**:1–11.

Boubakri H, Beuf M, Simonet P, Vogel TM. 2006. Development of metagenomic DNA shuffling for the construction of a xenobiotic gene. *Gene* **375**:87–94.

Courtois S, FrostegAArd A, Goransson P, Depret G, Jeannin P, Simonet P. 2001. Quantification of bacterial subgroups in soil: Comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environ. Microbiol*. **3**:431–439.

Courtois S, Cappellano CM, Ball M, et al., 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl. Environ. Microbiol*. **69**:49–55.

Curtis TP, Sloan WT, Scannell JW. 2002. Estimating prokaryotic diversity and its limits. *Proc. Natl. Acad. Sci. USA* **99**:10494–10499.

Davis KE, Joseph SJ, Janssen PH. 2005. Effects of growth medium, inoculum size, and incubation time on culturability and isolation of soil bacteria. *Appl. Environ. Microbiol*. **71**:826–834.

Dcosta VM, Griffiths E, Wright GD. 2007. Expanding the soil antibiotic resistome: Exploring environmental diversity. *Curr. Opin. Microbiol*. **10**:481–489.

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM. 2011. Accessing the soil metagenome for studies of microbial diversity. *Appl. Environ. Microbiol*. **77**:1315–1324.

Demaneche S, Sanguin H, Poté J, Navarro E, Bernillon D, Mavingui P et al., 2008. Antibiotic-resistant soil bacteria in transgenic plant fields. *Proc. Natl. Acad. Sci. USA* **105**:3957–3962.

Dib J, Motok J, Zenoff VF, Ordonez O, Farias ME. 2008. Occurrence of resistance to antibiotics, UV-B, and arsenic in bacteria isolated from extreme environments in high-altitude [above 4400m] Andean wetlands. *Curr. Microbiol*. **56**:510–517.

Dinsdale, EA, Edwards RA, Hall D et al., 2008. Functional metagenomic comparison profiling of nine biomes. *Nature* **452**: 629–632.

FALKOWSKI PG. 2001. Biogeochemical cycles. *Encyclopedia Biodivers*. **1**:437–453.

GALVAO TC, MOHN WW, DE LORENZO V. 2005. Exploring the microbial biodegradation and biotransformation gene pool. *Trends Biotechnol*. **23**:497–506.

GANS J, WOLINSKY M, DUNBAR J. 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**:1387–1390.

GEBERT J, STRALIS-PAVESE N, ALAWI M, BODROSSY L. 2008. Analysis of methanotrophic communities in landfill biofilters using diagnostic microarray. *Environ. Microbiol*. **10**:1175–1188.

GINOLHAC A, JARRIN C, GILLET B, ROBE P, et al., 2004. Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones. *Appl. Environ. Microbiol*. **70**:5522–5527.

GINOLHAC A, et al., 2005. Type I polyketide synthases may have evolved through horizontal gene transfer. *J. Mol. Evol*. **60**:716–725.

GRIFFITHS RI, WHITELY AS, O'DONNELL AG, BAILEY MJ. 2000. Rapid method for co-extraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Appl. Environ. Microbiol*. **66**:5488–5491.

HANDELSMAN J, RONDON MR, BRADY SF, CLARDY J, GOODMAN RM. 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol*. **5**: R245–R249.

HANDELSMAN J, WACKETT LP. 2002. Ecology and industrial microbiology: Microbial diversity—sustaining the Earth and industry. *Curr. Opin. Microbiol*. **5**:237–239.

HERY M, NAZARET S, JAFFRE T, NORMAND P, NAVARRO E, 2003. Adaptation to nickel spiking of bacterial communities in neocaledonian soils. *Environ. Microbiol*. **5**:3–12.

HUBER JA, MARK WELCH DB, MORRISON HG, HUSE SM, NEAL PR, BUTTERFIELD DA, SOGIN ML. 2007. Microbial population structures in the deep marine biosphere. *Science* **318**:97–100.

HUYGHE A et al. 2008. Novel microarray design strategy to study complex bacterial communities. *Appl. Environ. Microbiol*. **74**:1876–1885.

KNIETCH A, WASCHKOWITZ T, BOWIEN S, HENNE A and DANIEL R. 2003. Metagenomes of complex microbial consortia derived from different soils as sources for nobel genes conferring formation of carbonyls from short-chain polyols on *Echerichia coli*. *J. Microbiol. Biotechnol*. **5**:46–56.

KRASOVA-WADE T, NEYRA M. 2007. Optimization of DNA isolation from legume nodules. *Lett. Appl. Microbiol*. **45**:95–99.

LAROSE C, BERGER S, FERRARI C, NAVARRO E, DOMMERGUE A, SCHNEIDER D, VOGEL TM. 2010. Microbial sequences retrieved from environmental samples from seasonal Arctic snow and meltwater from Svalbard, Norway. *Extremophiles* **14**:205–212.

LEFEVRE F, ROBE P, JARRIN C, GINOLHAC A, ZAGO C, AURIOL D, et al. 2008. Drugs from hidden bugs: their discovery via untapped resources. *Res. Microbiol*. **159**:153–161.

LEMARCHAND K, BERTHIAUME F, MAYNARD C, HAREL J, PAYMENT P, BAYARDELLE P, MASSON L, BROUSSEAU R. 2005. Optimization of microbial DNA extraction and purification from raw wastewater samples for downstream pathogen detection by microarrays. *J. Microbiol. Methods* **63**:115–126.

LEVEAU JHJ. 2007. The magic and menace of metagenomics: Prospects for the study of plant growth-promoting rhizobacteria. *Eur. J. Plant Pathology* **11**:279–300.

MEYER F, PAARMANN D, D'SOUZA M, OLSON R, GLASS EM, et al. 2008. The metagenomics RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes *BMC Bioinformatics* **9**:386.

MORALES SE, COSART TF, JOHNSON JV, HOLBEN WE. 2009. Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters. *Appl. Environn. Microbiol*. **75**:668–675.

ORGAM A, SAYLER GS, and BARBAY T. 1987. The extraction and purification of microbial DNA from sediments. *J. Microb. Methods* **7**:57–66.

ORTÍZ-CASTRO R, CONTRERAS-CORNEJO HA, MACÍAS-RODRÍGUEZ L, LÓPEZ-BUCIO J. 2009. The role of microbial signals in plant growth and development. *Plant Signal Behav*. **4**:701–712.

PASIC L, KOVCE B, SKET B, HERZOG-VELIKONJA B. 2010. Diversity of microbial communities colonizing the walls of a Karstic cave in Slovenia. *FEMS Microbiol. Ecol*. **71**:50–60.

POINTING SB, CHAN Y, LACAP DC, LAU MC, JURGENS JA, FARRELL RL. 2009. Highly specialized microbial diversity in hyper-arid polar desert. *Proc. Natl. Acad. Sci. USA* **106**:19964–19969.

Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME Journal*. **1**:283–290.

RAJENDHRAN J, GUNASEKARAN P. 2008. Strategies for accessing soil metagenome for desired applications. *Biotech. Adv*. **26**:576–590.

RANJARD L, RICHAUME AS. 2001. Quantitative and qualitative microscale distribution of bacteria in soil. *Res. Microbiol*. **152**:707–716.

RICHARD JY, VOGEL TM. 1999. Characterization of a soil bacterial consortium capable of degrading diesel fuel. *Intern. Biodet. Biodegrad*. **44**:93–100.

ROBE P, NALIN R, CAPELLANO C, VOGEL TM, SIMONET P. 2003. Extraction of DNA from soil. *Euc. J. Soil Biol*. **39**:183–190.

ROESCH LL, FULTHORPE RR, RIVA A, CASELLA G, HADWIN AKM, KENT AD, et al., 2007.

SANGUIN H et al., 2006. Development and validation of a prototype 16S rRNA-based taxonomic microarray for Alphaproteobacteria. *Environ. Microbiol*. **8**:289–307.

SHENDURE J, JI H. 2008. Next-generation DNA sequencing. *Nat. Biotech*. **26**:1135–1145.

SILVERTOWN J, et al., 2006. The Park Grass Experiment 1856–2006: Its contribution to ecology. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*. **1**:283–290.

SINGH S, KANG SH, MULCHANDANI A, CHEN W. 2008. Bioremediation: Environmental clean-up through pathway engineering. *Curr. Opin. Biotechnol*. **19**:437–444.

STALEY JT, KONOPKA A. 1995. Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol*. **39**:321–346.

TORSVIK V, GOKSOYR J, DAAE FL. 1990. High diversity in DNA of soil bacteria. *Appl. Environ. Microbiol*. **56**:782–787.

TORSVIK V, OVREAS L, THINGSTAD TF. 2002. Prokaryotic diversity—Magnitude, dynamics, and controlling factors. *Science* **296**:1064–1066.

TRINGE SG, MERING CV, KOBAYASHI A, SALAMOV AA, CHEN K, CHANG HW et al., 2005. Comparative metagenomics of microbial communities. *Science* **308**:554–557.

TURNBAUGH PJ et al., 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**:1027–1031.

VAN ELSAS JD, MANTYNEN V, WOLTERS AC. 1997. Soil DNA extraction and assessment of the fate of *Mycobacterium cholorophenolicum* strain PC-1 in different soils by 16S ribosomal gene sequence based most probable number PCR and immunofluorescence. *Biol. Fertil. Soils* **24**:188–195.

VAN ELSAS JD, JANSSON JK, TREVORS JT. 2006. *Modern Soil Microbiology II*, Boca Raton, FL: CRC Press.

VOGEL TM, SIMONET P, JANSSON JK, HIRSCH PR, TIEDJE JM, VAN ELSAS JD, et al. 2009. TerraGenome: A consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol*. **7**:2.

WAINWRIGHT M, WICKRAMASINGHE NC, NARLIKAR JV, RAJARATNAM P. 2003. Microorganisms cultured from stratospheric air samples obtained at 41km. *FEMS Microbiol Lett*. **218**:161–165.

WARD BB et al. 2007. Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environ. Microbiol*. **9**:2522–2538.

WHITMAN WB, COLEMAN DC, WIEBE WJ. 1998. Prokaryotes: The unseen majority. *Proc. Natl. Acad. Sci. USA* **95**:6578–6583.

WILLNER D, THURBER RV, ROHWER F. 2009. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ. Microbiol*. **11**:1752–1766.

Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Microbiology

Metagenomic exploration of antibiotic resistance in soil

Jean-Michel Monier, Sandrine Demanèche, Tom O Delmont, Alban Mathieu, Timothy M Vogel and Pascal Simonet

The ongoing development of metagenomic approaches is providing the means to explore antibiotic resistance in nature and address questions that could not be answered previously with conventional culture-based strategies. The number of available environmental metagenomic sequence datasets is rapidly expanding and henceforth offer the ability to gain a more comprehensive understanding of antibiotic resistance at the global scale. Although there is now evidence that the environment constitutes a vast reservoir of antibiotic resistance gene determinants (ARGDs) and that the majority of ARGDs acquired by human pathogens may have an environmental origin, a better understanding of their diversity, prevalence and ecological significance may help predict the emergence and spreading of newly acquired resistances. Recent applications of metagenomic approaches to the study of ARGDs in natural environments such as soil should help overcome challenges concerning expanding antibiotic resistances.

Address
Environmental Microbial Genomics Group, Laboratoire AMPERE, UMR CNRS 5005, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 Ecully, France

Corresponding author: Simonet, Pascal (pascal.simonet@ec-lyon.fr)

Current Opinion in Microbiology 2011, **14**:1–7

This review comes from a themed issue on
Ecology and Industrial Microbiology
Edited by Eva Top and David Wilson

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2011.04.010

Introduction

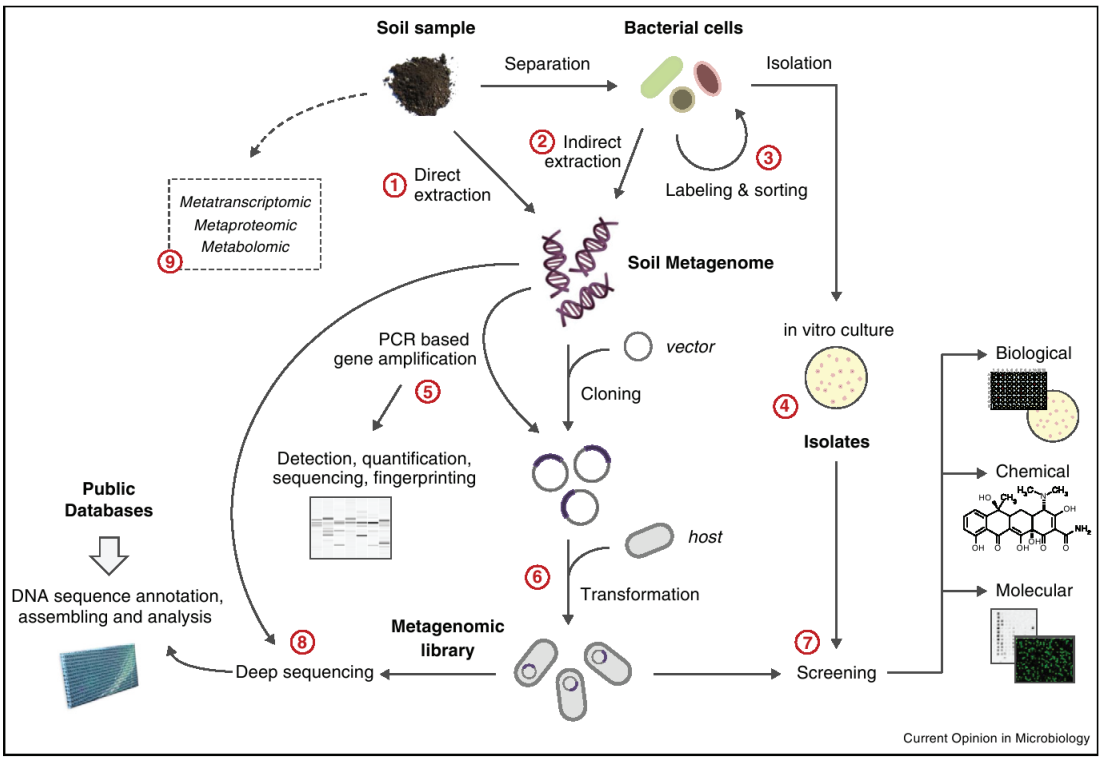
Diseases have over the centuries ravaged the human population without any systematic method for reducing the threat until last century when antibiotic use was put into practice. The discovery and application of antibiotics brought hope for a global eradication of infectious diseases, but this hope rapidly disappeared when antibiotic resistance began to spread in part as a function of antibiotic use increasing concern for the future of antibiotic-based health care. The importance of antibiotic resistance on human health led to studies to determine the origin and dispersion/dissemination of antibiotic resistance gene determinants (ARGDs). In addition, the ecology of ARGDs is critical to understanding the future of antibiotic resistance. For example, the 11 600 different 'bioactive' molecules, as calculated by Wright [1], based on data from Schloss and Handelsmann [2], from one group of bacteria (the *Actinobacteria*) in 1 g of soil would probably overwhelm any current human pathogen. Yet, these soil bacteria resist their own antibiotics [3] and, therefore, contain the ARGDs that could provide a reservoir from which ARGDs could be transferred to other bacteria including human pathogens. The size of this reservoir of ARGDs is, of course, much larger than the example provided and includes those found within the wide expanse of all life on our planet, including the plants, small animals, protists, fungi and other bacteria found in soil. The importance of soil as the arguably largest reservoir of ARGDs cannot be ignored due to the quantity and diversity of the ARGDs already uncovered [4]. Given the intricacy of accessing the entire microbial diversity in such complex environments (e.g. cell extraction yield and cultivability below 1% [5]), the importance – soil has been the source of the majority of ARGDs discovered [4] – magnifies the potential especially when the difficulties to isolate and define the complex and heterogeneous diversity of soil are considered. Some estimates suggest that only a small fraction (about 1%) of the soil diversity has been uncovered [5], and thus, the true size of the soil ARGD reservoir could be a hundred times larger. In addition, this reservoir is subjected to a wide range of processes responsible for increasing functional diversity including bacterial adaptation mechanisms such as gene transfer and genetic rearrangements and genetic and microbial exchanges with other ecosystems including medical facilities. The objective here is to illustrate how metagenomic approaches could help us understand and manage the dispersion and impact of antibiotic resistance genes (or ARGDs) within soil microorganisms and human pathogens.

Ecological significance of ARGD

Although culture-based approaches have revealed that antibiotic resistance was widespread in bacteria [6], over the past few years, the development of metagenomic approaches (Figure 1) has brought new insight into the prevalence and diversity of antibiotic producers and ARGDs in anthropogenically disturbed and natural pristine environments [7••,8]. Such observations are obviously of major concern regarding the dissemination of resistance genes and human health [1,9], but their ubiquity also raises important ecological questions concerning the primary role ARGDs in natural habitats. Why are there so many ARGDs? Why are they maintained in microorganisms in natural habitats? What role do they play? Why are certain resistances more successful than

Please cite this article in press as: Monier J-M, et al. Metagenomic exploration of antibiotic resistance in soil, Curr Opin Microbiol (2011), doi:10.1016/j.mib.2011.04.010

Figure 1



Overview of metagenomic approaches available to study ARGDs in the environment. Metagenomic DNA can be obtained by (1) direct extraction of total DNA from soil or (2) by indirect extraction after recovery of bacterial cells. (3) Before DNA extraction, targeted populations can be selected by labeling and sorting bacterial cells (e.g. FISH and flow cytometry). This additional step allows to perform single cell genome sequencing or to obtain metagenomes with reduced complexity. (4) Bacterial isolates can be obtained on synthetic growth media. While limited to a small fraction of cells present in soil, culture-based approaches offer the ability to study the physiology of the cell as well as to perform conventional molecular or chemical screenings. (5) Soil metagenome can be screened directly for genes of interest (e.g. ARGDs, *rrs*. . .) using PCR-based approaches or/and (6) cloned into vectors used to transform host cells to obtain metagenomic libraries. Targeted metagenomics can be performed by subtractive hybridization or capture of specific DNA fragments. (7) Libraries can be screened for biological activities, characterization of bioactive compounds, detection of genes of interest, although heterologous expression can sometimes be an issue. (8) Next generation sequencing of metagenomic DNA is providing large datasets of DNA sequences; however, annotation is often limited and assembly difficult. (9) Although not addressed in this article, metagenomics are used in combination with complementary approaches such as metatranscriptomics (RNA), metaproteomics (proteins) or metabolomics (metabolites).

others in specific habitats? Metagenomic tools and next-generation deep sequencing should prove useful in assessing the extent of ARGDs and the ecological roles of antibiotics in the environment.

Although there are few examples and evidence of the weapon/shield function of antibiotics and ARGDs in nature [10,11], our anthropogenic vision of antibiotic and ARGDs in the clinical environment may not be transposable to the natural environment. Therapeutic concentrations of antibiotics in the clinical environment are most likely higher than those encountered in the natural environment and antibiotics may exhibit distinct and different functions at sub-inhibitory concentrations. Although their 'historical' functions may provide a partial explanation for their ubiquity, it is now largely accepted that the primary role of antibiotics and ARGDs in nature may be completely different.

Phylogenetically diverse bacteria isolated from different soils were shown to subsist on several antibiotics as the sole source of carbon [12]. Furthermore, each isolate was resistant to several antibiotics at therapeutic concentrations. Although the selective advantages and ecological fitness of antibiotic use for growth substrates are unclear, the microbes subsisting on antibiotics might be a source of ARGDs for pathogens [12].

Please cite this article in press as: Monier J-M, et al. Metagenomic exploration of antibiotic resistance in soil, Curr Opin Microbiol (2011), doi:10.1016/j.mib.2011.04.010

Another major discovery in recent years is the hormetic effect of antibiotics, where transcription of bacterial genes is antibiotic dose-dependent. Their role as signaling agents at sub-inhibitory concentrations in the environment is based on observed transcription profiles [13,14]. Although most studies demonstrating the role of antibiotics as signals were performed with bacterial isolates, in a recent study, antibiotics and other pharmaceutical products were shown to induce transcriptional changes in a complex microbial community at low concentrations [15•]. While signal receptors are often the same as those targeted at therapeutic concentrations, this is not always the case, and the nature and relevance of the genes either induced or repressed by bioactive compounds at sub-inhibitory concentrations in complex communities still need to be addressed. Some bacteria respond to the presence of antibiotics by increasing their mutation frequency, where antibiotics act as an evolutionary force for resistance mechanisms [16–19].

Although there is less direct evidence of the ecological role of ARGDs in nature besides their ability to confer resistance to a wide range of toxic compounds, it has been reported that ARGDs could be implicated in important cell processes, such as homeostasis, detoxification, virulence, regulation of biosynthesis pathways [20••,21] or growth and survival [22]. Inversely, the number of genes involved in antibiotic resistance might be underestimated as several studies based on single gene-knockouts reported the unexpected implication of a significant percentage of genes in the resistance to antibiotics [23–25]. Still, the weapon/shield vision of antibiotics and ARGDs in the environment might be insignificant as illustrated by an elegant study using a differential assay for compounds that select against antibiotic resistance showing that selection-inverting compounds are secreted by soil microbes as part of their repertory of chemicals that counteract antibiotic resistance [26,27•]. The increasing amount of publically available genomic and metagenomic sequences will help us address the extent and role of ARGDs in nature; however, understanding the full extent of the environmental resistome and its role in shaping microbial communities still remains challenging. Although the role of ARGDs as receptors of chemical signals or actors in the regulation of cell processes is still hypothetical, an integrated approach combining metagenomic sequence analyses, functional screenings and spatial distribution of antibiotic producers and neighboring organisms in natural habitats should be considered when addressing the ecological relevance of ARGDs.

From functional screening to high-throughput sequencing: how metagenomic approaches can help?

Metagenomic approaches (Figure 1), although relatively recent, have proven to be of considerable help in the discovery of new antibiotics [4,28], tracking ARGDs across the planet and in exploring the dissemination of ARGDs. To investigate ARGDs among uncultured bacteria in an undisturbed soil environment, a functional analysis conducted on a remote Alaskan soil showed that this soil was a reservoir for ARGDs, thus implying that even in the absence of selective pressure imposed by anthropogenic activity, the soil microbial community harbors unique and ancient resistance determinants [7••]. The screening of 13 Gb of apple orchard soil metagenomic DNA led to the detection of two new genes for seftazidime and kanamycine resistance [8]. A new chloramphenicol-florfenicol resistant gene was discovered by screening an Alaska soil metagenomic DNA clone library [29]. This phenotypic screening methodology is a powerful approach in the discovery of novel genes, but suffers from the disadvantage that gene expression is necessary for selection during screening. In order to overcome this limitation, the genetic exploration of the metagenomic sequence data can lead to the recovery of an additional ARGD diversity from uncultivated microbiota and shows that some of the environmental sequences are interspersed with genes from commensal and pathogenic bacteria [30,31].

Quantification of genes by quantitative PCR is another powerful and sensitive tool to compare the distribution of antibiotic resistant genes in different environments and to collect information about their representation in different biotopes [32,33]. Recently, up to 10^{11} copies of sulfonamides resistance genes (respectively *sul1* and *sul2*) per gram of sediment of the Haihe River were measured [34]. The highest concentrations of *sul1* and *sul2* were correlated to the sulfonamide concentration, and gene copy numbers in the water column were related to the concentration of resistance genes in the sediment. Clearly, the quantity of resistance genes is not fixed, and fluxes exist with some environments being connected and interdependent.

Previously discovered genes can also be used as probes to hybridize metagenomic DNA clones under low stringency conditions in order to detect novel ARGDs. In addition to information of the ARGD sequence itself and its divergence level with the probe, the complete sequencing of positive clone inserts provides invaluable data on its genetic environment, including the presence of other ARGDs or sequences susceptible to promote its transfer to other bacteria [35].

High-throughput sequencing of metagenomic DNA offers the advantage of random sequencing. So despite some biases [36], estimates of ARGD abundance in various environments can be determined. Thousands of metagenomes have already been generated, and hundreds of them are available to the international community (e.g. http://metagenomics.nmpdr.org/; http://img.jgi.doe.gov/cgi-bin/m/main.cgi). These data can be

Please cite this article in press as: Monier J-M, et al. Metagenomic exploration of antibiotic resistance in soil, Curr Opin Microbiol (2011), doi:10.1016/j.mib.2011.04.010

used to test various hypotheses concerning ARGDs. As an example, available metagenomes corresponding to several ocean, soil and human feces microbial populations were annotated (E-value $<10^{-5}$) and classed into different functional subsystems using MG RAST [37] (Figure 2). Some of these functions correspond to antibiotic resistance or biosynthesis; others correspond to mobile genetic elements that can potentially disseminate ARGDs among different species and environments (e.g. integrons) [38]. The sequences related to tetracycline and fluoroquinolone resistance genes are detected more frequently in human microbial populations than in soil. In contrast, sequences related to the biosynthesis of pyoverdine are found more frequently in soil DNA sequences than in ocean and human feces metagenomic data. Finally, the distribution of sequences related to the *intI* gene from integrons was quantified in these metagenomes and this mobile genetic element appears to be more common in soil and humans than in ocean microbial populations, where a significant part of the predominant microorganisms exist as unattached cells suspended in the water column [39]. A planktonic life style most likely limits contacts between cells, and thus, the efficiency of mobile genetic elements that are absent, in *Pelagibacter ubique*, for example [40]. Metagenomic comparisons provide information about the prevalence of species of interest, ARGDs and mobile genetic elements in different environments, and thus, could stimulate the study of ARGDs.

Lost in translation

The development and application of metagenomic approaches are of major interest and show great promise for the study of ARGDs in the natural environment because the often critical culturing steps are avoided ('domesticated' microbes represent only about 1% of the total diversity). Environmental metagenomic sequence datasets are accumulating at an impressive pace and next generation sequencing technology will accelerate that pace even further [41–44]; however, these datasets still represent a minute fraction of the actual diversity. Although minimum sequencing was required in the pioneering metagenomic work of Tyson and collaborators [45], who characterized an acid mine drainage community containing few species, such a limited approach would not be applicable to more complex and diverse microbial communities, such as those present in soils. To address relevant ecological questions and to assess the prevalence and diversity of ARGDs in the environment, innovative approaches will have to be implemented. Major challenges ahead of us are related to the necessity to increase the resolution of metagenomic data by either accumulating more data or generating specialized/enriched metagenomic libraries. The later could avoid extensive and often laborious functional screening of large metagenomic libraries; however, functional screening of metagenomic libraries still remains a major step since vectors developed are limited [46,47] and heterologous expression is often required.

Another challenge of major relevance when studying the acquisition and dissemination of ARGDs is the discovery of the links between ARGDs, their genetic environment, and their host. Promising approaches (e.g. subtractive hybridization) [48], PCR-DGGE combined to metagenome walking [49], *in situ* rolling circle amplification [50] and FISH combined with cell sorting [51] have been developed to address such questions, but to the best of our knowledge, they have not yet been applied to ARGDs in soil habitats.

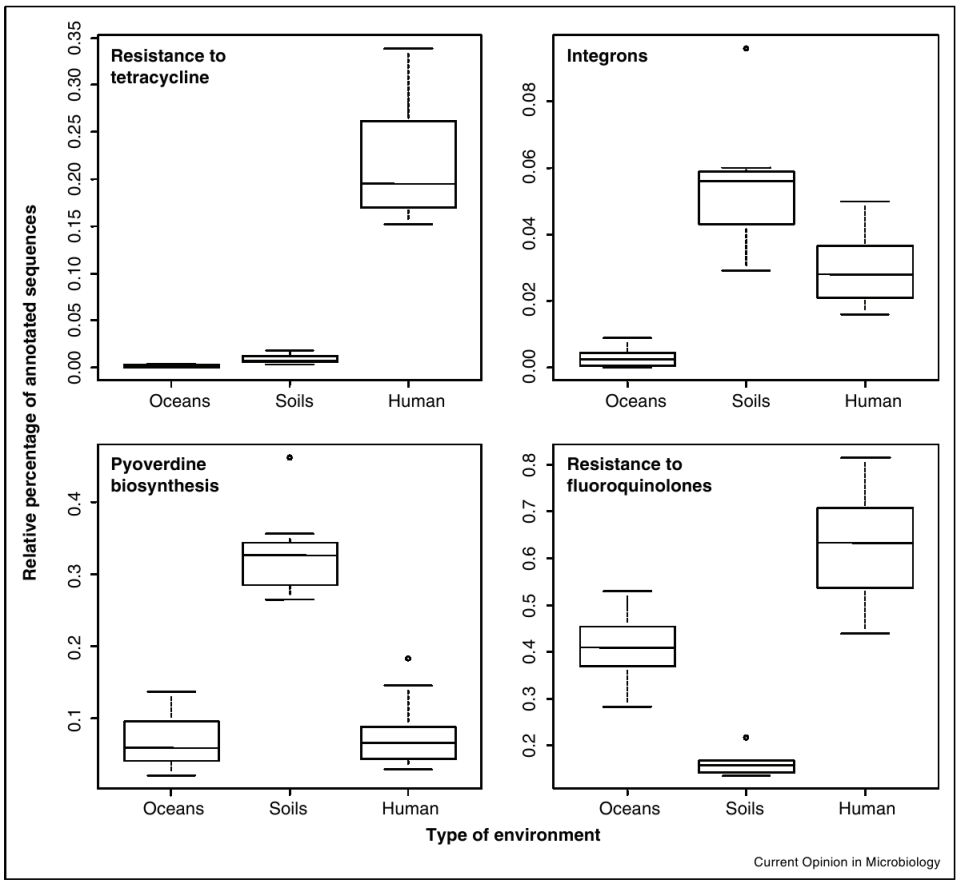
ARGDs have to be tracked in the environment to understand their diversity, functions, and dissemination capacities; however, the length of generated sequences is limited and varies from 100 to 800 pb as a function of the sequencing technology used. In addition, due to the considerable biodiversity in a majority of environments (e.g. in soil), metagenomes are difficult to assemble. As a consequence, only fragments of genes can be analyzed when an environmental metagenome is generated and it is difficult to study the genetic environment of detected ARGDs (e.g. mobile genetic elements or other ARDG) and the phylogeny of species possessing these functions (pathogenic or not). Technological limitations prevent the accurate identification of ARGDs in the environment and need to be improved (especially the length and number of generated sequences). Finally, the challenge is to assemble quickly and at acceptable costs the individual genomes in metagenomes. *In silico* data will then provide considerable information about ARGD location in both environmental and clinical microorganisms.

Future prospects

Metagenomic approaches have the potential to play a crucial role in our understanding of ARDGs at the global scale. To stimulate the study of ARGDs using these tools, additional environments need to be sequenced and deposited in public databases. This sequencing effort across the planet coupled to global metagenomic comparisons will provide information about the prevalence of specific ARGDs in different ecosystems (e.g. Figure 2). These comparisons will provide information about ecosystems with unusually high concentrations of ARGDs. These ecosystems could be further explored via clone libraries and functional screening. Important sequencing projects from specific environments (e.g. Terragenome project for the Rothamsted soil [52]) will also help their study (especially their genetic environment) by stimulating the assembly of complex metagenomes; however, when considering the genetic diversity present in soil (up to 4×10^{13} bp/g, the equivalent of 10^9 pyrosequencing runs and 2.7×10^6 HiSeq complete runs) and the difficulty to access a soil metagenome using classical DNA extraction approaches [53], actual strategies and

Please cite this article in press as: Monier J-M, et al. Metagenomic exploration of antibiotic resistance in soil, Curr Opin Microbiol (2011), doi:10.1016/j.mib.2011.04.010

Figure 2



Relative distribution of four functional subsystems among 32 metagenomes grouped as a function of three environments. Box-plots were obtained using 12, 9 and 11 metagenomes for ocean, soil and human feces, respectively (MG-RAST annotation, *E*-value = 10^{-5}). Box plots represent interquartile ranges (upper, median and lower quartile) and the minimum and maximum values which can be considered as outliers (abnormal distance). The 25 metagenomes used for ocean and human feces are available at http://metagenomics.nmpdr.org [29] (accession numbers for Oceans: 4441573.3, 4441574.3, 4441576.3, 4441577.3, 4441591.3, 4443688.3, 4443697.3, 4443713.3, 4443714.3, 4443716.3, 4443725.3, 4443729.3 and for Human feces: 4440825.3, 4440460.5, 4440614.3, 4440611.3, 4440613.3, 4440616.3, 4440595.4, 4440452.7, 4440939.3, 4440942.3, 4440943.3). The nine metagenomes used for the analysis are available at http://metagenomics.nmpdr.org (accession numbers: 4441091.3, 4446153.3) and at http://metasoil.univ-lyon1.fr (Rothamsted Park Grass soil).

technologies to study prokaryote communities from this environment are limited and need to be improved. But the assembly of uncultured soil microorganisms using a massive sequencing approach like was recently done with the cow rumen [54] will stimulate the study of ARGDs in natural and highly biodiverse environments.

Although waiting for better sequencing technologies, additional screening of clone libraries needs to be initiated to help study the genetic environment of ARGDs in natural ecosystems and to discover new bioactive compounds. With this in mind, we created a two million fosmid library using DNA extracted from the Rothamsted soil under both natural conditions and from microcosms to study the usually inaccessible DNA. This library is accessible to the international community (contact the Libragen Company for more details) for further screening of the equivalent of 20 000 bacterial genomes.

Finally, after studying ARGD structures, mechanisms, diversities and roles for microorganisms, strategies need to be tested to counter or at least limit antibiotic resistance emergence and dissemination in hospitals and human-associated microbial populations.

Please cite this article in press as: Monier J-M, et al. Metagenomic exploration of antibiotic resistance in soil, Curr Opin Microbiol (2011), doi:10.1016/j.mib.2011.04.010

Acknowledgements

This work was financed partly by ADEME (project 'Generique', convention de financement no. 0975C0007), Ministère de l'Ecologie (project 'Septante pro' convention no. 2010-0027) and CNRS-INEE (contract INEE 10-444).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Wright GD: **Antibiotic resistance in the environment: a link to the clinic?** *Curr Opin Microbiol* 2010, **13**:589-594.

2. Schloss PD, Handelsman J: **Toward a census of bacteria in soil**. *PLoS Comput Biol* 2006, **2**:e92.

3. Hopwood DA: **How do antibiotic-producing bacteria ensure their self-resistance before antibiotic biosynthesis incapacitates them?** *Mol Microbiol* 2007, **63**:937-940.

4. Lefevre F, Robe P, Jarrin C, Ginolhac A, Zago C, Auriol D, Vogel TM, Simonet P, Nalin R: **Drugs from hidden bugs: their discovery via untapped resources**. *Res Microbiol* 2008, **159**:153-161.

5. Amann RI, Ludwig W, Schleifer KH: **Phylogenetic identification and in situ detection of individual microbial cells without cultivation**. *Microbiol Rev* 1995, **59**:143-169.

6. Dcosta VM, McGrann KM, Hughes DW, Wright GD: **Sampling the antibiotic resistome**. *Science* 2006, **311**:374-377.

7. •• Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J: **Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil**. *ISME J* 2009, **3**:243-251.

Investigation of antibiotic resistance in a pristine Alaskan soil using a functional metagenomic approach showing that an unpolluted environment can constitute a reservoir of unique and ancient antibiotic resistance gene determinants.

8. Donato JJ, Moe LA, Converse BJ, Smart KD, Berklein FC, McManus PS, Handelsman J: **Metagenomics reveals antibiotic resistance genes encoding predicted bifunctional proteins in apple orchard soil**. *Appl Environ Microbiol* 2010, **76**:4396-4401.

9. Wright GD: **The antibiotic resistome: the nexus of chemical and genetic diversity**. *Nat Rev Microbiol* 2007, **5**:175-186.

10. Currie CR, Scott JA, Summerbell RC, Malloch D: **Fungus-growing ants use antibioticproducing bacteria to control garden parasites**. *Nature* 1999, **398**:701-704.

11. Neeno EC, Kinkel LL, Schottel JL: **Competition and antibiosis in the biological control of potato scab**. *Can J Microbiol* 2001, **47**:332-340.

12. Dantas G, Sommer MO, Oluwasegun RD, Church GM: **Bacteria subsisting on antibiotics**. *Science* 2008, **320**:100-103.

13. Davies J, Spiegelman GB, Yim G: **The world of subinhibitory antibiotic concentrations**. *Curr Opin Microbiol* 2006, **9**:445-453.

14. Linares JF, Gustafsson I, Baquero F, Martinez JL: **Antibiotics as intermicrobial signaling agents instead of weapons**. *Proc Natl Acad Sci U S A* 2006, **103**:19484-19489.

15. • Yergeau E, Lawrence JR, Waiser MJ, Korber DR, Greer CW: **Meta-transcriptomic analysis of the response of river biofilms to pharmaceutical products using anonymous DNA microarrays**. *Appl Environ Microbiol* 2010, **76**:5432-5439.

Evidence of the impact of pharmaceutical products at low concentrations on the induction of transcriptional responses in complex microbial communities in aquatic environments.

16. Davies J, Davies D: **Origins and evolution of antibiotic resistance**. *Microbiol Mol Biol Rev* 2010, **74**:417-433.

17. Martinez JL, Baquero F: **Mutation frequencies and antibiotic resistance**. *Antimicrob Agents Chemother* 2000, **44**:1771-1777.

18. Roth JR: **The joys and terrors of fast adaptation: new findings elucidate antibiotic resistance and natural selection**. *Mol Microbiol* 2011, **79**:279-282.

19. Wang P, Zhang XN, Wang L, Zhen Z, Tang ML, Li JB: **Subinhibitory concentrations of ciprofloxacin induce SOS response and mutations of antibiotic resistance in bacteria**. *Ann Microbiol* 2010, **60**:511-517.

20. •• Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J: **Call of the wild: antibiotic resistance genes in natural environments**. *Nat Rev Microbiol* 2010, **8**:251-259.

An excellent review on the presence and dissemination of antibiotic resistance genes in natural environments highlighting how little is known about the ecology of antibiotic resistance genes in nature and its implication to human health.

21. Martinez JL, Sanchez MB, Martinez-Solano L, Hernandez A, Garmendia L, Fajardo A, Alvarez-Ortega C: **Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems**. *FEMS Microbiol Rev* 2009, **33**:430-439.

22. Groh JL, Luo Q, Ballard JD, Krumholz LR: **Genes that enhance the ecological fitness of *Shewanella oneidensis* MR-1 in sediments reveal the value of antibiotic resistance**. *Appl Environ Microbiol* 2007, **73**:492-498.

23. Breidenstein EB, Khaira BK, Wiegand I, Overhage J, Hancock RE: **Complex ciprofloxacin resistome revealed by screening a *Pseudomonas aeruginosa* mutant library for altered susceptibility**. *Antimicrob Agents Chemother* 2008, **52**:4486-4491.

24. Gomez MJ, Neyfakh AA: **Genes involved in intrinsic antibiotic resistance of *Acinetobacter baylyi***. *Antimicrob Agents Chemother* 2006, **50**:3562-3567.

25. Tamae C, Liu A, Kim K, Sitz D, Hong J, Becket E, Bui A, Solaimani P, Tran KP, Yang H *et al.*: **Determination of antibiotic hypersensitivity among 4,000 single-gene-knockout mutants of *Escherichia coli***. *J Bacteriol* 2008, **190**:5981-5988.

26. Chait R, Craney A, Kishony R: **Antibiotic interactions that select against resistance**. *Nature* 2007, **446**:668-671.

27. • Chait R, Shrestha S, Shah AK, Michel JB, Kishony R: **A differential drug screen for compounds that select against antibiotic resistance**. *PLoS ONE* 2010, **5**:e15179.

An elegant approach with soil microbes to identify secreted compounds selecting against specific resistance mechanisms or whose resistance is based on their physiological interaction with antibiotics.

28. Van Elsas JD, Costa R, Jansson J, Sjoling S, Bailey M, Nalin R, Vogel TM, Van Overbeek L: **The metagenomics of disease-suppressive soils – experiences from the METACONTROL project**. *Trends Biotechnol* 2009, **26**:591-601.

29. Lang KS, Anderson JM, Schwarz S, Williamson L, Handelsman J, Singer RS: **Novel florfenicol and chloramphenicol resistance gene discovered in Alaskan soil by using functional metagenomics**. *Appl Environ Microbiol* 2010, **76**:5321-5326.

30. Aminov RI, MacKie RI: **Evolution and ecology of antibiotic resistance genes**. *FEMS Microbiol Lett* 2007, **271**:147-161.

31. Demanèche S, Sanguin H, Poté J, Navarro E, Bernillon D, Mavingui P, Wildi W, Vogel TM, Simonet P: **Antibiotic resistant soil bacteria in transgenic plant fields**. *Proc Natl Acad Sci U S A* 2008, **105**:3957-3962.

32. Heuer H, Solehati Q, Zimmerling U, Kleineidam K, Schloter M, Müller T, Focks A, Thiele-Bruhn S, Smalla K: **Accumulation of sulfonamide resistance genes in arable soils due to repeated application of manure containing sulfadiazine**. *Appl Environ Microbiol* 2011, **77**:2527-2530.

33. Tamminen M, Karkman A, Lõhmus A, Muziasari WI, Takasu H, Wada S, Suzuki S, Virta M: **Tetracycline resistance genes persist at aquaculture farms in the absence of selection pressure**. *Environ Sci Technol* 2011, **45**:386-391.

34. Luo Y, Mao D, Rysz M, Zhou Q, Zhang H, Xu L, Alvarez JJP: **Trends in antibiotic resistance genes occurrence in the Haihe River, China**. *Environ Sci Technol* 2010, **44**:7220-7225.

35. Demanèche S, David MM, Navarro E, Simonet P, Vogel TM: **Evaluation of functional gene enrichment in a soil metagenomic clone library**. *J Microbiol Methods* 2009, **76**:105-107.

Please cite this article in press as: Monier J-M, et al. Metagenomic exploration of antibiotic resistance in soil, Curr Opin Microbiol (2011), doi:10.1016/j.mib.2011.04.010

36. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing**. *Nucleic Acids Res* 2008, **36**:e105.

37. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A *et al.*: **The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes**. *BMC Bioinform* 2008, **9**:386.

38. Gillings MR, Krishnan S, Worden PJ, Hardwick SA: **Recovery of diverse genes for class 1 integron-integrases from environmental DNA samples**. *FEMS Microbiol Lett* 2008, **287**:56-62.

39. Morris RM, Rappé MS, Connon SA, Vergin KL, Siebold WA, Carlson CA, Giovannoni SJ: **SAR11 clade dominates ocean surface bacterioplankton communities**. *Nature* 2002, **420**:806-810.

40. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M *et al.*: **Genome streamlining in a cosmopolitan oceanic bacterium**. *Science* 2005, **19**:1242-1245.

41. Cardenas E, Tiedje JM: **New tools for discovering and characterizing microbial diversity**. *Curr Opin Biotechnol* 2008, **19**:544-549.

42. Chistoserdova L: **Recent progress and new challenges in metagenomics for biotechnology**. *Biotechnol Lett* 2010, **32**:1351-1359.

43. Fox S, Filichkin S, Mockler TC: **Applications of ultra-high-throughput sequencing**. *Methods Mol Biol* 2009, **553**:79-108.

44. Nagarajan N, Pop M: **Sequencing and genome assembly using next-generation technologies**. *Methods Mol Biol* 2010, **673**:1-17.

45. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF *et al.*: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**:37-43.

46. Craig JW, Chang FY, Kim JH, Obiajulu SC, Brady SF: **Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse Proteobacteria**. *Appl Environ Microbiol* 2010, **76**:1633-1641.

47. Komatsu M, Uchiyama T, Omura S, Cane DE, Ikeda H: **Genome-minimized *Streptomyces* host for the heterologous expression of secondary metabolism**. *Proc Natl Acad Sci U S A* 2010, **107**:2646-2651.

48. Chew YV, Holmes AJ: **Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest**. *J Microbiol Methods* 2009, **78**:136-143.

49. Morimoto S, Fujii T: **A new approach to retrieve full lengths of functional genes from soil by PCR-DGGE and metagenome walking**. *Appl Microbiol Biotechnol* 2009, **83**:389-396.

50. Hoshino T, Schramm A: **Detection of denitrification genes by in situ rolling circle amplification-fluorescence in situ hybridization to link metabolic potential with identity inside bacterial cells**. *Environ Microbiol* 2010, **12**:2508-2517.

51. Kalyuzhnaya MG, Zabinsky R, Bowerman S, Baker DR, Lidstrom ME, Chistoserdova L: **Fluorescence in situ hybridization-flow cytometry-cell sorting-based method for separation and enrichment of type I and type II methanotroph populations**. *Appl Environ Microbiol* 2006, **72**:4293-4301.

52. Vogel TM, Simonet P, Jansson JK, Hirsh PR, Tiedje JM, Van Elsas JD, Bailey MJ, Nalin R, Philippot L: **TerraGenome: a consortium for the sequencing of a soil metagenome**. *Nat Rev Microbiol* 2009, **7**:252.

53. Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P, Hirsch PR, Vogel TM: **Accessing the soil metagenome for studies of microbial diversity**. *Appl Environ Microbiol* 2011, **77**:1315-1324.

54. Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T *et al.*: **Metagenomic discovery of biomass-degrading genes and genomes from cow rumen**. *Nature* 2011, **331**:463-467.

Please cite this article in press as: Monier J-M, et al. Metagenomic exploration of antibiotic resistance in soil, Curr Opin Microbiol (2011), doi:10.1016/j.mib.2011.04.010

Generation of high and low GC content soil metagenomes provides access to distinct genetic diversities.

Tom O Delmont, William E. Holben, and Timothy M. Vogel

Abstract: In complex environments like soils, metagenomes represent a wide range of genetic structures corresponding to different phyla. However, in some cases only one specific part of the global diversity can be studied to discover genes of interest and to stimulate assembly efficiency. Some phyla are known to possess a relatively low GC content (*e.g.*, Acidobacteria) while others possess a high GC content (*e.g.*, Actinobacteria). In order to generate metagenomes from the same soil environment, but representing different phylum distributions, 500 micrograms of purified DNA were fractionated as a function of the GC content. Six fractions were defined and sequenced using the Titanium pyrosequencing technology (one complete run per fraction). Results demonstrated the separation power of this approach (*e.g.*, Actinobacteria distribution varied from 7.47% to 42.60% between the different fractions). The distribution of functions varied as well, so it is possible to correlate functions and taxa. Unexpectedly, the subsystem related to secondary metabolism known to be highly present among Actinobacteria did not increase in the high GC content metagenome fractions. In addition, subsystems related to cell walls and capsules were negatively correlated to this phylum. Finally, assembly efforts were unable to produce long contigs. Thus, this strategy accessed different microbial distributions, but was too complex for genome assembly efforts.

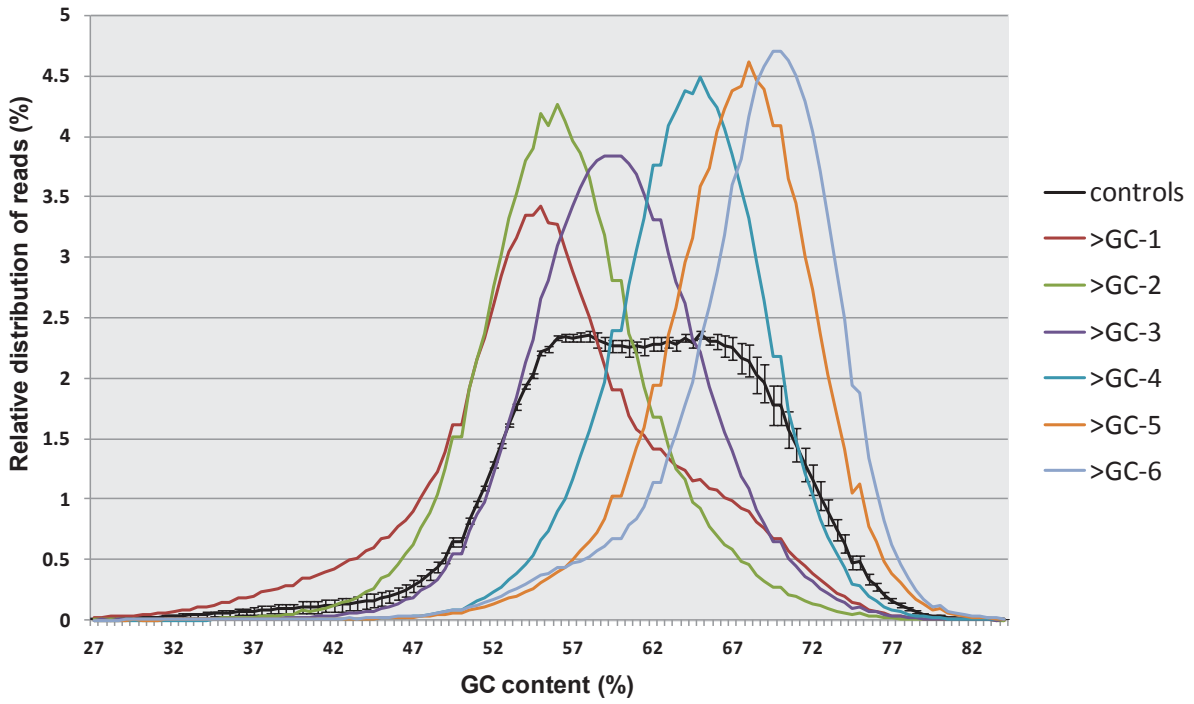


Figure 1: Relative distribution of reads (%) as a function of the GC content in the six metagenomes generated after the DNA separation and in four metagenomes corresponding to natural communities (controls).

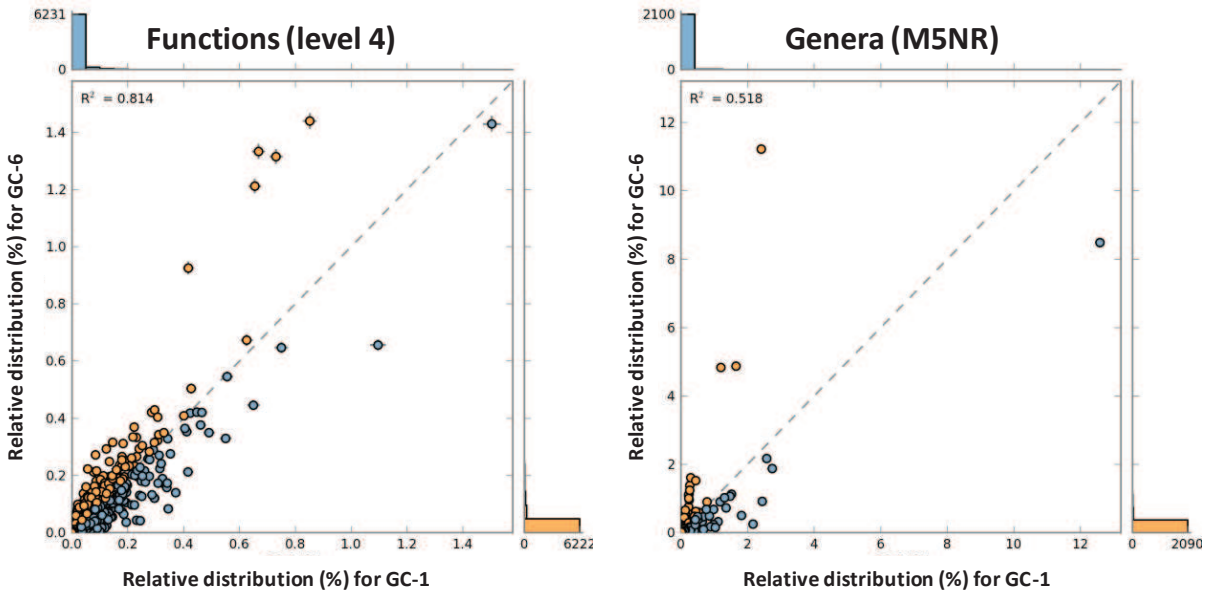


Figure 2: Relative distribution (%) of functions and genera between metagenomes corresponding to the lowest and highest GC content separation.

	GC-1	GC-2	GC-3	GC-4	GC-5	GC-6
GC CONTENT AVERAGE (%)	55.00	57.00	60.00	64.00	68.00	70.00
Taxo (%)						
Acidobacteria	5.80	7.51	6.62	3.45	2.04	1.86
Actinobacteria	10.82	7.47	8.76	17.52	36.74	42.60
Ascomycota	1.48	0.67	0.44	0.34	0.27	0.35
Bacteroidetes	5.75	3.45	2.22	1.52	1.34	1.26
Chloroflexi	1.77	1.67	1.33	1.40	1.86	1.94
Euryarchaeota	0.79	0.74	0.60	0.59	0.78	0.81
Proteobacteria	36.68	39.51	50.10	50.89	34.60	29.36
Verrucomicrobia	6.35	7.84	3.03	1.27	0.84	0.86
Functions (%)						
Cell Wall and Capsule	4.37	4.14	3.64	3.52	3.35	3.41
Fatty Acids, Lipids, and Isoprenoids	3.61	3.51	3.94	4.31	4.64	4.64
RNA Metabolism	4.86	4.83	4.35	4.25	4.30	4.37
Secondary Metabolism	0.36	0.48	0.51	0.42	0.37	0.38
Virulence, Disease and Defense	3.22	3.82	3.63	3.08	2.34	2.15

Table: Relative distribution (%) of specific phyla and general functional subsystems among metagenomes corresponding to different GC contents.

Metagenomes extracted from dry soil samples archived for decades provide access to highly unusual nucleic diversities.

Tom O Delmont, Ian Clark, Eric Pelletier, Denis LePaslier, Pascal Simonet, Penny Hirsch and Timothy M. Vogel

Abstract: Rothamsted is the oldest experimental station in the world and started archiving soil samples 160 years ago. These samples are of interest to study the changes in agricultural practices on soil microbial communitoes. In addition, old samples could also be used to study the effect of relatively long term climate fluctuations on the structure of microbial communities. Thus, metagenomes corresponding to soil samples stored in 1876, 1923, 1959, 1991, 2002, and 2008 were generated to study the variation of the soil microbial community of the Rothamsted Research station Parkgrass control soil as a function of time. The DNA diversity extracted from soil stored for 135 years is statistically different from those corresponding to fresh soils. These observations (*e.g.*, predominance of Proteobacteria) cannot be explained simply by environmental variations and might represent both a microbial adaptation to dry storage or a long term storage effect on DNA. The archived samples cannot be used to study the effect of environmental parameters. However, the corresponding communities are unusual and provided access to a subset of the original soil metagenome.

Remark:

Due to the low reproducibility of generated datasets from the two samples stored in 1876, we decided to study additional plots stored the same year. Results are ongoing.

Thus only two preliminary figures are presented at this time to show principal trends.

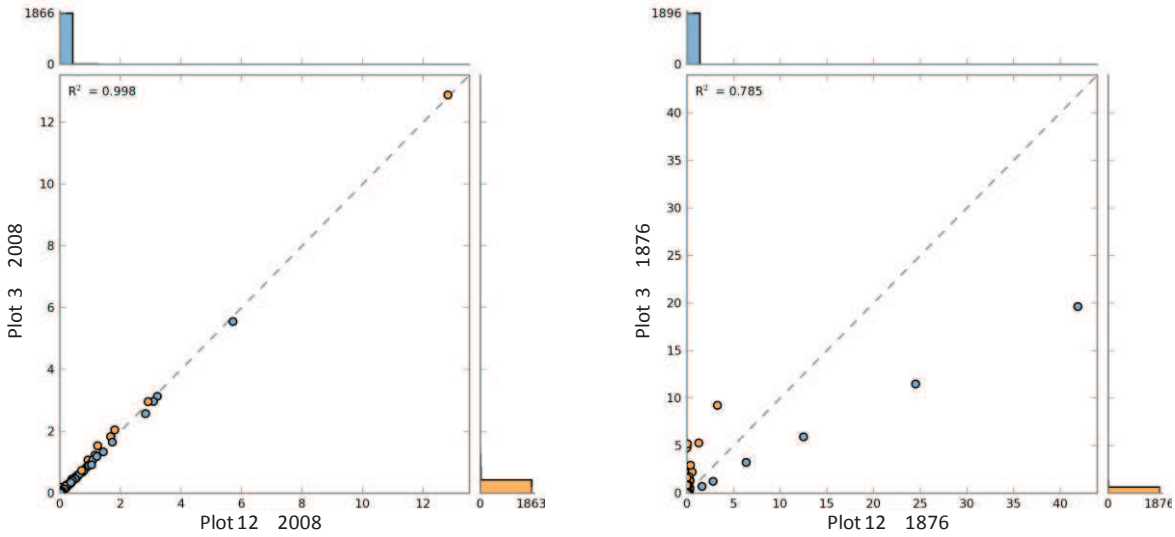


Figure 1: Reproducibility of the dry storage on the structure of soil microbial communities. Graphs are based on the relative distribution of genera annotated on MG-RAST (E-value cut-off of 10^{-5}) and exported to the STAMP software. There are no replicates on these two graphs.

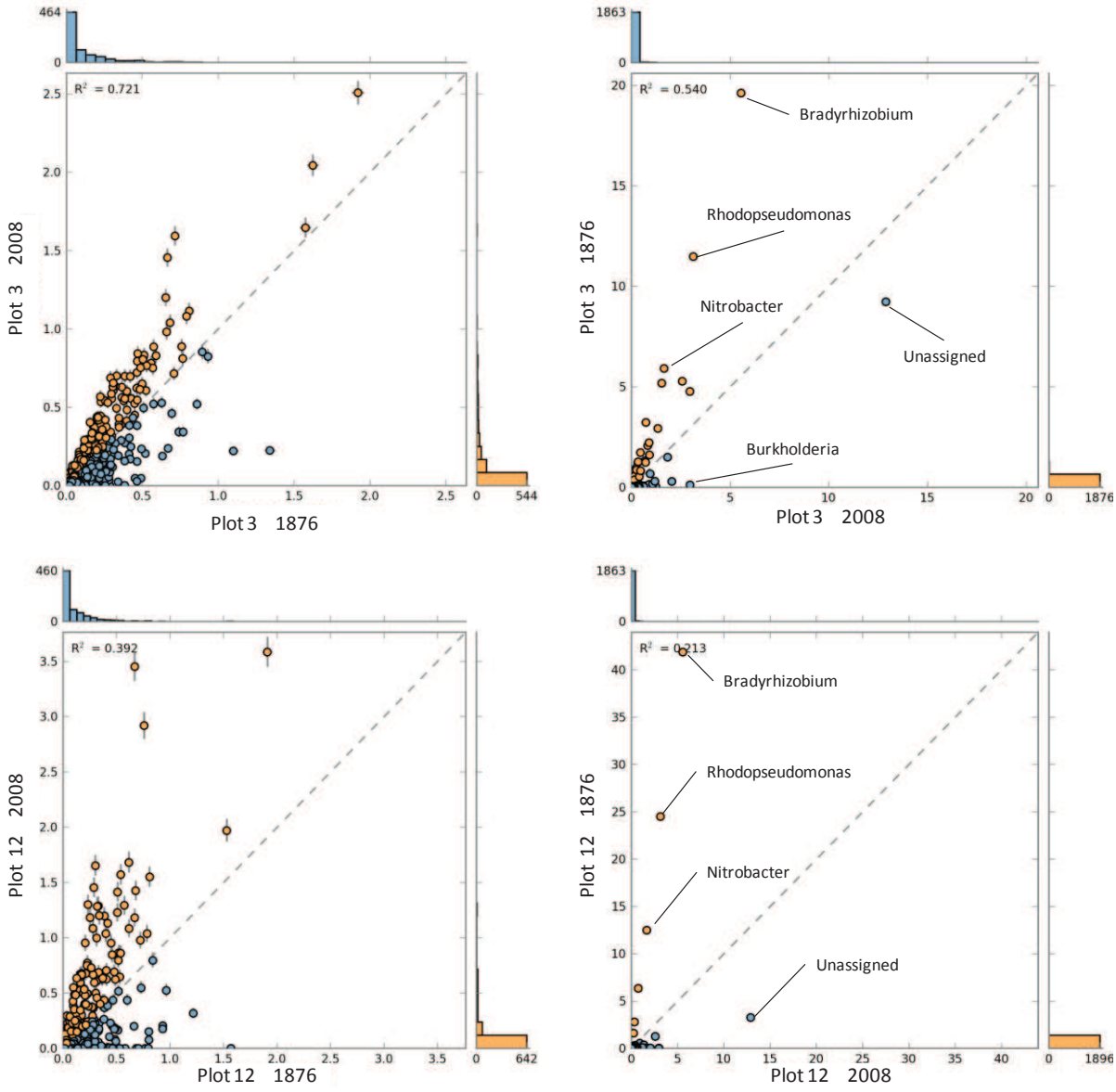


Figure 2: Comparison of samples stored in 2008 and 1876 from two distinct Rothamsted plots. Graphs represent both functional (level 3) and taxonomical (genera level) differences.

Scripts for bioinformatics analyses:

HSPextraction.py
programm that extracts bank seq ids from a blast result default flat file. Typical use : ./HSPextraction.blast file.blast ncbi use 454 instead of ncbi if the blastdb has been formated from a 454 generated fasta file.

SubsetQual.py
program that subsets a qual file giving a subseted fasta file. This works if the sequence ordering is preserved between the original fasta, and the (typically with cd-hit-454) subseted one example of use: SubsetQual.py F64.454.fasta F64.qual

MaskFasta.py
needs BioPython libraries from a file.fasta and a corresponding file.qual masks (i.e. replace by "N") the nucleotides below a quality cutoff (arg 2), removes the leading and trailing poly "N" and then delete sequences that contain more than (arg 3) percent of "N"

cleanFastaHeaders.py
from a fasta file cleans up the sequences headers removing spaces and adding a "|"

countFasta.py
Counts several elements in a fasta file: the nb of sequences, of residues, and the mean sequence length

fastaLenDistrib.py
Draws an histogramm of sequence lengths from a fasta file (arg1) , and giving a number of histo bins (arg2)

fastq2fasta.py
reads a fastq file and write a fasta file if mean sequence quality is below the (arg 2) treshold.

example 1 : fastq2fasta.py seq.fastq 20
example 2 : fastq2fasta.py seq.fastq 0 (to keep all reads)

GM2fasta.py
GeneMark default protein output to fasta

SFF2fastq.py
Generates a fastq file from a 454 SFF file

SubsetFasta.py

extracts sequences from a fasta file (arg 1)

whose id is in the IDs file (arg 2)

GCfasta.py name.fasta 100 (nb barres dans histogramme) courve GC d'un fichier fasta

cat ... ... > name pool different fasta

cd-hit-454 -i name.fasta -o namecdhit.fasta enleve les duplicats
cd-hit -i -o

--help
Man name program
less parcourir un fichier
head pour voir le debut d'un fichier
tail la queue (du fichier)

diviser un fasta en n sous fchiers de nb total / n sequences :
pyfasta split -n 10 mesSequences.fasta

pour changer le nom de gros fichiers ::
perl -pi -e "s/>NCBI/>AntarcticLakes-4443684.3-NCBI/g;" 4443684.3.fasta

pour formatdb un fichier fasta avant blast:
formatdb -i /home/tom/Desktop/4443725.3.fasta -o T -p F -n Ocean4443725.3

comment recuperer des sequences nucleiques a partir de leur nom:

cut -f 3 /home/tom/Desktop/Eval-40_Ocean45_chidb.blast > queryDsDatabank

fastacmd -i queryDsDatabank -d /home/tom/Desktop/78metchhit -o kitine.fasta

References

"The Universal Protein Resource (UniProt) in 2010." Nucleic Acids Res **38**(Database issue): D142-148.

(2007). "The Universal Protein Resource (UniProt)." Nucleic Acids Res **35**(Database issue): D193-197.

(2008). "The universal protein resource (UniProt)." Nucleic Acids Res **36**(Database issue): D190-195.

(2009). "The Universal Protein Resource (UniProt) 2009." Nucleic Acids Res **37**(Database issue): D169-174.

(2010). "A sequence of changes." Nat Rev Microbiol **8**(2): 85.

Abulencia, C. B., D. L. Wyborski, et al. (2006). "Environmental whole-genome amplification to access microbial populations in contaminated sediments." Appl Environ Microbiol **72**(5): 3291-3301.

Agarwal, N. and W. R. Bishai (2009). "cAMP signaling in Mycobacterium tuberculosis." Indian J Exp Biol **47**(6): 393-400.

Ajithdoss, D. K., B. F. Porter, et al. (2009). "Septicemia in a neonatal calf associated with Chromobacterium violaceum." Vet Pathol **46**(1): 71-74.

Akhter, Y., S. Yellaboina, et al. (2008). "Genome scale portrait of cAMP-receptor protein (CRP) regulons in mycobacteria points to their role in pathogenesis." Gene **407**(1-2): 148-158.

Allers, E., C. Niesner, et al. (2008). "Microbes enriched in seawater after addition of coral mucus." Appl Environ Microbiol **74**(10): 3274-3278.

Allwood, A. C., M. R. Walter, et al. (2006). "Stromatolite reef from the Early Archaean era of Australia." Nature **441**(7094): 714-718.

Alonso-Saez, L., O. Sanchez, et al. (2008). "Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes." Environ Microbiol **10**(9): 2444-2454.

Altermann, W. and J. Kazmierczak (2003). "Archean microfossils: a reappraisal of early life on Earth." Res Microbiol **154**(9): 611-617.

Amann, R. I., W. Ludwig, et al. (1995). "Phylogenetic identification and in situ detection of individual microbial cells without cultivation." Microbiol Rev **59**(1): 143-169.

Amato, P. and B. C. Christner (2009). "Energy metabolism response to low-temperature and frozen conditions in Psychrobacter cryohalolentis." Appl Environ Microbiol **75**(3): 711-718.

Amundsen, S. K., J. Fero, et al. (2009). "Dual nuclease and helicase activities of Helicobacter pylori AddAB are required for DNA repair, recombination, and mouse infectivity." J Biol Chem **284**(25): 16759-16766.

Anton, A., C. Grosse, et al. (1999). "CzcD is a heavy metal ion transporter involved in regulation of heavy metal resistance in Ralstonia sp. strain CH34." J Bacteriol **181**(22): 6876-6881.

Arbogast, S. and A. Ferreiro (2010). "Selenoproteins and protection against oxidative stress: selenoprotein N as a novel player at the crossroads of redox signaling and calcium homeostasis." Antioxid Redox Signal **12**(7): 893-904.

Ariza, R. R., S. P. Cohen, et al. (1994). "Repressor mutations in the marRAB operon that activate oxidative stress genes and multiple antibiotic resistance in Escherichia coli." J Bacteriol **176**(1): 143-148.

Arumugam, M., J. Raes, et al. (2011). "Enterotypes of the human gut microbiome." Nature **473**(7346): 174-180.
Asako, H., H. Nakajima, et al. (1997). "Organic solvent tolerance and antibiotic resistance increased by overexpression of marA in Escherichia coli." Appl Environ Microbiol **63**(4): 1428-1433.
Ashby, M. N., J. Rine, et al. (2007). "Serial analysis of rRNA genes and the unexpected dominance of rare members of microbial communities." Appl Environ Microbiol **73**(14): 4532-4542.
Atlas, R. M. (1999). "Legionella: from environmental habitats to disease pathology, detection and control." Environ Microbiol **1**(4): 283-293.
Ayala-Castro, C., A. Saini, et al. (2008). "Fe-S cluster assembly pathways in bacteria." Microbiol Mol Biol Rev **72**(1): 110-125, table of contents.
Ayala-del-Rio, H. L., P. S. Chain, et al. (2010). "The genome sequence of Psychrobacter arcticus 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth." Appl Environ Microbiol **76**(7): 2304-2312.
B., N. J., P. T., et al. (1980). "High Affinity Iron Transport in Microorganisms." AMERICAN CHEMICAL SOCIETY **140**: 263-278.
Backhed, F., H. Ding, et al. (2004). "The gut microbiota as an environmental factor that regulates fat storage." Proc Natl Acad Sci U S A **101**(44): 15718-15723.
Badger, J. H., T. R. Hoover, et al. (2006). "Comparative genomic evidence for a close relationship between the dimorphic prosthecate bacteria Hyphomonas neptunium and Caulobacter crescentus." J Bacteriol **188**(19): 6841-6850.
Bagramyan, K. and A. Trchounian (2003). "Structural and functional features of formate hydrogen lyase, an enzyme of mixed-acid fermentation from Escherichia coli." Biochemistry (Mosc) **68**(11): 1159-1170.
Bailey, J. S., L. C. Blankenship, et al. (1991). "Effect of fructooligosaccharide on Salmonella colonization of the chicken intestine." Poult Sci **70**(12): 2433-2438.
Baker-Austin, C., M. Dopson, et al. (2005). "Molecular insight into extreme copper resistance in the extremophilic archaeon 'Ferroplasma acidarmanus' Fer1." Microbiology **151**(Pt 8): 2637-2646.
Baker-Austin, C., J. Potrykus, et al. (2010). "Biofilm development in the extremely acidophilic archaeon 'Ferroplasma acidarmanus' Fer1." Extremophiles **14**(6): 485-491.
Bakermans, C., H. L. Ayala-del-Rio, et al. (2006). "Psychrobacter cryohalolentis sp. nov. and Psychrobacter arcticus sp. nov., isolated from Siberian permafrost." Int J Syst Evol Microbiol **56**(Pt 6): 1285-1291.
Bakken, L. R. (1985). "Separation and purification of bacteria from soil." Appl Environ Microbiol **49**(6): 1482-1487.
Balch, W. E., G. E. Fox, et al. (1979). "Methanogens: reevaluation of a unique biological group." Microbiol Rev **43**(2): 260-296.
Baldermann, C., A. Lupas, et al. (1998). "The regulated outer membrane protein Omp21 from Comamonas acidovorans is identified as a member of a new family of eight-stranded beta-sheet proteins by its sequence and properties." J Bacteriol **180**(15): 3741-3749.
Baliga, N. S., R. Bonneau, et al. (2004). "Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea." Genome Res **14**(11): 2221-2234.
Balkwill, D. L., G. R. Drake, et al. (1997). "Taxonomic study of aromatic-degrading bacteria from deep-terrestrial-subsurface sediments and description of Sphingomonas aromaticivorans sp. nov., Sphingomonas subterranea sp. nov., and Sphingomonas stygia sp. nov." Int J Syst Bacteriol **47**(1): 191-201.

Barker, E. S., R. A. Schorn, et al. (1970). "Mars: Detection of Atmospheric Water Vapor during the Southern Hemisphere Spring and Summer Season." Science **170**(3964): 1308-1310.
Barnes, M. R., W. A. Duetz, et al. (1997). "A 3-(3-hydroxyphenyl)propionic acid catabolic pathway in Rhodococcus globerulus PWD1: cloning and characterization of the hpp operon." J Bacteriol **179**(19): 6145-6153.
Bartlett, J. G. (2002). "Clinical practice. Antibiotic-associated diarrhea." N Engl J Med **346**(5): 334-339.
Bassford, P. J., Jr., C. Bradbeer, et al. (1976). "Transport of vitamin B12 in tonB mutants of Escherichia coli." J Bacteriol **128**(1): 242-247.
Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32**(Database issue): D138-141.
Beinert, H., R. H. Holm, et al. (1997). "Iron-sulfur clusters: nature's modular, multipurpose structures." Science **277**(5326): 653-659.
Beinert, H. and P. J. Kiley (1999). "Fe-S proteins in sensing and regulatory functions." Curr Opin Chem Biol **3**(2): 152-157.
Beja, O., L. Aravind, et al. (2000). "Bacterial rhodopsin: evidence for a new type of phototrophy in the sea." Science **289**(5486): 1902-1906.
Belaich, J. P., C. Tardif, et al. (1997). "The cellulolytic system of Clostridium cellulolyticum." J Biotechnol **57**(1-3): 3-14.
Benson, D. R. and W. B. Silvester (1993). "Biology of Frankia strains, actinomycete symbionts of actinorhizal plants." Microbiol Rev **57**(2): 293-319.
Berlemont, R., D. Pipers, et al. (2011). "Exploring the Antarctic soil metagenome as a source of novel cold-adapted enzymes and genetic mobile elements." Rev Argent Microbiol **43**(2): 94-103.
Bernard, P. and M. Couturier (1992). "Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes." J Mol Biol **226**(3): 735-745.
Bernard, P., K. E. Kezdy, et al. (1993). "The F plasmid CcdB protein induces efficient ATP-dependent DNA cleavage by gyrase." J Mol Biol **234**(3): 534-541.
Berry, A. E., C. Chiocchini, et al. (2003). "Isolation of high molecular weight DNA from soil for cloning into BAC vectors." FEMS Microbiol Lett **223**(1): 15-20.
Bertin, P. N., A. Heinrich-Salmeron, et al. (2011). "Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics." ISME J **5**(11): 1735-1747.
Bertrand, H., F. Poly, et al. (2005). "High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction." J Microbiol Methods **62**(1): 1-11.
Beswick, P. H., G. H. Hall, et al. (1976). "Copper toxicity: evidence for the conversion of cupric to cuprous copper in vivo under anaerobic conditions." Chem Biol Interact **14**(3-4): 347-356.
Bevers, L. E., P. L. Hagedoorn, et al. (2006). "Tungsten transport protein A (WtpA) in Pyrococcus furiosus: the first member of a new class of tungstate and molybdate transporters." J Bacteriol **188**(18): 6498-6505.
Beynon, J., A. Ally, et al. (1987). "Comparative organization of nitrogen fixation-specific genes from Azotobacter vinelandii and Klebsiella pneumoniae: DNA sequence of the nifUSV genes." J Bacteriol **169**(9): 4024-4029.
Biddle, J. F., S. Fitz-Gibbon, et al. (2008). "Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment." Proc Natl Acad Sci U S A **105**(30): 10583-10588.

Biebl, H., M. Allgaier, et al. (2005). "Dinoroseobacter shibae gen. nov., sp. nov., a new aerobic phototrophic bacterium isolated from dinoflagellates." Int J Syst Evol Microbiol **55**(Pt 3): 1089-1096.
Binga, E. K., R. S. Lasken, et al. (2008). "Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology." ISME J **2**(3): 233-241.
Birolo, L., M. L. Tutino, et al. (2000). "Aspartate aminotransferase from the Antarctic bacterium Pseudoalteromonas haloplanktis TAC 125. Cloning, expression, properties, and molecular modelling." Eur J Biochem **267**(9): 2790-2802.
Bock, A., K. Forchhammer, et al. (1991). "Selenocysteine: the 21st amino acid." Mol Microbiol **5**(3): 515-520.
Boente, R. F., L. Q. Ferreira, et al. (2010). "Detection of resistance genes and susceptibility patterns in Bacteroides and Parabacteroides strains." Anaerobe **16**(3): 190-194.
Bohin, J. P. (2000). "Osmoregulated periplasmic glucans in Proteobacteria." FEMS Microbiol Lett **186**(1): 11-19.
Bonam, D., L. Lehman, et al. (1989). "Regulation of carbon monoxide dehydrogenase and hydrogenase in Rhodospirillum rubrum: effects of CO and oxygen on synthesis and activity." J Bacteriol **171**(6): 3102-3107.
Bond, D. R., D. E. Holmes, et al. (2002). "Electrode-reducing microorganisms that harvest energy from marine sediments." Science **295**(5554): 483-485.
Boos, W. and H. Shuman (1998). "Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation." Microbiol Mol Biol Rev **62**(1): 204-229.
Borella, P., E. Guerrieri, et al. (2005). "Water ecology of Legionella and protozoan: environmental and public health perspectives." Biotechnol Annu Rev **11**: 355-380.
Borodin, V. B., A. A. Tsygankov, et al. (2000). "Hydrogen production by Anabaena variabilis PK84 under simulated outdoor conditions." Biotechnol Bioeng **69**(5): 478-485.
Boubakri, H., M. Beuf, et al. (2006). "Development of metagenomic DNA shuffling for the construction of a xenobiotic gene." Gene **375**: 87-94.
Boyd, P. W., T. Jickells, et al. (2007). "Mesoscale iron enrichment experiments 1993-2005: synthesis and future directions." Science **315**(5812): 612-617.
Bozal, N., M. J. Montes, et al. (2003). "Characterization of several Psychrobacter strains isolated from Antarctic environments and description of Psychrobacter luti sp. nov. and Psychrobacter fozii sp. nov." Int J Syst Evol Microbiol **53**(Pt 4): 1093-1100.
Brandes, N., A. Rinck, et al. (2007). "Nitrosative stress treatment of E. coli targets distinct set of thiol-containing proteins." Mol Microbiol **66**(4): 901-914.
Bratlie, M. S., J. Johansen, et al. (2010). "Gene duplications in prokaryotes can be associated with environmental adaptation." BMC Genomics **11**: 588.
Breezee, J., N. Cady, et al. (2004). "Subfreezing growth of the sea ice bacterium "Psychromonas ingrahamii"." Microb Ecol **47**(3): 300-304.
Breitbart, M., P. Salamon, et al. (2002). "Genomic analysis of uncultured marine viral communities." Proc Natl Acad Sci U S A **99**(22): 14250-14255.
Breitung, J., G. Borner, et al. (1992). "Salt dependence, kinetic properties and catalytic mechanism of N-formylmethanofuran:tetrahydromethanopterin formyltransferase from the extreme thermophile Methanopyrus kandleri." Eur J Biochem **210**(3): 971-981.
Brim, H., A. Venkateswaran, et al. (2003). "Engineering Deinococcus geothermalis for bioremediation of high-temperature radioactive waste environments." Appl Environ Microbiol **69**(8): 4575-4582.
Bruce, K. D., W. D. Hiorns, et al. (1992). "Amplification of DNA from native populations of soil bacteria by using the polymerase chain reaction." Appl Environ Microbiol **58**(10): 3413-3416.

Bruggemann, H., S. Baumer, et al. (2003). "The genome sequence of Clostridium tetani, the causative agent of tetanus disease." Proc Natl Acad Sci U S A **100**(3): 1316-1321.
Bruins, M. R., S. Kapil, et al. (2000). "Microbial resistance to metals in the environment." Ecotoxicol Environ Saf **45**(3): 198-207.
Brun, A. and E. Englund (1986). "Brain changes in dementia of Alzheimer's type relevant to new imaging diagnostic methods." Prog Neuropsychopharmacol Biol Psychiatry **10**(3-5): 297-308.
Bryant, M. P. (1974). "Nutritional features and ecology of predominant anaerobic bacteria of the intestinal tract." Am J Clin Nutr **27**(11): 1313-1319.
Buchan, A., J. M. Gonzalez, et al. (2005). "Overview of the marine roseobacter lineage." Appl Environ Microbiol **71**(10): 5665-5677.
Buddington, K. K., J. B. Donahoo, et al. (2002). "Dietary oligofructose and inulin protect mice from enteric and systemic pathogens and tumor inducers." J Nutr **132**(3): 472-477.
Byrne-Bailey, K. G., K. A. Weber, et al. (2010). "Completed genome sequence of the anaerobic iron-oxidizing bacterium Acidovorax ebreus strain TPSY." J Bacteriol **192**(5): 1475-1476.
Cai, S., J. Li, et al. (2010). "Cellulosilyticum ruminicola, a newly described rumen bacterium that possesses redundant fibrolytic-protein-encoding genes and degrades lignocellulose with multiple carbohydrate- borne fibrolytic enzymes." Appl Environ Microbiol **76**(12): 3818-3824.
Cane, D. E., C. T. Walsh, et al. (1998). "Harnessing the biosynthetic code: combinations, permutations, and mutations." Science **282**(5386): 63-68.
Cao, X., X. Liu, et al. (2003). "Alkaliphilus crotonatoxidans sp. nov., a strictly anaerobic, crotonate-dismutating bacterium isolated from a methanogenic environment." Int J Syst Evol Microbiol **53**(Pt 4): 971-975.
Carepo, M. S., J. S. Azevedo, et al. (2004). "Identification of Chromobacterium violaceum genes with potential biotechnological application in environmental detoxification." Genet Mol Res **3**(1): 181-194.
Champoux, J. J. (2001). "DNA topoisomerases: structure, function, and mechanism." Annu Rev Biochem **70**: 369-413.
Charlson, R. J., J. E. Lovelock, et al. (1987). "Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate." Nature **326**(6114): 655-661.
Chen, C. M., Q. Z. Ye, et al. (1990). "Molecular biology of carbon-phosphorus bond cleavage. Cloning and sequencing of the phn (psiD) genes involved in alkylphosphonate uptake and C-P lyase activity in Escherichia coli B." J Biol Chem **265**(8): 4461-4471.
Chesson, A., C. S. Stewart, et al. (1986). "Degradation of isolated grass mesophyll, epidermis and fibre cell walls in the rumen and by cellulolytic rumen bacteria in axenic culture." journal of applied microbiology **60**(4): 327–336.
Chopra, I. and M. Roberts (2001). "Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance." Microbiol Mol Biol Rev **65**(2): 232-260 ; second page, table of contents.
Claverie, J. M. and C. Abergel (2009). "Mimivirus and its virophage." Annu Rev Genet **43**: 49-66.
Clifford, S., A. Treiman, et al. (1998). "Introduction to special section: early Mars." J Geophys Res **103**(E13): 31405.
Clokie, M. R. and N. H. Mann (2006). "Marine cyanophages and light." Environ Microbiol **8**(12): 2074-2082.

Coates, J. D., R. Chakraborty, et al. (2001). "Anaerobic benzene oxidation coupled to nitrate reduction in pure culture by two strains of Dechloromonas." Nature **411**(6841): 1039-1043.
Cohen, S. P., L. M. McMurry, et al. (1989). "Cross-resistance to fluoroquinolones in multiple-antibiotic-resistant (Mar) Escherichia coli selected by tetracycline or chloramphenicol: decreased drug accumulation associated with membrane changes in addition to OmpF reduction." Antimicrob Agents Chemother **33**(8): 1318-1325.
Connell, S. R., D. M. Tracz, et al. (2003). "Ribosomal protection proteins and their mechanism of tetracycline resistance." Antimicrob Agents Chemother **47**(12): 3675-3681.
Courtois, S., C. M. Cappellano, et al. (2003). "Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products." Appl Environ Microbiol **69**(1): 49-55.
Courtois, S., A. Frostegard, et al. (2001). "Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation." Environ Microbiol **3**(7): 431-439.
Couturier, M., M. Bahassi el, et al. (1998). "Bacterial death by DNA gyrase poisoning." Trends Microbiol **6**(7): 269-275.
Cozzarelli, N. R. (1980). "DNA gyrase and the supercoiling of DNA." Science **207**(4434): 953-960.
Crane, B. R., J. Sudhamsu, et al. (2010). "Bacterial nitric oxide synthases." Annu Rev Biochem **79**: 445-470.
Crawley, M. J., A. E. Johnston, et al. (2005). "Determinants of species richness in the Park Grass Experiment." Am Nat **165**(2): 179-192.
Creczynski-Pasa, T. B. and R. V. Antonio (2004). "Energetic metabolism of Chromobacterium violaceum." Genet Mol Res **3**(1): 162-166.
Crossman, L. and J. M. Dow (2004). "Biofilm formation and dispersal in Xanthomonas campestris." Microbes Infect **6**(6): 623-629.
Crossman, L. C., V. C. Gould, et al. (2008). "The complete genome, comparative and functional analysis of Stenotrophomonas maltophilia reveals an organism heavily shielded by drug resistance determinants." Genome Biol **9**(4): R74.
Curtis, T. P., W. T. Sloan, et al. (2002). "Estimating prokaryotic diversity and its limits." Proc Natl Acad Sci U S A **99**(16): 10494-10499.
Dabrowska, G., J. Prusiniska, et al. (2006). "[The stringent response--bacterial mechanism of an adaptive stress response]." Postepy Biochem **52**(1): 87-93.
Dagley, S., P. J. Chapman, et al. (1965). "The metabolism of beta-phenylpropionic acid by an Achromobacter." Biochem J **97**(3): 643-650.
Dao-Thi, M. H., L. Van Melderen, et al. (2005). "Molecular basis of gyrase poisoning by the addiction toxin CcdB." J Mol Biol **348**(5): 1091-1102.
Dartnell, L. R., S. J. Hunter, et al. (2010). "Low-temperature ionizing radiation resistance of Deinococcus radiodurans and Antarctic Dry Valley bacteria." Astrobiology **10**(7): 717-732.
Dassarma, S., S. P. Kennedy, et al. (2001). "Genomic perspective on the photobiology of Halobacterium species NRC-1, a phototrophic, phototactic, and UV-tolerant haloarchaeon." Photosynth Res **70**(1): 3-17.
Davis, K. E., P. Sangwan, et al. (2011). "Acidobacteria, Rubrobacteridae and Chloroflexi are abundant among very slow-growing and mini-colony-forming soil bacteria." Environ Microbiol **13**(3): 798-805.
De la Cerda, B., O. Castielli, et al. (2007). "A proteomic approach to iron and copper homeostasis in cyanobacteria." Brief Funct Genomic Proteomic **6**(4): 322-329.

De Rosa, C. T., R. Nickle, et al. (2003). "The impact of toxicology on public health policy and service: an update." Toxicol Ind Health **19**(2-6): 115-124.
de Vera, J. P., D. Mohlmann, et al. (2010). "Survival potential and photosynthetic activity of lichens under Mars-like conditions: a laboratory study." Astrobiology **10**(2): 215-227.
Delmont, T. O., C. Malandain, et al. "Metagenomic mining for microbiologists." ISME J.
Delmont, T. O., P. Robe, et al. (2011). "Accessing the soil metagenome for studies of microbial diversity." Appl Environ Microbiol **77**(4): 1315-1324.
Delmont, T. O., P. Robe, et al. (2011). "Metagenomic comparison of direct and indirect soil DNA extraction approaches." J Microbiol Methods **86**(3): 397-400.
DeLong, E. F. (2000). "Extreme genomes." Genome Biol **1**(6): REVIEWS1029.
DeLong, E. F. and D. M. Karl (2005). "Genomic perspectives in microbial oceanography." Nature **437**(7057): 336-342.
DeLong, E. F., C. M. Preston, et al. (2006). "Community genomics among stratified microbial assemblages in the ocean's interior." Science **311**(5760): 496-503.
Demaneche, S., H. Sanguin, et al. (2008). "Antibiotic-resistant soil bacteria in transgenic plant fields." Proc Natl Acad Sci U S A **105**(10): 3957-3962.
Denton, M. and K. G. Kerr (1998). "Microbiological and clinical aspects of infection associated with Stenotrophomonas maltophilia." Clin Microbiol Rev **11**(1): 57-80.
Derrien, M., M. C. Collado, et al. (2008). "The Mucin degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract." Appl Environ Microbiol **74**(5): 1646-1648.
Derrien, M., E. E. Vaughan, et al. (2004). "Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium." Int J Syst Evol Microbiol **54**(Pt 5): 1469-1476.
Diaz, B. and D. Schulze-Makuch (2006). "Microbial survival rates of Escherichia coli and Deinococcus radiodurans under low temperature, low pressure, and UV-Irradiation conditions, and their relevance to possible Martian life." Astrobiology **6**(2): 332-347.
Diaz, E. (2004). "Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility." Int Microbiol **7**(3): 173-180.
Diaz, E., A. Ferrandez, et al. (1998). "Characterization of the hca cluster encoding the dioxygenolytic pathway for initial catabolism of 3-phenylpropionic acid in Escherichia coli K-12." J Bacteriol **180**(11): 2915-2923.
Dickschat, J. S., C. Zell, et al. (2010). "Pathways and substrate specificity of DMSP catabolism in marine bacteria of the Roseobacter clade." Chembiochem **11**(3): 417-425.
Dinsdale, E. A., R. A. Edwards, et al. (2008). "Functional metagenomic profiling of nine biomes." Nature **452**(7187): 629-632.
Dinsdale, E. A., O. Pantos, et al. (2008). "Microbial ecology of four coral atolls in the Northern Line Islands." PLoS One **3**(2): e1584.
Doi, R. H., M. Goldstein, et al. (1994). "The Clostridium cellulovorans cellulosome." Crit Rev Microbiol **20**(2): 87-93.
Dopson, M., C. Baker-Austin, et al. (2005). "Analysis of differential protein expression during growth states of Ferroplasma strains and insights into electron transport for iron oxidation." Microbiology **151**(Pt 12): 4127-4137.
Driks, A. (1999). "Bacillus subtilis spore coat." Microbiol Mol Biol Rev **63**(1): 1-20.
Duine, J. A. (1999). "The PQQ story." J Biosci Bioeng **88**(3): 231-236.
Duran, N. and C. F. Menck (2001). "Chromobacterium violaceum: a review of pharmacological and industiral perspectives." Crit Rev Microbiol **27**(3): 201-222.

Dürre, P. (1998). "New insights and novel developments in clostridial acetone/butanol/isopropanol fermentation." Applied Microbiology and Biotechnology **49**(6): 639-648.
Dwyer, D. J., M. A. Kohanski, et al. (2007). "Gyrase inhibitors induce an oxidative damage cellular death pathway in Escherichia coli." Mol Syst Biol **3**: 91.
Eckburg, P. B., E. M. Bik, et al. (2005). "Diversity of the human intestinal microbial flora." Science **308**(5728): 1635-1638.
Edwards, K. J., P. L. Bond, et al. (2000). "An archaeal iron-oxidizing extreme acidophile important in acid mine drainage." Science **287**(5459): 1796-1799.
Ehlers, K. T., M. Rusan, et al. (2009). "Fusobacterium necrophorum: most prevalent pathogen in peritonsillar abscess in Denmark." Clin Infect Dis(49): 1467-1472.
Elbein, A. D. (1974). "The metabolism of alpha,alpha-trehalose." Adv Carbohydr Chem Biochem **30**: 227-256.
Elbein, A. D., Y. T. Pan, et al. (2003). "New insights on trehalose: a multifunctional molecule." Glycobiology **13**(4): 17R-27R.
Elshahed, M. S., N. H. Youssef, et al. (2008). "Novelty and uniqueness patterns of rare members of the soil biosphere." Appl Environ Microbiol **74**(17): 5422-5428.
Engelhardt, B., M. T. Martin-Simonet, et al. (1998). "Adhesion molecule phenotype of T lymphocytes in inflamed CNS." J Neuroimmunol **84**(1): 92-104.
Erable, B., N. M. Duteanu, et al. (2010). "Application of electro-active biofilms." Biofouling **26**(1): 57-71.
Fajardo-Cavazos, P., A. C. Schuerger, et al. (2010). "Exposure of DNA and Bacillus subtilis spores to simulated martian environments: use of quantitative PCR (qPCR) to measure inactivation rates of DNA to function as a template molecule." Astrobiology **10**(4): 403-411.
Falb, M., F. Pfeiffer, et al. (2005). "Living with two extremes: conclusions from the genome sequence of Natronomonas pharaonis." Genome Res **15**(10): 1336-1343.
Falkowski, P. G. and Y. Rosenthal (2001). "Biological diversity and resource plunder in the geological record: casual correlations or causal relationships?" Proc Natl Acad Sci U S A **98**(8): 4290-4292.
Fang, F. C. (1997). "Perspectives series: host/pathogen interactions. Mechanisms of nitric oxide-related antimicrobial activity." J Clin Invest **99**(12): 2818-2825.
Feinstein, L. M., W. J. Sul, et al. (2009). "Assessment of bias associated with incomplete extraction of microbial DNA from soil." Appl Environ Microbiol **75**(16): 5428-5433.
Feller, G. and C. Gerday (2003). "Psychrophilic enzymes: hot topics in cold adaptation." Nat Rev Microbiol **1**(3): 200-208.
Fendrihan, S., A. Berces, et al. (2009). "Investigating the effects of simulated martian ultraviolet radiation on Halococcus dombrowskii and other extremely halophilic archaebacteria." Astrobiology **9**(1): 104-112.
Ferreira, A. C., M. F. Nobre, et al. (1997). "Deinococcus geothermalis sp. nov. and Deinococcus murrayi sp. nov., two extremely radiation-resistant and slightly thermophilic species from hot springs." Int J Syst Bacteriol **47**(4): 939-947.
Field, C. B., M. J. Behrenfeld, et al. (1998). "Primary production of the biosphere: integrating terrestrial and oceanic components." Science **281**(5374): 237-240.
Filee, J. (2009). "Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses." J Invertebr Pathol **101**(3): 169-171.
Finneran, K. T., C. V. Johnsen, et al. (2003). "Rhodoferax ferrireducens sp. nov., a psychrotolerant, facultatively anaerobic bacterium that oxidizes acetate with the reduction of Fe(III)." Int J Syst Evol Microbiol **53**(Pt 3): 669-673.

Fischer, B., G. Rummel, et al. (2002). "The FtsH protease is involved in development, stress response and heat shock control in Caulobacter crescentus." Mol Microbiol **44**(2): 461-478.
Fischer, R. J., J. Helms, et al. (1993). "Cloning, sequencing, and molecular analysis of the sol operon of Clostridium acetobutylicum, a chromosomal locus involved in solventogenesis." J Bacteriol **175**(21): 6959-6969.
Formisano, V., S. Atreya, et al. (2004). "Detection of methane in the atmosphere of Mars." Science **306**(5702): 1758-1761.
Frias-Lopez, J., Y. Shi, et al. (2008). "Microbial community gene expression in ocean surface waters." Proc Natl Acad Sci U S A **105**(10): 3805-3810.
Friedmann, E. I. and R. Ocampo-Friedmann (1995). "A primitive cyanobacterium as pioneer microorganism for terraforming Mars." Adv Space Res **15**(3): 243-246.
Fromin, N., J. Hamelin, et al. (2002). "Statistical analysis of denaturing gel electrophoresis (DGE) fingerprinting patterns." Environ Microbiol **4**(11): 634-643.
Fromme, P. and I. Grotjohann (2008). "Structure of Photosystems I and II." Results Probl Cell Differ **45**: 33-72.
Frostegard, A., S. Courtois, et al. (1999). "Quantification of bias related to the extraction of DNA directly from soils." Appl Environ Microbiol **65**(12): 5409-5420.
Frostl, J. M. and J. Overmann (1998). "Physiology and tactic response of the phototrophic consortium "Chlorochromatium aggregatum"." Arch Microbiol **169**(2): 129-135.
Fuhrman, J. A., I. Hewson, et al. (2006). "Annually reoccurring bacterial communities are predictable from ocean conditions." Proc Natl Acad Sci U S A **103**(35): 13104-13109.
Fuhrman, J. A., M. S. Schwalbach, et al. (2008). "Proteorhodopsins: an array of physiological roles?" Nat Rev Microbiol **6**(6): 488-494.
Futterer, O., A. Angelov, et al. (2004). "Genome sequence of Picrophilus torridus and its implications for life around pH 0." Proc Natl Acad Sci U S A **101**(24): 9091-9096.
Galbally, I. E. and W. Kirstine (2002). "The Production of Methanol by Flowering Plants and the Global Cycle of Methanol." Journal of Atmospheric Chemistry **43**(3): 195-229.
Galindo, C. L., C. Gutierrez, Jr., et al. (2006). "Potential involvement of galectin-3 and SNAP23 in Aeromonas hydrophila cytotoxic enterotoxin-induced host cell apoptosis." Microb Pathog **40**(2): 56-68.
Galvao, T. C., W. W. Mohn, et al. (2005). "Exploring the microbial biodegradation and biotransformation gene pool." Trends Biotechnol **23**(10): 497-506.
Gans, J., M. Wolinsky, et al. (2005). "Computational improvements reveal great bacterial diversity and high metal toxicity in soil." Science **309**(5739): 1387-1390.
Garcia Martin, H., N. Ivanova, et al. (2006). "Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities." Nat Biotechnol **24**(10): 1263-1269.
Garcia Sanchez, R., K. Karhumaa, et al. (2010). "Improved xylose and arabinose utilization by an industrial recombinant Saccharomyces cerevisiae strain using evolutionary engineering." Biotechnol Biofuels **3**: 13.
Gauthier, M. J., B. Lafay, et al. (1992). "Marinobacter hydrocarbonoclasticus gen. nov., sp. nov., a new, extremely halotolerant, hydrocarbon-degrading marine bacterium." Int J Syst Bacteriol **42**(4): 568-576.
Gellert, M., K. Mizuuchi, et al. (1976). "DNA gyrase: an enzyme that introduces superhelical turns into DNA." Proc Natl Acad Sci U S A **73**(11): 3872-3876.
Genin, S. and C. Boucher (2004). "Lessons learned from the genome analysis of ralstonia solanacearum." Annu Rev Phytopathol **42**: 107-134.

George, A. M. and S. B. Levy (1983). "Amplifiable resistance to tetracycline, chloramphenicol, and other antibiotics in Escherichia coli: involvement of a non-plasmid-determined efflux of tetracycline." J Bacteriol **155**(2): 531-540.
Gerber, S., M. Comellas-Bigler, et al. (2008). "Structural basis of trans-inhibition in a molybdate/tungstate ABC transporter." Science **321**(5886): 246-250.
Gerlach, W., S. Junemann, et al. (2009). "WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads." BMC Bioinformatics **10**: 430.
Gibson, D. G., J. I. Glass, et al. (2010). "Creation of a bacterial cell controlled by a chemically synthesized genome." Science **329**(5987): 52-56.
Gibson, G. R., E. R. Beatty, et al. (1995). "Selective stimulation of bifidobacteria in the human colon by oligofructose and inulin." Gastroenterology **108**(4): 975-982.
Gibson, G. R. and X. Wang (1994). "Enrichment of bifidobacteria from human gut contents by oligofructose using continuous culture." FEMS Microbiol Lett **118**(1-2): 121-127.
Gihring, T. M., P. L. Bond, et al. (2003). "Arsenic resistance in the archaeon "Ferroplasma acidarmanus": new insights into the structure and evolution of the ars genes." Extremophiles **7**(2): 123-130.
Gilad, J. (2007). "Burkholderia mallei and Burkholderia pseudomallei: the causative micro-organisms of glanders and melioidosis." Recent Pat Antiinfect Drug Discov **2**(3): 233-241.
Gilbert, J. A., F. Meyer, et al. (2010). "The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010." Stand Genomic Sci **3**(3): 249-253.
Gill, S. R., M. Pop, et al. (2006). "Metagenomic analysis of the human distal gut microbiome." Science **312**(5778): 1355-1359.
Gillespie, D. E., S. F. Brady, et al. (2002). "Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA." Appl Environ Microbiol **68**(9): 4301-4306.
Ginolhac, A., C. Jarrin, et al. (2004). "Phylogenetic analysis of polyketide synthase I domains from soil metagenomic libraries allows selection of promising clones." Appl Environ Microbiol **70**(9): 5522-5527.
Ginzburg, M., L. Sachs, et al. (1970). "Ion metabolism in a Halobacterium. I. Influence of age of culture on intracellular concentrations." J Gen Physiol **55**(2): 187-207.
Giovannoni, S. J., L. Bibbs, et al. (2005). "Proteorhodopsin in the ubiquitous marine bacterium SAR11." Nature **438**(7064): 82-85.
Goldmark, P. J. and S. Linn (1972). "Purification and properties of the recBC DNase of Escherichia coli K-12." J Biol Chem **247**(6): 1849-1860.
Goll, J., D. B. Rusch, et al. (2010). "METAREP: JCVI metagenomics reports--an open source tool for high-performance comparative metagenomics." Bioinformatics **26**(20): 2631-2632.
Golyshin, P. N., V. A. Martins Dos Santos, et al. (2003). "Genome sequence completed of Alcanivorax borkumensis, a hydrocarbon-degrading bacterium that plays a global role in oil removal from marine systems." J Biotechnol **106**(2-3): 215-220.
Gonzalez, J. E. and M. M. Marketon (2003). "Quorum sensing in nitrogen-fixing rhizobia." Microbiol Mol Biol Rev **67**(4): 574-592.
Gonzalez, O., T. Oberwinkler, et al. (2010). "Characterization of growth and metabolism of the haloalkaliphile Natronomonas pharaonis." PLoS Comput Biol **6**(6): e1000799.
Goodner, B., G. Hinkle, et al. (2001). "Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58." Science **294**(5550): 2323-2328.

Gratia, J. P. (1964). "[Resistance to Colicin B in Escherichia Coli. Specificity Relations among Colicins B, I and V and Phage T-4. Genetic Study]." Ann Inst Pasteur (Paris) **107**: SUPPL:132-151.
Gray, S. J. (1984). "Aeromonas hydrophila in livestock: incidence, biochemical characteristics and antibiotic susceptibility." J Hyg (Lond) **92**(3): 365-375.
Green, D. H., P. R. Wakeley, et al. (1999). "Characterization of two Bacillus probiotics." Appl Environ Microbiol **65**(9): 4288-4291.
Greenberg, J. T., J. H. Chou, et al. (1991). "Activation of oxidative stress genes by mutations at the soxQ/cfxB/marA locus of Escherichia coli." J Bacteriol **173**(14): 4433-4439.
Griffiths, R. I., A. S. Whiteley, et al. (2000). "Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition." Appl Environ Microbiol **66**(12): 5488-5491.
Grzymski, J. J., A. E. Murray, et al. (2008). "Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility." Proc Natl Acad Sci U S A **105**(45): 17516-17521.
Guarner, F. and J. R. Malagelada (2003). "Gut flora in health and disease." Lancet **361**(9356): 512-519.
Guettler, M. V., D. Rumler, et al. (1999). "Actinobacillus succinogenes sp. nov., a novel succinic-acid-producing strain from the bovine rumen." Int J Syst Bacteriol **49 Pt 1**: 207-216.
Halbleib, C. M. and P. W. Ludden (2000). "Regulation of biological nitrogen fixation." J Nutr **130**(5): 1081-1084.
Hall, R. M. and C. M. Collis (1995). "Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination." Mol Microbiol **15**(4): 593-600.
Hamp, T. J., W. J. Jones, et al. (2009). "Effects of experimental choices and analysis noise on surveys of the "rare biosphere"." Appl Environ Microbiol **75**(10): 3263-3270.
Hanawalt, P. C. (1966). "The U.V. sensitivity of bacteria: its relation to the DNA replication cycle." Photochem Photobiol **5**(1): 1-12.
Handelsman, J., M. R. Rondon, et al. (1998). "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products." Chem Biol **5**(10): R245-249.
Hanson, R. S., J. A. Peterson, et al. (1970). "Unique biochemical events in bacterial sporulation." Annu Rev Microbiol **24**: 53-90.
Hao, B., W. Gong, et al. (2002). "A new UAG-encoded residue in the structure of a methanogen methyltransferase." Science **296**(5572): 1462-1466.
Happe, T., K. Schutz, et al. (2000). "Transcriptional and mutational analysis of the uptake hydrogenase of the filamentous cyanobacterium Anabaena variabilis ATCC 29413." J Bacteriol **182**(6): 1624-1631.
Hara, A., K. Syutsubo, et al. (2003). "Alcanivorax which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation." Environ Microbiol **5**(9): 746-753.
Harrison, F. H. and C. S. Harwood (2005). "The pimFABCDE operon from Rhodopseudomonas palustris mediates dicarboxylic acid degradation and participates in anaerobic benzoate degradation." Microbiology **151**(Pt 3): 727-736.
Hassett, D. J., J. Cuppoletti, et al. (2002). "Anaerobic metabolism and quorum sensing by Pseudomonas aeruginosa biofilms in chronically infected cystic fibrosis airways: rethinking antibiotic treatment strategies and drug targets." Adv Drug Deliv Rev **54**(11): 1425-1443.

Haveman, S. A., R. J. DiDonato, Jr., et al. (2008). "Genome-wide gene expression patterns and growth requirements suggest that Pelobacter carbinolicus reduces Fe(III) indirectly via sulfide production." Appl Environ Microbiol **74**(14): 4277-4284.
Hayes, F. (2003). "Toxins-antitoxins: plasmid maintenance, programmed cell death, and cell cycle arrest." Science **301**(5639): 1496-1499.
Hayes, J. M. and J. R. Waldbauer (2006). "The carbon cycle and associated redox processes through time." Philos Trans R Soc Lond B Biol Sci **361**(1470): 931-950.
Hazen, T. C., E. A. Dubinsky, et al. (2010). "Deep-sea oil plume enriches indigenous oil-degrading bacteria." Science **330**(6001): 204-208.
Head, I. M., J. R. Saunders, et al. (1998). "Microbial Evolution, Diversity, and Ecology: A Decade of Ribosomal RNA Analysis of Uncultivated Microorganisms." Microb Ecol **35**(1): 1-21.
Hebbeln, P., D. A. Rodionov, et al. (2007). "Biotin uptake in prokaryotes by solute transporters with an optional ATP-binding cassette-containing module." Proc Natl Acad Sci U S A **104**(8): 2909-2914.
Hecker, M. and U. Volker (1998). "Non-specific, general and multiple stress resistance of growth-restricted Bacillus subtilis cells by the expression of the sigmaB regulon." Mol Microbiol **29**(5): 1129-1136.
Henrichsen, J. (1972). "Bacterial surface translocation: a survey and a classification." Bacteriol Rev **36**(4): 478-503.
Hess, M., A. Sczyrba, et al. (2011). "Metagenomic discovery of biomass-degrading genes and genomes from cow rumen." Science **331**(6016): 463-467.
Heusser, L. E. and N. J. Shackleton (1979). "Direct marine-continental correlation: 150,000-year oxygen isotope--pollen record from the north pacific." Science **204**(4395): 837-839.
Higginbottom, J., K. M. Bagnall, et al. (1976). "Ultrasound monitoring of fetal movements. A method for assessing fetal development?" Lancet **1**(7962): 719-721.
Higgins, C. F. (1992). "ABC transporters: from microorganisms to man." Annu Rev Cell Biol **8**: 67-113.
Hippe, H., D. Caspari, et al. (1979). "Utilization of trimethylamine and other N-methyl compounds for growth and methane formation by Methanosarcina barkeri." Proc Natl Acad Sci U S A **76**(1): 494-498.
Hobbie, J. E., R. J. Daley, et al. (1977). "Use of nuclepore filters for counting bacteria by fluorescence microscopy." Appl Environ Microbiol **33**(5): 1225-1228.
Hoff, K. J., T. Lingner, et al. (2009). "Orphelia: predicting genes in metagenomic sequencing reads." Nucleic Acids Res **37**(Web Server issue): W101-105.
Holmes, D. E., D. R. Bond, et al. (2004). "Microbial communities associated with electrodes harvesting electricity from a variety of aquatic sediments." Microb Ecol **48**(2): 178-190.
Holtzendorff, J., F. Partensky, et al. (2008). "Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium Prochlorococcus marinus PCC 9511." J Biol Rhythms **23**(3): 187-199.
Hong, S., J. Bunge, et al. (2009). "Polymerase chain reaction primers miss half of rRNA microbial diversity." ISME J **3**(12): 1365-1373.
Hooper, L. V., T. Midtvedt, et al. (2002). "How host-microbial interactions shape the nutrient environment of the mammalian intestine." Annu Rev Nutr **22**: 283-307.
Hopkins, M. J. and G. T. Macfarlane (2002). "Changes in predominant bacterial populations in human faeces with age and with Clostridium difficile infection." J Med Microbiol **51**(5): 448-454.

Hopkins, M. J. and G. T. Macfarlane (2003). "Nondigestible oligosaccharides enhance bacterial colonization resistance against Clostridium difficile in vitro." Appl Environ Microbiol **69**(4): 1920-1927.
Hu, G. Q., J. T. Guo, et al. (2009). "MetaTISA: Metagenomic Translation Initiation Site Annotator for improving gene start prediction." Bioinformatics **25**(14): 1843-1845.
Hu, Y., C. Fu, et al. "Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment." FEMS Microbiol Ecol **72**(2): 228-237.
Huber, R., M. Kurr, et al. (1989). "A novel group of abyssal methanogenic archaebacteria (Methanopyrus) growing at 110 [deg]C." **342**(6251): 833-834.
Huson, D. H., A. F. Auch, et al. (2007). "MEGAN analysis of metagenomic data." Genome Res **17**(3): 377-386.
Huson, D. H., D. C. Richter, et al. (2009). "Methods for comparative metagenomics." BMC Bioinformatics **10 Suppl 1**: S12.
Huu, N. B., E. B. Denner, et al. (1999). "Marinobacter aquaeolei sp. nov., a halophilic bacterium isolated from a Vietnamese oil-producing well." Int J Syst Bacteriol **49 Pt 2**: 367-375.
Irbis, C. and K. Ushida (2004). "Detection of methanogens and proteobacteria from a single cell of rumen ciliate protozoa." J Gen Appl Microbiol **50**(4): 203-212.
Ireland, M. M., J. A. Karty, et al. (2002). "Proteomic analysis of the Caulobacter crescentus stalk indicates competence for nutrient uptake." Mol Microbiol **45**(4): 1029-1041.
Ishii, S., K. Watanabe, et al. (2008). "Comparison of electrode reduction activities of Geobacter sulfurreducens and an enriched consortium in an air-cathode microbial fuel cell." Appl Environ Microbiol **74**(23): 7348-7355.
Iwai, Y. and S. Omura (1982). "Culture conditions for screening of new antibiotics." J Antibiot (Tokyo) **35**(2): 123-141.
Jacob, C., G. I. Giles, et al. (2003). "Sulfur and selenium: the role of oxidation state in protein structure and function." Angew Chem Int Ed Engl **42**(39): 4742-4758.
Jacoby, G. A. and L. S. Munoz-Price (2005). "The new beta-lactamases." N Engl J Med **352**(4): 380-391.
Jaffe, A., T. Ogura, et al. (1985). "Effects of the ccd function of the F plasmid on bacterial growth." J Bacteriol **163**(3): 841-849.
Jakosky, B. M., K. H. Nealson, et al. (2003). "Subfreezing activity of microorganisms and the potential habitability of Mars' polar regions." Astrobiology **3**(2): 343-350.
Janssen, P. H. and M. Kirs (2008). "Structure of the archaeal community of the rumen." Appl Environ Microbiol **74**(12): 3619-3625.
Janssen, P. J., R. Van Houdt, et al. (2010). "The complete genome sequence of Cupriavidus metallidurans strain CH34, a master survivalist in harsh and anthropogenic environments." PLoS One **5**(5): e10433.
Jernberg, C., S. Lofmark, et al. (2010). "Long-term impacts of antibiotic exposure on the human intestinal microbiota." Microbiology **156**(Pt 11): 3216-3223.
Ji, X. B. and T. C. Hollocher (1988). "Reduction of nitrite to nitric oxide by enteric bacteria." Biochem Biophys Res Commun **157**(1): 106-108.
Jin, D. J. and J. E. Cabrera (2006). "Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in Escherichia coli." J Struct Biol **156**(2): 284-291.
Johnson, D. C., D. R. Dean, et al. (2005). "Structure, function, and formation of biological iron-sulfur clusters." Annu Rev Biochem **74**: 247-281.
Johnston, A. W., J. D. Todd, et al. (2008). "Molecular diversity of bacterial production of the climate-changing gas, dimethyl sulphide, a molecule that impinges on local and global symbioses." J Exp Bot **59**(5): 1059-1067.

Jones, D. T. and D. R. Woods (1986). "Acetone-butanol fermentation revisited." Microbiol Rev **50**(4): 484-524.
Juarez, J. F., M. T. Zamarro, et al. (2010). "Identification of the Geobacter metallireducens bamVW two-component system, involved in transcriptional regulation of aromatic degradation." Appl Environ Microbiol **76**(1): 383-385.
Jung, Y. H., J. Y. Yi, et al. (2010). "Overexpression of cold shock protein A of Psychromonas arctica KOPRI 22215 confers cold-resistance." Protein J **29**(2): 136-142.
Kahvejian, A., J. Quackenbush, et al. (2008). "What would you do if you could sequence everything?" Nat Biotechnol **26**(10): 1125-1133.
Kang, K. H., S. K. Jang, et al. (1994). "Antibacterial phenylpropanoid glycosides from Paulownia tomentosa Steud." Arch Pharm Res **17**(6): 470-475.
Kanzler, B. E., K. R. Pfannes, et al. (2005). "Molecular characterization of the nonphotosynthetic partner bacterium in the consortium "Chlorochromatium aggregatum"." Appl Environ Microbiol **71**(11): 7434-7441.
Kaserer, W. A., X. Jiang, et al. (2008). "Insight from TonB hybrid proteins into the mechanism of iron transport through the outer membrane." J Bacteriol **190**(11): 4001-4016.
Katoch, V. M. (2004). "Infections due to non-tuberculous mycobacteria (NTM)." Indian J Med Res **120**(4): 290-304.
Kawashima, T., N. Amano, et al. (2000). "Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium." Proc Natl Acad Sci U S A **97**(26): 14257-14262.
Kay, C. W., B. Mennenga, et al. (2006). "Structure of the pyrroloquinoline quinone radical in quinoprotein ethanol dehydrogenase." J Biol Chem **281**(3): 1470-1476.
Kendrick, M. G. and T. A. Kral (2006). "Survival of methanogens during desiccation: implications for life on Mars." Astrobiology **6**(4): 546-551.
Kennaway, E. L. and I. Hieger (1930). "Carcinogenic Substances and Their Fluorescence Spectra." Br Med J **1**(3622): 1044-1046.
Kerby, R. L., P. W. Ludden, et al. (1995). "Carbon monoxide-dependent growth of Rhodospirillum rubrum." J Bacteriol **177**(8): 2241-2244.
Kerr, R. A. (2010). "Planetary science. Liquid water found on Mars, but it's still a hard road for life." Science **330**(6004): 571.
Kerscher, L., S. Nowitzki, et al. (1982). "Thermoacidophilic archaebacteria contain bacterial-type ferredoxins acting as electron acceptors of 2-oxoacid:ferredoxin oxidoreductases." Eur J Biochem **128**(1): 223-230.
Kiene, R. P., R. S. Oremland, et al. (1986). "Metabolism of reduced methylated sulfur compounds in anaerobic sediments and by a pure culture of an estuarine methanogen." Appl Environ Microbiol **52**(5): 1037-1045.
Kim, S., P. Singh, et al. (2011). "Genetic and molecular characterization of a blue light photoreceptor MGWC-1 in Magnaporth oryzae." Fungal Genet Biol **48**(4): 400-407.
Kim, W. and W. B. Whitman (1999). "Isolation of acetate auxotrophs of the methane-producing archaeon Methanococcus maripaludis by random insertional mutagenesis." Genetics **152**(4): 1429-1437.
Kimura, T. and H. Nishioka (1997). "Intracellular generation of superoxide by copper sulphate in Escherichia coli." Mutat Res **389**(2-3): 237-242.
Kitaoka, M., J. Tian, et al. (2005). "Novel putative galactose operon involving lacto-N-biose phosphorylase in Bifidobacterium longum." Appl Environ Microbiol **71**(6): 3158-3162.
Klebba, P. E. (2003). "Three paradoxes of ferric enterobactin uptake." Front Biosci **8**: s1422-1436.

Klotz, M. G., D. J. Arp, et al. (2006). "Complete genome sequence of the marine, chemolithoautotrophic, ammonia-oxidizing bacterium Nitrosococcus oceani ATCC 19707." Appl Environ Microbiol **72**(9): 6299-6315.
Klotz, M. G. and L. Y. Stein (2008). "Nitrifier genomics and evolution of the nitrogen cycle." FEMS Microbiol Lett **278**(2): 146-156.
Kneip, C., P. Lockhart, et al. (2007). "Nitrogen fixation in eukaryotes--new models for symbiosis." BMC Evol Biol **7**: 55.
Knietsch, A., T. Waschkowitz, et al. (2003). "Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on Escherichia coli." Appl Environ Microbiol **69**(3): 1408-1416.
Knoblauch, C., K. Sahm, et al. (1999). "Psychrophilic sulfate-reducing bacteria isolated from permanently cold arctic marine sediments: description of Desulfofrigus oceanense gen. nov., sp. nov., Desulfofrigus fragile sp. nov., Desulfofaba gelida gen. nov., sp. nov., Desulfotalea psychrophila gen. nov., sp. nov. and Desulfotalea arctica sp. nov." Int J Syst Bacteriol **49 Pt 4**: 1631-1643.
Koehler, T. M. (2009). "Bacillus anthracis physiology and genetics." Mol Aspects Med **30**(6): 386-396.
Kolosova, N., N. Gorenstein, et al. (2001). "Regulation of circadian methyl benzoate emission in diurnally and nocturnally emitting plants." Plant Cell **13**(10): 2333-2347.
Konstantinidis, K. T., J. Braff, et al. (2009). "Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre." Appl Environ Microbiol **75**(16): 5345-5355.
Konstantinidis, K. T. and J. M. Tiedje (2004). "Trends between gene content and genome size in prokaryotic species with larger genomes." Proc Natl Acad Sci U S A **101**(9): 3160-3165.
Kooistra, J. and G. Venema (1991). "Cloning, sequencing, and expression of Bacillus subtilis genes involved in ATP-dependent nuclease synthesis." J Bacteriol **173**(12): 3644-3655.
Kooistra, J., B. Vosman, et al. (1988). "Cloning and characterization of a Bacillus subtilis transcription unit involved in ATP-dependent DNase synthesis." J Bacteriol **170**(10): 4791-4797.
Kostka, J. E., O. Prakash, et al. (2011). "Hydrocarbon-degrading bacteria and the bacterial community response in Gulf of Mexico beach sands impacted by the Deepwater Horizon oil spill." Appl Environ Microbiol.
Kovacs, A. T., G. Rakhely, et al. (2005). "Anaerobic regulation of hydrogenase transcription in different bacteria." Biochem Soc Trans **33**(Pt 1): 36-38.
Krajmalnik-Brown, R., T. Holscher, et al. (2004). "Genetic identification of a putative vinyl chloride reductase in Dehalococcoides sp. strain BAV1." Appl Environ Microbiol **70**(10): 6347-6351.
Kral, T. A., C. R. Bekkum, et al. (2004). "Growth of methanogens on a Mars soil simulant." Orig Life Evol Biosph **34**(6): 615-626.
Kreuzer, K. N. and N. R. Cozzarelli (1979). "Escherichia coli mutants thermosensitive for deoxyribonucleic acid gyrase subunit A: effects on deoxyribonucleic acid replication, transcription, and bacteriophage growth." J Bacteriol **140**(2): 424-435.
Kristiansson, E., P. Hugenholtz, et al. (2009). "ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes." Bioinformatics **25**(20): 2737-2738.
Krzycki, J. A. (2004). "Function of genetically encoded pyrrolysine in corrinoid-dependent methylamine methyltransferases." Curr Opin Chem Biol **8**(5): 484-491.

Kunin, V., A. Engelbrektson, et al. (2010). "Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates." Environ Microbiol **12**(1): 118-123.
Kunin, V., J. Raes, et al. (2008). "Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat." Mol Syst Biol **4**: 198.
Kupsch, J., J. C. Alonso, et al. (1989). "Analysis of structural and biological parameters affecting plasmid deletion formation in Bacillus subtilis." Mol Gen Genet **218**(3): 402-408.
Lakay, F. M., A. Botha, et al. (2007). "Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils." J Appl Microbiol **102**(1): 265-273.
LaMontagne, M. G., F. C. Michel, Jr., et al. (2002). "Evaluation of extraction and purification methods for obtaining PCR-amplifiable DNA from compost for microbial community analysis." J Microbiol Methods **49**(3): 255-264.
Landis, G. A. (2001). "Martian water: are there extant halobacteria on Mars?" Astrobiology **1**(2): 161-164.
Landt, S. G., J. A. Lesley, et al. (2010). "CrfA, a small noncoding RNA regulator of adaptation to carbon starvation in Caulobacter crescentus." J Bacteriol **192**(18): 4763-4775.
Larimer, F. W., P. Chain, et al. (2004). "Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris." Nat Biotechnol **22**(1): 55-61.
Larose, C., S. Berger, et al. (2010). "Microbial sequences retrieved from environmental samples from seasonal arctic snow and meltwater from Svalbard, Norway." Extremophiles **14**(2): 205-212.
Lauber, C. L., M. Hamady, et al. (2009). "Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale." Appl Environ Microbiol **75**(15): 5111-5120.
Lauber, C. L., N. Zhou, et al. (2010). "Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples." FEMS Microbiol Lett **307**(1): 80-86.
Leahy, S. C., W. J. Kelly, et al. (2010). "The genome sequence of the rumen methanogen Methanobrevibacter ruminantium reveals new possibilities for controlling ruminant methane emissions." PLoS One **5**(1): e8926.
Lechevalier, H. A. and M. P. Lechevalier (1967). "Biology of actinomycetes." Annu Rev Microbiol **21**: 71-100.
Lee, S., E. Cho, et al. (2009). "Periplasmic glucans isolated from Proteobacteria." BMB Rep **42**(12): 769-775.
Lefevre, F., P. Robe, et al. (2008). "Drugs from hidden bugs: their discovery via untapped resources." Res Microbiol **159**(3): 153-161.
Leff, L. G., J. R. Dana, et al. (1995). "Comparison of methods of DNA extraction from stream sediments." Appl Environ Microbiol **61**(3): 1141-1143.
Legendre, G., F. Fay, et al. (2011). "Evaluation of antibacterial activity against Salmonella Enteritidis." J Microbiol **49**(3): 349-354.
Lenz, O., M. Ludwig, et al. (2010). "H2 conversion in the presence of O2 as performed by the membrane-bound [NiFe]-hydrogenase of Ralstonia eutropha." Chemphyschem **11**(6): 1107-1119.
Lessner, D. J., R. E. Parales, et al. (2003). "Expression of the nitroarene dioxygenase genes in Comamonas sp. strain JS765 and Acidovorax sp. strain JS42 is induced by multiple aromatic compounds." J Bacteriol **185**(13): 3895-3904.

Ley, R. E., F. Backhed, et al. (2005). "Obesity alters gut microbial ecology." Proc Natl Acad Sci U S A **102**(31): 11070-11075.
Ley, R. E., P. J. Turnbaugh, et al. (2006). "Microbial ecology: human gut microbes associated with obesity." Nature **444**(7122): 1022-1023.
Li, X. Z. and H. Nikaido (2004). "Efflux-mediated drug resistance in bacteria." Drugs **64**(2): 159-204.
Lidstrom, M. (2006). Aerobic Methylotrophic Prokaryotes. The Prokaryotes. M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer and E. Stackebrandt, Springer New York**:** 618-634.
Lipschultz, F., O. C. Zafiriou, et al. (1981). "Production of NO and N2O by soil nitrifying bacteria." Nature **294**(641-643): 641-643.
Liu, Y., F. Harnisch, et al. (2008). "Improvement of the anodic bioelectrocatalytic activity of mixed culture biofilms by a simple consecutive electrochemical selection procedure." Biosens Bioelectron **24**(4): 1012-1017.
Loh, J. and G. Stacey (2003). "Nodulation gene regulation in Bradyrhizobium japonicum: a unique integration of global regulatory circuits." Appl Environ Microbiol **69**(1): 10-17.
Lorber, B. (2005). "Treatment of brain abscess due to Listeria monocytogenes." Clin Infect Dis **41**(3): 419.
Lovley, D. R. and J. D. Coates (2000). "Novel forms of anaerobic respiration of environmental relevance." Curr Opin Microbiol **3**(3): 252-256.
Lovley, D. R., S. J. Giovannoni, et al. (1993). "Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals." Arch Microbiol **159**(4): 336-344.
Lowy, F. D. (1998). "Staphylococcus aureus infections." N Engl J Med **339**(8): 520-532.
Mackey, S. R. and S. S. Golden (2007). "Winding up the cyanobacterial circadian clock." Trends Microbiol **15**(9): 381-388.
Makarova, K. S., M. V. Omelchenko, et al. (2007). "Deinococcus geothermalis: the pool of extreme radiation resistance genes shrinks." PLoS One **2**(9): e955.
Mallozzi, M., V. K. Viswanathan, et al. (2010). "Spore-forming Bacilli and Clostridia in human disease." Future Microbiol **5**(7): 1109-1123.
Mandlik, A., A. Swierczynski, et al. (2008). "Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development." Trends Microbiol **16**(1): 33-40.
Maniloff, J. and H. W. Ackermann (1998). "Taxonomy of bacterial viruses: establishment of tailed virus genera and the order Caudovirales." Arch Virol **143**(10): 2051-2063.
Mao, H., S. A. Hart, et al. (2004). "Sortase-mediated protein ligation: a new method for protein engineering." J Am Chem Soc **126**(9): 2670-2671.
Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.
Markowitz, V. M., N. Ivanova, et al. (2006). "An experimental metagenome data management and analysis system." Bioinformatics **22**(14): e359-367.
Markowitz, V. M., N. N. Ivanova, et al. (2008). "IMG/M: a data management and analysis system for metagenomes." Nucleic Acids Res **36**(Database issue): D534-538.
Martin-Cuadrado, A. B., P. Lopez-Garcia, et al. (2007). "Metagenomics of the deep Mediterranean, a warm bathypelagic habitat." PLoS One **2**(9): e914.
Martin-Laurent, F., L. Philippot, et al. (2001). "DNA extraction from soils: old bias for new microbial diversity analysis methods." Appl Environ Microbiol **67**(5): 2354-2359.
Martinez, J. L., M. B. Sanchez, et al. (2009). "Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems." FEMS Microbiol Rev **33**(2): 430-449.

Maruyama, A., D. Honda, et al. (2000). "Phylogenetic analysis of psychrophilic bacteria isolated from the Japan Trench, including a description of the deep-sea species Psychrobacter pacificensis sp. nov." Int J Syst Evol Microbiol **50 Pt 2**: 835-846.
Mastrangelo, G., E. Fadda, et al. (1996). "Polycyclic aromatic hydrocarbons and cancer in man." Environ Health Perspect **104**(11): 1166-1170.
Mathes, A. and H. Engelhardt (1998). "Nonlinear and asymmetric open channel characteristics of an ion-selective porin in planar membranes." Biophys J **75**(3): 1255-1262.
Matsushiro, A. (1963). "Specialized transduction of tryptophan markers in Escherichia coli K12 by bacteriophage phi-80." Virology **19**: 475-482.
Maxwell, A. (1997). "DNA gyrase as a drug target." Trends Microbiol **5**(3): 102-109.
Mazmanian, S. K., C. H. Liu, et al. (2005). "An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system." Cell **122**(1): 107-118.
Mazmanian, S. K., J. L. Round, et al. (2008). "A microbial symbiosis factor prevents intestinal inflammatory disease." Nature **453**(7195): 620-625.
Medigue, C., E. Krin, et al. (2005). "Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125." Genome Res **15**(10): 1325-1335.
Meilleur, C., J. F. Hupe, et al. (2009). "Isolation and characterization of a new alkali-thermostable lipase cloned from a metagenomic library." J Ind Microbiol Biotechnol **36**(6): 853-861.
Meima, R., B. J. Haijema, et al. (1997). "Role of enzymes of homologous recombination in illegitimate plasmid recombination in Bacillus subtilis." J Bacteriol **179**(4): 1219-1229.
Meima, R., B. J. Haijema, et al. (1995). "Overproduction of the ATP-dependent nuclease AddAB improves the structural stability of a model plasmid system in Bacillus subtilis." Mol Gen Genet **248**(4): 391-398.
Meldrum, F. C., S. Mann, et al. (1993). "Electron Microscopy Study of Magnetosomes in a Cultured Coccoid Magnetotactic Bacterium." Proc. R. Soc. Lond. B **251**(1332): 231-236
Mendez, C. and J. A. Salas (2001). "Altering the glycosylation pattern of bioactive compounds." Trends Biotechnol **19**(11): 449-456.
Mendum, T. A., B. Z. Chilima, et al. (2000). "The PCR amplification of non-tuberculous mycobacterial 16S rRNA sequences from soil." FEMS Microbiol Lett **185**(2): 189-192.
Mergeay, M., C. Houba, et al. (1978). "Extrachromosomal inheritance controlling resistance to cadmium, cobalt, copper and zinc ions: evidence from curing in a Pseudomonas [proceedings]." Arch Int Physiol Biochim **86**(2): 440-442.
Mergeay, M., S. Monchy, et al. (2003). "Ralstonia metallidurans, a bacterium specifically adapted to toxic metals: towards a catalogue of metal-responsive genes." FEMS Microbiol Rev **27**(2-3): 385-410.
Messens, J. and S. Silver (2006). "Arsenate reduction: thiol cascade chemistry with convergent evolution." J Mol Biol **362**(1): 1-17.
Meyer, F., D. Paarmann, et al. (2008). "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes." BMC Bioinformatics **9**: 386.
Miki, T., J. A. Park, et al. (1992). "Control of segregation of chromosomal DNA by sex factor F in Escherichia coli. Mutants of DNA gyrase subunit A suppress letD (ccdB) product growth inhibition." J Mol Biol **225**(1): 39-52.

Miller, T. L., M. J. Wolin, et al. (1986). "Characteristics of methanogens isolated from bovine rumen." Appl Environ Microbiol **51**(1): 201-202.
Miller, T. R., A. L. Delcher, et al. (2010). "Genome sequence of the dioxin-mineralizing bacterium Sphingomonas wittichii RW1." J Bacteriol **192**(22): 6101-6102.
Mitra, S., B. Klar, et al. (2009). "Visual and statistical comparison of metagenomes." Bioinformatics **25**(15): 1849-1855.
Mitsui, A., S. Kumazawa, et al. (1986). "Strategy by which nitrogen-fixing unicellular cyanobacteria grow photoautotrophically." Nature **323**(6090): 720-722.
Mizuuchi, K., M. H. O'Dea, et al. (1978). "DNA gyrase: subunit structure and ATPase activity of the purified enzyme." Proc Natl Acad Sci U S A **75**(12): 5960-5963.
Moken, M. C., L. M. McMurry, et al. (1997). "Selection of multiple-antibiotic-resistant (mar) mutants of Escherichia coli by using the disinfectant pine oil: roles of the mar and acrAB loci." Antimicrob Agents Chemother **41**(12): 2770-2772.
Moracci, M., B. Cobucci Ponzano, et al. (2000). "Identification and molecular characterization of the first alpha -xylosidase from an archaeon." J Biol Chem **275**(29): 22082-22089.
Morales, S. E., T. F. Cosart, et al. (2009). "Extensive phylogenetic analysis of a soil bacterial community illustrates extreme taxon evenness and the effects of amplicon length, degree of coverage, and DNA fractionation on classification and ecological parameters." Appl Environ Microbiol **75**(3): 668-675.
Morales, S. E. and W. E. Holben (2009). "Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by in silico analysis." Appl Environ Microbiol **75**(9): 2677-2683.
Morris, C. E., M. Bardin, et al. (2002). "Microbial biodiversity: approaches to experimental design and hypothesis testing in primary scientific literature from 1975 to 1999." Microbiol Mol Biol Rev **66**(4): 592-616, table of contents.
Muller, R. H. and W. Babel (2004). "Delftia acidovorans MC1 resists high herbicide concentrations--a study of nutristat growth on (RS)-2-(2,4-Dichlorophenoxy)propionate and 2,4-dichlorophenoxyacetate." Biosci Biotechnol Biochem **68**(3): 622-630.
Muller, R. H., S. Jorks, et al. (1999). "Comamonas acidovorans strain MC1: a new isolate capable of degrading the chiral herbicides dichlorprop and mecoprop and the herbicides 2,4-D and MCPA." Microbiol Res **154**(3): 241-246.
Murray, R. G. and S. W. Watson (1965). "Structure of Nitrosocystis Oceanus and Comparison with Nitrosomonas and Nitrobacter." J Bacteriol **89**: 1594-1609.
Mushtaq, N., M. Ezzati, et al. (2011). "Adhesion of Streptococcus pneumoniae to human airway epithelial cells exposed to urban particulate matter." J Allergy Clin Immunol **127**(5): 1236-1242 e1232.
Nacke, H., C. Will, et al. "Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German Biodiversity Exploratories." FEMS Microbiol Ecol.
Narasingarao, P., S. Podell, et al. (2011). "De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities." ISME J.
Nascimento, A. M. and E. Chartone-Souza (2003). "Operon mer: bacterial resistance to mercury and potential for bioremediation of contaminated environments." Genet Mol Res **2**(1): 92-101.
Nathan, C. (1992). "Nitric oxide as a secretory product of mammalian cells." FASEB J **6**(12): 3051-3064.

Navarre, W. W. and O. Schneewind (1999). "Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope." Microbiol Mol Biol Rev **63**(1): 174-229.
Nealson, K. H. and J. C. Venter (2007). "Metagenomics and the global ocean survey: what's in it for us, and why should we care?" ISME J **1**(3): 185-187.
Nelson, K. E., R. D. Fleischmann, et al. (2003). "Complete genome sequence of the oral pathogenic Bacterium porphyromonas gingivalis strain W83." J Bacteriol **185**(18): 5591-5601.
Newell, C. A. (2000). "Plant transformation technology. Developments and applications." Mol Biotechnol **16**(1): 53-65.
Ng, W. V., S. P. Kennedy, et al. (2000). "Genome sequence of Halobacterium species NRC-1." Proc Natl Acad Sci U S A **97**(22): 12176-12181.
Nicholson, W. L., P. Fajardo-Cavazos, et al. (2010). "Exploring the low-pressure growth limit: evolution of Bacillus subtilis in the laboratory to enhanced growth at 5 kilopascals." Appl Environ Microbiol **76**(22): 7559-7565.
Nies, D., M. Mergeay, et al. (1987). "Cloning of plasmid genes encoding resistance to cadmium, zinc, and cobalt in Alcaligenes eutrophus CH34." J Bacteriol **169**(10): 4865-4868.
Nies, D. H. (1992). "Resistance to cadmium, cobalt, zinc, and nickel in microbes." Plasmid **27**(1): 17-28.
Niu, B., L. Fu, et al. (2010). "Artificial and natural duplicates in pyrosequencing reads of metagenomic data." BMC Bioinformatics **11**: 187.
Noguchi, H., J. Park, et al. (2006). "MetaGene: prokaryotic gene finding from environmental genome shotgun sequences." Nucleic Acids Res **34**(19): 5623-5630.
Noonan, J. P., G. Coop, et al. (2006). "Sequencing and analysis of Neanderthal genomic DNA." Science **314**(5802): 1113-1118.
O'Brien, J. M., R. H. Wolkin, et al. (1984). "Association of hydrogen metabolism with unitrophic or mixotrophic growth of Methanosarcina barkeri on carbon monoxide." J Bacteriol **158**(1): 373-375.
Oelgeschlager, E. and M. Rother (2008). "Carbon monoxide-dependent energy metabolism in anaerobic bacteria and archaea." Arch Microbiol **190**(3): 257-269.
Oishi, M. (1969). "An ATP-dependent deoxyribonuclease from Escherichia coli with a possible role in genetic recombination." Proc Natl Acad Sci U S A **64**(4): 1292-1299.
Olsen, G. J., D. J. Lane, et al. (1986). "Microbial ecology and evolution: a ribosomal RNA approach." Annu Rev Microbiol **40**: 337-365.
Oltmann, L., E. van der Beek, et al. (1975). "Reduction of inorganic sulphur compounds by facultatively aerobic bacteria." Plant and Soil **43**(1): 153-169.
Oppermann, H., A. D. Levinson, et al. (1981). "The structure and protein kinase activity of proteins encoded by nonconditional mutants and back mutants in the sec gene of avian sarcoma virus." Virology **108**(1): 47-70.
Ortiz-Castro, R., H. A. Contreras-Cornejo, et al. (2009). "The role of microbial signals in plant growth and development." Plant Signal Behav **4**(8): 701-712.
Outten, F. W., D. L. Huffman, et al. (2001). "The independent cue and cus systems confer copper tolerance during aerobic and anaerobic growth in Escherichia coli." J Biol Chem **276**(33): 30670-30677.
Overbeek, R., T. Begley, et al. (2005). "The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." Nucleic Acids Res **33**(17): 5691-5702.
Overmann, J. (2010). "The phototrophic consortium "Chlorochromatium aggregatum" - a model for bacterial heterologous multicellularity." Adv Exp Med Biol **675**: 15-29.

Palas, K. M. and S. R. Kushner (1990). "Biochemical and physical characterization of exonuclease V from Escherichia coli. Comparison of the catalytic activities of the RecBC and RecBCD enzymes." J Biol Chem **265**(6): 3447-3454.
Papa, R., E. Parrilli, et al. (2009). "Engineered marine Antarctic bacterium Pseudoalteromonas haloplanktis TAC125: a promising micro-organism for the bioremediation of aromatic compounds." J Appl Microbiol **106**(1): 49-56.
Papavassiliou, J., V. Samaraki-Lyberopoulou, et al. (1969). "Production of tetrathionate reductase by Salmonella." Can J Microbiol **15**(2): 238-240.
Parkhill, J., B. W. Wren, et al. (2000). "The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences." Nature **403**(6770): 665-668.
Parks, D. H. and R. G. Beiko (2010). "Identifying biologically relevant differences between metagenomic communities." Bioinformatics **26**(6): 715-721.
Parsons, R. and R. J. Sunley (2001). "Nitrogen nutrition and the role of root-shoot nitrogen signalling particularly in symbiotic systems." J Exp Bot **52**(Spec Issue): 435-443.
Patankar, A. V. and J. E. Gonzalez (2009). "Orphan LuxR regulators of quorum sensing." FEMS Microbiol Rev **33**(4): 739-756.
Patel, M. R., A. Berces, et al. (2004). "Annual solar UV exposure and biological effective dose rates on the Martian surface." Adv Space Res **33**(8): 1247-1252.
Peijnenburg, A. A., S. Bron, et al. (1987). "Structural plasmid instability in recombination- and repair-deficient strains of Bacillus subtilis." Plasmid **17**(2): 167-170.
Pesaro, M., G. Nicollier, et al. (2004). "Impact of soil drying-rewetting stress on microbial communities and activities and on degradation of two crop protection products." Appl Environ Microbiol **70**(5): 2577-2587.
Peterson, R. B. and C. P. Wolk (1978). "High recovery of nitrogenase activity and of Fe-labeled nitrogenase in heterocysts isolated from Anabaena variabilis." Proc Natl Acad Sci U S A **75**(12): 6271-6275.
Phadtare, S. (2004). "Recent developments in bacterial cold-shock response." Curr Issues Mol Biol **6**(2): 125-136.
Phale, P. S., A. Basu, et al. (2007). "Metabolic diversity in bacterial degradation of aromatic compounds." OMICS **11**(3): 252-279.
Pham, C. A., S. J. Jung, et al. (2003). "A novel electrochemically active and Fe(III)-reducing bacterium phylogenetically related to Aeromonas hydrophila, isolated from a microbial fuel cell." FEMS Microbiol Lett **223**(1): 129-134.
Phillips, J. E. (1961). "The commensal role of Actinobacillus lignieresi." J Pathol Bacteriol **82**: 205-208.
Piddock, L. J. (2006). "Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria." Clin Microbiol Rev **19**(2): 382-402.
Pierson, B. K. and R. W. Castenholz (1974). "A phototrophic gliding filamentous bacterium of hot springs, Chloroflexus aurantiacus, gen. and sp. nov." Arch Microbiol **100**(1): 5-24.
Pittard, J. (1964). "Effect of phage-controlled restriction on genetic linkage in bacterial crosses." J Bacteriol **87**(5): 1256-1257.
Pohlmann, A., W. F. Fricke, et al. (2006). "Genome sequence of the bioplastic-producing "Knallgas" bacterium Ralstonia eutropha H16." Nat Biotechnol **24**(10): 1257-1262.
Poinar, H. N., C. Schwarz, et al. (2006). "Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA." Science **311**(5759): 392-394.
Pointing, S. B., Y. Chan, et al. (2009). "Highly specialized microbial diversity in hyper-arid polar desert." Proc Natl Acad Sci U S A **106**(47): 19964-19969.
Pollock, M. R., R. Knox, et al. (1942). "Bacterial Reduction of Tetrathionate." Nature(150): 94-94.

Poole, K. (2007). "Efflux pumps as antimicrobial resistance mechanisms." Ann Med **39**(3): 162-176.
Poole, L. B. (2005). "Bacterial defenses against oxidants: mechanistic features of cysteine-based peroxidases and their flavoprotein reductases." Arch Biochem Biophys **433**(1): 240-254.
Porter, J. R. (1976). "Antony van Leeuwenhoek: tercentenary of his discovery of bacteria." Bacteriol Rev **40**(2): 260-269.
Postgate, J. R. (1970). "Biological nitrogen fixation." Nature **226**(5240): 25-27.
Potrykus, K. and M. Cashel (2008). "(p)ppGpp: still magical?" Annu Rev Microbiol **62**: 35-51.
Potter, M. C. (1911). "Electrical Effects Accompanying the Decomposition of Organic Compounds." Proc. R. Soc. Lond.(84): 260-276.
Potts, M. (1994). "Desiccation tolerance of prokaryotes." Microbiol Rev **58**(4): 755-805.
Prade, R. A. (1996). "Xylanases: from biology to biotechnology." Biotechnol Genet Eng Rev **13**: 101-131.
Pringault, O., H. Viret, et al. (2011). "Interactions between Zn and bacteria in marine tropical coastal sediments." Environ Sci Pollut Res Int.
Prosser, J. I. (2010). "Replicate or lie." Environ Microbiol **12**(7): 1806-1810.
Pumbwe, L., C. A. Skilbeck, et al. (2007). "Induction of multiple antibiotic resistance in Bacteroides fragilis by benzene and benzene-derived active compounds of commonly used analgesics, antiseptics and cleaning agents." J Antimicrob Chemother **60**(6): 1288-1297.
Qin, J., R. Li, et al. (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.
Rabaey, K., N. Boon, et al. (2004). "Biofuel cells select for microbial consortia that self-mediate electron transfer." Appl Environ Microbiol **70**(9): 5373-5382.
Rabaey, K., S. T. Read, et al. (2008). "Cathodic oxygen reduction catalyzed by bacteria in microbial fuel cells." ISME J **2**(5): 519-527.
Rabus, R., M. Kube, et al. (2005). "The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1." Arch Microbiol **183**(1): 27-36.
Rabus, R., A. Ruepp, et al. (2004). "The genome of Desulfotalea psychrophila, a sulfate-reducing bacterium from permanently cold Arctic sediments." Environ Microbiol **6**(9): 887-902.
Ragsdale, S. W. (1991). "Enzymology of the acetyl-CoA pathway of CO2 fixation." Crit Rev Biochem Mol Biol **26**(3-4): 261-300.
Rajendhran, J. and P. Gunasekaran (2008). "Strategies for accessing soil metagenome for desired applications." Biotechnol Adv **26**(6): 576-590.
Ramirez, K. S., C. L. Lauber, et al. (2010). "Consistent effects of nitrogen fertilization on soil bacterial communities in contrasting systems." Ecology **91**(12): 3463-3470; discussion 3503-3414.
Ramos, J. L., E. Duque, et al. (2002). "Mechanisms of solvent tolerance in gram-negative bacteria." Annu Rev Microbiol **56**: 743-768.
Randall, L. P. and M. J. Woodward (2001). "Multiple antibiotic resistance (mar) locus in Salmonella enterica serovar typhimurium DT104." Appl Environ Microbiol **67**(3): 1190-1197.
Randall, L. P. and M. J. Woodward (2002). "The multiple antibiotic resistance (mar) locus and its significance." Res Vet Sci **72**(2): 87-93.
Ranjard, L., E. Brothier, et al. (2000). "Sequencing bands of ribosomal intergenic spacer analysis fingerprints for characterization and microscale distribution of soil bacterium

populations responding to mercury spiking." Appl Environ Microbiol **66**(12): 5334-5339.
Raoult, D., S. Audic, et al. (2004). "The 1.2-megabase genome sequence of Mimivirus." Science **306**(5700): 1344-1350.
Rappe, M. S. and S. J. Giovannoni (2003). "The uncultured microbial majority." Annu Rev Microbiol **57**: 369-394.
Rawlings, D. E. (1999). "Proteic toxin-antitoxin, bacterial plasmid addiction systems and their evolution with special reference to the pas system of pTF-FC2." FEMS Microbiol Lett **176**(2): 269-277.
Regnier, P. and E. Hajnsdorf (2009). "Poly(A)-assisted RNA decay and modulators of RNA stability." Prog Mol Biol Transl Sci **85**: 137-185.
Reid, G. (2004). "When microbe meets human." Clin Infect Dis **39**(6): 827-830.
Revsbech, N. P., B. Thamdrup, et al. (2011). "Construction of STOX oxygen sensors and their application for determination of O2 concentrations in oxygen minimum zones." Methods Enzymol **486**: 325-341.
Riley, M., J. T. Staley, et al. (2008). "Genomics of an extreme psychrophile, Psychromonas ingrahamii." BMC Genomics **9**: 210.
Risso, C., J. Sun, et al. (2009). "Genome-scale comparison and constraint-based metabolic reconstruction of the facultative anaerobic Fe(III)-reducer Rhodoferax ferrireducens." BMC Genomics **10**: 447.
Rodionov, D. A., P. Hebbeln, et al. (2009). "A novel class of modular transporters for vitamins in prokaryotes." J Bacteriol **191**(1): 42-51.
Rodionov, D. A., P. Hebbeln, et al. (2006). "Comparative and functional genomic analysis of prokaryotic nickel and cobalt uptake transporters: evidence for a novel group of ATP-binding cassette transporters." J Bacteriol **188**(1): 317-327.
Roesch, L. F., R. R. Fulthorpe, et al. (2007). "Pyrosequencing enumerates and contrasts soil microbial diversity." ISME J **1**(4): 283-290.
Rohlin, L. and R. P. Gunsalus (2010). "Carbon-dependent control of electron transfer and central carbon pathway genes for methane biosynthesis in the Archaean, Methanosarcina acetivorans strain C2A." BMC Microbiol **10**: 62.
Rohwer, F., V. Seguritan, et al. (2001). "Production of shotgun libraries using random amplification." Biotechniques **31**(1): 108-112, 114-106, 118.
Rokyta, D. R., C. L. Burch, et al. (2006). "Horizontal gene transfer and the evolution of microvirid coliphage genomes." J Bacteriol **188**(3): 1134-1142.
Romanenko, L. A., P. Schumann, et al. (2002). "Psychrobacter submarinus sp. nov. and Psychrobacter marincola sp. nov., psychrophilic halophiles from marine environments." Int J Syst Evol Microbiol **52**(Pt 4): 1291-1297.
Rondon, M. R., P. R. August, et al. (2000). "Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms." Appl Environ Microbiol **66**(6): 2541-2547.
Ronto, G., A. Berces, et al. (2003). "Solar UV irradiation conditions on the surface of Mars." Photochem Photobiol **77**(1): 34-40.
Rosado, P., L. Gallego, et al. (2009). "Lemierre's syndrome: a serious complication of an odontogenic infection." Med Oral Patol Oral Cir Bucal **14**(8): e398-401.
Rother, M. and W. W. Metcalf (2004). "Anaerobic growth of Methanosarcina acetivorans C2A on carbon monoxide: an unusual way of life for a methanogenic archaeon." Proc Natl Acad Sci U S A **101**(48): 16929-16934.
Rouch, D. A., B. T. Lee, et al. (1995). "Understanding cellular responses to toxic agents: a model for mechanism-choice in bacterial metal resistance." J Ind Microbiol **14**(2): 132-141.

Rousk, J., E. Baath, et al. (2010). "Soil bacterial and fungal communities across a pH gradient in an arable soil." ISME J **4**(10): 1340-1351.
Ruepp, A., W. Graml, et al. (2000). "The genome sequence of the thermoacidophilic scavenger Thermoplasma acidophilum." Nature **407**(6803): 508-513.
Ryder, C., M. Byrd, et al. (2007). "Role of polysaccharides in Pseudomonas aeruginosa biofilm development." Curr Opin Microbiol **10**(6): 644-648.
Sa-Nogueira, I., T. V. Nogueira, et al. (1997). "The Bacillus subtilis L-arabinose (ara) operon: nucleotide sequence, genetic organization and expression." Microbiology **143 (Pt 3)**: 957-969.
Sader, H. S. and R. N. Jones (2005). "Antimicrobial susceptibility of uncommonly isolated non-enteric Gram-negative bacilli." Int J Antimicrob Agents **25**(2): 95-109.
Sakamoto, M., M. Kitahara, et al. (2007). "Parabacteroides johnsonii sp. nov., isolated from human faeces." Int J Syst Evol Microbiol **57**(Pt 2): 293-296.
Salyers, A. A., N. B. Shoemaker, et al. (1995). "In the driver's seat: the Bacteroides conjugative transposons and the elements they mobilize." J Bacteriol **177**(20): 5727-5731.
Salyers, A. A., N. B. Shoemaker, et al. (1995). "Conjugative transposons: an unusual and diverse set of integrated gene transfer elements." Microbiol Rev **59**(4): 579-590.
Sanguin, H., B. Remenant, et al. (2006). "Potential of a 16S rRNA-based taxonomic microarray for analyzing the rhizosphere effects of maize on Agrobacterium spp. and bacterial communities." Appl Environ Microbiol **72**(6): 4302-4312.
Saunders, J. R. (1984). "Genetics and evolution of antibiotic resistance." Br Med Bull **40**(1): 54-60.
Savage, D. C. (1977). "Microbial ecology of the gastrointestinal tract." Annu Rev Microbiol **31**: 107-133.
Schleheck, D., T. P. Knepper, et al. (2007). "Parvibaculum lavamentivorans DS-1T degrades centrally substituted congeners of commercial linear alkylbenzenesulfonate to sulfophenyl carboxylates and sulfophenyl dicarboxylates." Appl Environ Microbiol **73**(15): 4725-4732.
Schleheck, D., B. J. Tindall, et al. (2004). "Parvibaculum lavamentivorans gen. nov., sp. nov., a novel heterotroph that initiates catabolism of linear alkylbenzenesulfonate." Int J Syst Evol Microbiol **54**(Pt 5): 1489-1497.
Schleif, R. (2000). "Regulation of the L-arabinose operon of Escherichia coli." Trends Genet **16**(12): 559-565.
Schleper, C., G. Puehler, et al. (1995). "Picrophilus gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0." J Bacteriol **177**(24): 7050-7059.
Schlesner, H., C. Rensmann, et al. (2004). "Taxonomic heterogeneity within the Planctomycetales as derived by DNA-DNA hybridization, description of Rhodopirellula baltica gen. nov., sp. nov., transfer of Pirellula marina to the genus Blastopirellula gen. nov. as Blastopirellula marina comb. nov. and emended description of the genus Pirellula." Int J Syst Evol Microbiol **54**(Pt 5): 1567-1580.
Schloss, P. D. and J. Handelsman (2003). "Biotechnological prospects from metagenomics." Curr Opin Biotechnol **14**(3): 303-310.
Schloss, P. D. and J. Handelsman (2005). "Metagenomics for studying unculturable microorganisms: cutting the Gordian knot." Genome Biol **6**(8): 229.
Schonert, S., S. Seitz, et al. (2006). "Maltose and maltodextrin utilization by Bacillus subtilis." J Bacteriol **188**(11): 3911-3922.
Scott, J. R. and D. Zahner (2006). "Pili with strong attachments: Gram-positive bacteria do it differently." Mol Microbiol **62**(2): 320-330.

Seah, S. Y., J. Ke, et al. (2007). "Characterization of a C-C bond hydrolase from Sphingomonas wittichii RW1 with novel specificities towards polychlorinated biphenyl metabolites." J Bacteriol **189**(11): 4038-4045.
Sebat, J. L., F. S. Colwell, et al. (2003). "Metagenomic profiling: microarray analysis of an environmental genomic library." Appl Environ Microbiol **69**(8): 4927-4934.
Seshadri, R., S. W. Joseph, et al. (2006). "Genome sequence of Aeromonas hydrophila ATCC 7966T: jack of all trades." J Bacteriol **188**(23): 8272-8282.
Seshadri, R., S. A. Kravitz, et al. (2007). "CAMERA: a community resource for metagenomics." PLoS Biol **5**(3): e75.
Sessitsch, A., S. Gyamfi, et al. (2002). "RNA isolation from soil for bacterial community and functional analysis: evaluation of different extraction and soil conservation protocols." J Microbiol Methods **51**(2): 171-179.
Sevcenco, A. M., L. E. Bevers, et al. (2010). "Molybdenum incorporation in tungsten aldehyde oxidoreductase enzymes from Pyrococcus furiosus." J Bacteriol **192**(16): 4143-4152.
Sharma, S., F. G. Khan, et al. "Molecular cloning and characterization of amylase from soil metagenomic library derived from Northwestern Himalayas." Appl Microbiol Biotechnol **86**(6): 1821-1828.
Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.
Shoemaker, N. B., H. Vlamakis, et al. (2001). "Evidence for extensive resistance gene transfer among Bacteroides spp. and among Bacteroides and other genera in the human colon." Appl Environ Microbiol **67**(2): 561-568.
Siefert, J. L. (2009). "Defining the mobilome." Methods Mol Biol **532**: 13-27.
Silver, S. (1992). "Plasmid-determined metal resistance mechanisms: range and overview." Plasmid **27**(1): 1-3.
Silver, S. and T. Phung le (2005). "A bacterial view of the periodic table: genes and proteins for toxic inorganic ions." J Ind Microbiol Biotechnol **32**(11-12): 587-605.
Silver, S. and L. T. Phung (1996). "Bacterial heavy metal resistance: new surprises." Annu Rev Microbiol **50**: 753-789.
Sirevag, R. and R. Castenholz (1979). "Aspects of carbon metabolism in Chloroflexus." Archives of Microbiology **120**(2): 151-153.
Sleator, R. D., C. Shortall, et al. (2008). "Metagenomics." Lett Appl Microbiol **47**(5): 361-366.
Slesarev, A. I., J. A. Lake, et al. (1994). "Purification and characterization of DNA topoisomerase V. An enzyme from the hyperthermophilic prokaryote Methanopyrus kandleri that resembles eukaryotic topoisomerase I." J Biol Chem **269**(5): 3295-3303.
Smart, J. P., M. J. Cliff, et al. (2009). "A role for tungsten in the biology of Campylobacter jejuni: tungstate stimulates formate dehydrogenase activity and is transported via an ultra-high affinity ABC system distinct from the molybdate transporter." Mol Microbiol **74**(3): 742-757.
Smith, C. J., G. D. Tribble, et al. (1998). "Genetic elements of Bacteroides species: a moving story." Plasmid **40**(1): 12-29.
Smith, D. J., A. C. Schuerger, et al. (2009). "Survivability of Psychrobacter cryohalolentis K5 under simulated martian surface conditions." Astrobiology **9**(2): 221-228.
Smith, E. F. and C. O. Townsend (1907). "A Plant-Tumor of Bacterial Origin." Science **25**(643): 671-673.
Sogin, M. L., H. G. Morrison, et al. (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"." Proc Natl Acad Sci U S A **103**(32): 12115-12120.

Solomon, S. C., O. Aharonson, et al. (2005). "New perspectives on ancient Mars." Science **307**(5713): 1214-1220.
Sowers, K. R., S. F. Baron, et al. (1984). "Methanosarcina acetivorans sp. nov., an Acetotrophic Methane-Producing Bacterium Isolated from Marine Sediments." Appl Environ Microbiol **47**(5): 971-978.
Spaink, H. P. (2000). "Root nodulation and infection factors produced by rhizobial bacteria." Annu Rev Microbiol **54**: 257-288.
Spormann, A. M. (1999). "Gliding motility in bacteria: insights from studies of Myxococcus xanthus." Microbiol Mol Biol Rev **63**(3): 621-641.
Sram, R. J., B. Binkova, et al. (1999). "Adverse reproductive outcomes from exposure to environmental mutagens." Mutat Res **428**(1-2): 203-215.
Srinivasan, G., C. M. James, et al. (2002). "Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA." Science **296**(5572): 1459-1462.
Sugihara, T. F., L. Kline, et al. (1971). "Microorganisms of the San Francisco sour dough bread process. I. Yeasts responsible for the leavening action." Appl Microbiol **21**(3): 456-458.
Sunda, W., D. J. Kieber, et al. (2002). "An antioxidant function for DMSP and DMS in marine algae." Nature **418**(6895): 317-320.
Sunna, A. and G. Antranikian (1997). "Xylanolytic enzymes from fungi and bacteria." Crit Rev Biotechnol **17**(1): 39-67.
Suree, N., S. W. Yi, et al. (2009). "Discovery and structure-activity relationship analysis of Staphylococcus aureus sortase A inhibitors." Bioorg Med Chem **17**(20): 7174-7185.
Takahashi, Y. and U. Tokumoto (2002). "A third bacterial system for the assembly of iron-sulfur clusters with homologs in archaea and plastids." J Biol Chem **277**(32): 28380-28383.
Takai, K., D. P. Moser, et al. (2001). "Alkaliphilus transvaalensis gen. nov., sp. nov., an extremely alkaliphilic bacterium isolated from a deep South African gold mine." Int J Syst Evol Microbiol **51**(Pt 4): 1245-1256.
Tavares, P., A. S. Pereira, et al. (2006). "Metalloenzymes of the denitrification pathway." J Inorg Biochem **100**(12): 2087-2100.
Ternan, N. G., J. W. Mc Grath, et al. (1998). "Review: Organophosphonates: occurrence, synthesis and biodegradation by microorganisms." World Journal of Microbiology and Biotechnology **14**(5): 635-647.
Thioulouse, J. (1989). "Statistical analysis and graphical display of multivariate data on the Macintosh." Comput Appl Biosci **5**(4): 287-292.
Thomas, S. B. (1969). "Methods of assessing the psychrotrophic bacterial content of milk." J Appl Bacteriol **32**(3): 269-296.
Timoney, J. F. (2004). "The pathogenic equine streptococci." Vet Res **35**(4): 397-409.
Tolbert, N. E. (1997). "The C2 Oxidative Photosynthetic Carbon Cycle." Annu Rev Plant Physiol Plant Mol Biol **48**: 1-25.
Tomlinson, A. D. and C. Fuqua (2009). "Mechanisms and regulation of polar surface attachment in Agrobacterium tumefaciens." Curr Opin Microbiol **12**(6): 708-714.
Ton-That, H., G. Liu, et al. (1999). "Purification and characterization of sortase, the transpeptidase that cleaves surface proteins of Staphylococcus aureus at the LPXTG motif." Proc Natl Acad Sci U S A **96**(22): 12424-12429.
Topping, J. F., W. Wei, et al. (1995). "Agrobacterium-mediated transformation of Arabidopsis thaliana. Application in T-DNA tagging." Methods Mol Biol **49**: 63-76.
Torres-Cortes, G., V. Millan, et al. "Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples." Environ Microbiol **13**(4): 1101-1114.

Torsvik, V., J. Goksoyr, et al. (1990). "High diversity in DNA of soil bacteria." Appl Environ Microbiol **56**(3): 782-787.
Torsvik, V., L. Ovreas, et al. (2002). "Prokaryotic diversity--magnitude, dynamics, and controlling factors." Science **296**(5570): 1064-1066.
Tottey, S., D. R. Harvie, et al. (2005). "Understanding how cells allocate metals using metal sensors and metallochaperones." Acc Chem Res **38**(10): 775-783.
Tringe, S. G., C. von Mering, et al. (2005). "Comparative metagenomics of microbial communities." Science **308**(5721): 554-557.
Tringe, S. G., T. Zhang, et al. (2008). "The airborne metagenome in an indoor urban environment." PLoS One **3**(4): e1862.
Tripp, H. J., J. B. Kitner, et al. (2008). "SAR11 marine bacteria require exogenous reduced sulphur for growth." Nature **452**(7188): 741-744.
Troy, E. B. and D. L. Kasper (2010). "Beneficial effects of Bacteroides fragilis polysaccharides on the immune system." Front Biosci **15**: 25-34.
Tsang, P. H., G. Li, et al. (2006). "Adhesion of single bacterial cells in the micronewton range." Proc Natl Acad Sci U S A **103**(15): 5764-5768.
Turnbaugh, P. J. and J. I. Gordon (2009). "The core gut microbiome, energy balance and obesity." J Physiol **587**(Pt 17): 4153-4158.
Turnbaugh, P. J., R. E. Ley, et al. (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest." Nature **444**(7122): 1027-1031.
Tutino, M. L., A. Duilio, et al. (2001). "A novel replication element from an Antarctic plasmid as a tool for the expression of proteins at low temperature." Extremophiles **5**(4): 257-264.
Tuttle, J. H. (1980). "Thiosulfate Oxidation and Tetrathionate Reduction by Intact Cells of Marine Pseudomonad Strain 16B." Appl Environ Microbiol **39**(6): 1159-1166.
Tyson, G. W., J. Chapman, et al. (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature **428**(6978): 37-43.
Ulrich, R., T. Kral, et al. (2010). "Dynamic temperature fields under Mars landing sites and implications for supporting microbial life." Astrobiology **10**(6): 643-650.
van Belkum, A., D. C. Melles, et al. (2009). "Co-evolutionary aspects of human colonisation and infection by Staphylococcus aureus." Infect Genet Evol **9**(1): 32-47.
van de, V., A. J. Driessen, et al. (1998). "Bioenergetics and cytoplasmic membrane stability of the extremely acidophilic, thermophilic archaeon Picrophilus oshimae." Extremophiles **2**(2): 67-74.
Van der Werf, M. J., M. V. Guettler, et al. (1997). "Environmental and physiological factors affecting the succinate product ratio during carbohydrate fermentation by Actinobacillus sp. 130Z." Arch Microbiol **167**(6): 332-342.
Van Eldere, J. (2003). "Multicentre surveillance of Pseudomonas aeruginosa susceptibility patterns in nosocomial infections." J Antimicrob Chemother **51**(2): 347-352.
van Elsas, J. D., R. Costa, et al. (2008). "The metagenomics of disease-suppressive soils - experiences from the METACONTROL project." Trends Biotechnol **26**(11): 591-601.
Van Melderen, L. (2010). "Toxin-antitoxin systems: why so many, what for?" Curr Opin Microbiol **13**(6): 781-785.
Venter, J. C., K. Remington, et al. (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.
Vial, L., M. C. Groleau, et al. (2007). "Burkholderia diversity and versatility: an inventory of the extracellular products." J Microbiol Biotechnol **17**(9): 1407-1429.
Vijayan, A. P., M. R. Anand, et al. (2009). "Chromobacterium violaceum sepsis in an infant." Indian Pediatr **46**(8): 721-722.

Villemur, R., M. Lanthier, et al. (2006). "The Desulfitobacterium genus." FEMS Microbiol Rev **30**(5): 706-733.
Vladimirov, N. and V. Sourjik (2009). "Chemotaxis: how bacteria use memory." Biol Chem **390**(11): 1097-1104.
Vogel, T. M., P. Simonet, et al. (2010). TerraGenome: a consortium for the sequencing of a soil metagenome. Nat Rev Microbiol. **7:** 252.
Vogel, T. R., V. Y. Dombrovskiy, et al. (2009). "Has the implementation of EVAR for ruptured AAA improved outcomes?" Vasc Endovascular Surg **43**(3): 252-257.
Vreeland, R. H. and L. H. Hochstein (1993). "Biology of halophilic bacteria." CRC Press, Boca Raton, FL.
Wagner, J. K., S. Setayeshgar, et al. (2006). "A nutrient uptake role for bacterial cell envelope extensions." Proc Natl Acad Sci U S A **103**(31): 11772-11777.
Walcott, R. R., R. D. Gitaitis, et al. (2003). "Role of Blossoms in Watermelon Seed Infestation by Acidovorax avenae subsp. citrulli." Phytopathology **93**(5): 528-534.
Wallenius, K., H. Rita, et al. "Sample storage for soil enzyme activity and bacterial community profiles." J Microbiol Methods **81**(1): 48-55.
Wang, J. C. (1996). "DNA topoisomerases." Annu Rev Biochem **65**: 635-692.
Wang, Q., Y. Shao, et al. (2008). "Fine-scale population structure of Accumulibacter phosphatis in enhanced biological phosphorus removal sludge." J Microbiol Biotechnol **18**(7): 1290-1297.
Ward, D. M., M. J. Ferris, et al. (1998). "A natural view of microbial biodiversity within hot spring cyanobacterial mat communities." Microbiol Mol Biol Rev **62**(4): 1353-1370.
Ward, N. L., J. F. Challacombe, et al. (2009). "Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils." Appl Environ Microbiol **75**(7): 2046-2056.
Warnick, T. A., B. A. Methe, et al. (2002). "Clostridium phytofermentans sp. nov., a cellulolytic mesophile from forest soil." Int J Syst Evol Microbiol **52**(Pt 4): 1155-1160.
Watson, S. W. and J. B. Waterbury (1971). "Characteristics of two marine nitrite oxidizing bacteria, Nitrospina gracilis nov. gen. nov. sp. and Nitrococcus mobilis nov. gen. nov. sp. ." Archives of Microbiology **77**(3): 203-230.
Weese, J. S., C. Jarlot, et al. (2009). "Survival of Streptococcus equi on surfaces in an outdoor environment." Can Vet J **50**(9): 968-970.
Weinberg, E. D. (1984). "Iron withholding: a defense against infection and neoplasia." Physiol Rev **64**(1): 65-102.
Weinberg, M. V., G. J. Schut, et al. (2005). "Cold shock of a hyperthermophilic archaeon: Pyrococcus furiosus exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoproteins." J Bacteriol **187**(1): 336-348.
Wells, C. L. and T. D. Wilkins (1996). Clostridia: Sporeforming Anaerobic Bacilli. Medical Microbiology. S. Baron. Galveston (TX).
Wexler, H. M. (2007). "Bacteroides: the good, the bad, and the nitty-gritty." Clin Microbiol Rev **20**(4): 593-621.
White, A. K. and W. W. Metcalf (2004). "Two C-P lyase operons in Pseudomonas stutzeri and their roles in the oxidation of phosphonates, phosphite, and hypophosphite." J Bacteriol **186**(14): 4730-4739.
Whitford, M. F., R. M. Teather, et al. (2001). "Phylogenetic analysis of methanogens from the bovine rumen." BMC Microbiol **1**: 5.
Whitman, W. B., D. C. Coleman, et al. (1998). "Prokaryotes: the unseen majority." Proc Natl Acad Sci U S A **95**(12): 6578-6583.

Wierzchos, J., B. Camara, et al. (2011). "Microbial colonization of Ca-sulfate crusts in the hyperarid core of the Atacama Desert: implications for the search for life on Mars." Geobiology **9**(1): 44-60.
Williamson, S. J., D. B. Rusch, et al. (2008). "The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples." PLoS One **3**(1): e1456.
Willmott, C. J., S. E. Critchlow, et al. (1994). "The complex of DNA gyrase and quinolone drugs with DNA forms a barrier to transcription by RNA polymerase." J Mol Biol **242**(4): 351-363.
Willner, D., R. V. Thurber, et al. (2009). "Metagenomic signatures of 86 microbial and viral metagenomes." Environ Microbiol **11**(7): 1752-1766.
Wilmes, B., A. Hartung, et al. (2010). "Fed-batch process for the psychrotolerant marine bacterium Pseudoalteromonas haloplanktis." Microb Cell Fact **9**: 72.
Winther, L., R. M. Andersen, et al. (2010). "Association of Stenotrophomonas maltophilia infection with lower airway disease in the horse: a retrospective case series." Vet J **186**(3): 358-363.
Woese, C. R., O. Kandler, et al. (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-4579.
Wood, D. W., J. C. Setubal, et al. (2001). "The genome of the natural genetic engineer Agrobacterium tumefaciens C58." Science **294**(5550): 2317-2323.
Woyke, T., D. Tighe, et al. (2010). "One bacterial cell, one complete genome." PLoS One **5**(4): e10314.
Woyke, T., G. Xie, et al. (2009). "Assembling the marine metagenome, one cell at a time." PLoS One **4**(4): e5299.
Wu, X. Y., K. L. Shi, et al. (2010). "Alkaliphilus halophilus sp. nov., a strictly anaerobic and halophilic bacterium isolated from a saline lake, and emended description of the genus Alkaliphilus." Int J Syst Evol Microbiol **60**(Pt 12): 2898-2902.
Xie, G., D. C. Bruce, et al. (2007). "Genome sequence of the cellulolytic gliding bacterium Cytophaga hutchinsonii." Appl Environ Microbiol **73**(11): 3536-3546.
Xu, J. and J. I. Gordon (2003). "Honor thy symbionts." Proc Natl Acad Sci U S A **100**(18): 10452-10459.
Yabuuchi, E., H. Yamamoto, et al. (2001). "Proposal of Sphingomonas wittichii sp. nov. for strain RW1T, known as a dibenzo-p-dioxin metabolizer." Int J Syst Evol Microbiol **51**(Pt 2): 281-292.
Yakimov, M. M., P. N. Golyshin, et al. (1998). "Alcanivorax borkumensis gen. nov., sp. nov., a new, hydrocarbon-degrading and surfactant-producing marine bacterium." Int J Syst Bacteriol **48 Pt 2**: 339-348.
Yeh, Y. C., L. R. Comolli, et al. (2010). "The caulobacter Tol-Pal complex is essential for outer membrane integrity and the positioning of a polar localization factor." J Bacteriol **192**(19): 4847-4858.
Yergeau, E., S. Bokhorst, et al. (2011). "Shifts in soil microorganisms in response to warming are consistent across a range of Antarctic environments." ISME J.
Yilmaz, S., M. Allgaier, et al. (2010). "Multiple displacement amplification compromises quantitative analysis of metagenomes." Nat Methods **7**(12): 943-944.
Yooseph, S., G. Sutton, et al. (2007). "The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families." PLoS Biol **5**(3): e16.
Yoshida, T. and T. Sakamoto (2009). "Water-stress induced trehalose accumulation and control of trehalase in the cyanobacterium Nostoc punctiforme IAM M-15." J Gen Appl Microbiol **55**(2): 135-145.

Zeikus, J. G., R. Kerby, et al. (1985). "Single-carbon chemistry of acetogenic and methanogenic bacteria." Science **227**(4691): 1167-1173.
Zhang, X., J. E. Vrijenhoek, et al. (2011). "A genetic element present on megaplasmids allows Enterococcus faecium to use raffinose as carbon source." Environ Microbiol **13**(2): 518-528.
Zheng, L., V. L. Cash, et al. (1998). "Assembly of iron-sulfur clusters. Identification of an iscSUA-hscBA-fdx gene cluster from Azotobacter vinelandii." J Biol Chem **273**(21): 13264-13272.
Zheng, L., R. H. White, et al. (1993). "Cysteine desulfurase activity indicates a role for NIFS in metallocluster biosynthesis." Proc Natl Acad Sci U S A **90**(7): 2754-2758.
Zhou, J., M. A. Bruns, et al. (1996). "DNA recovery from soils of diverse composition." Appl Environ Microbiol **62**(2): 316-322.
Zhou, J., L. Wu, et al. (2011). "Reproducibility and quantitation of amplicon sequencing-based detection." ISME J **5**(8): 1303-1313.
Zhou, N. Y., S. L. Fuenmayor, et al. (2001). "nag genes of Ralstonia (formerly Pseudomonas) sp. strain U2 encoding enzymes for gentisate catabolism." J Bacteriol **183**(2): 700-708.
Zhulin, I. B. (2009). "It is computation time for bacteriology!" J Bacteriol **191**(1): 20-22.
Zijnge, V., M. B. van Leeuwen, et al. (2010). "Oral biofilm architecture on natural teeth." PLoS One **5**(2): e9321.
Zobell, C. E. (1946). "Action of microorganisms on hydrocarbons." Bacteriol Rev **10**(1): 1-49.
Zscheck, K. K. and B. E. Murray (1993). "Genes involved in the regulation of beta-lactamase production in enterococci and staphylococci." Antimicrob Agents Chemother **37**(9): 1966-1970.
Zumft, W. G. (1997). "Cell biology and molecular basis of denitrification." Microbiol Mol Biol Rev **61**(4): 533-616.