Détermination de sondes oligonucléotidiques pour outils moléculaires à haut débit : application pour le développement d'une nouvelle approche de capture de gènes pour l'écologie microbienne

Jérémie Denonfoux

UNIVERSITE BLAISE PASCAL
CLERMONT-FERRAND II

UNIVERSITE D'AUVERGNE
CLERMONT-FERRAND I

ANNEE 2013

ECOLE DOCTORALE DES SCIENCES DE LA VIE, SANTE, AGRONOMIE ET ENVIRONNEMENT

Thèse

Présentée à l'Université Blaise Pascal pour l'obtention du grade de

DOCTEUR D'UNIVERSITE
(Spécialité : Génomique et Ecologie Microbienne)

Présentée et soutenue publiquement le 9 janvier 2013

Jérémie DENONFOUX

DETERMINATION DE SONDES OLIGONUCLEOTIDIQUES POUR OUTILS MOLECULAIRES A HAUT-DEBIT : APPLICATION POUR LE DEVELOPPEMENT D'UNE NOUVELLE APPROCHE DE CAPTURE DE GENES POUR L'ECOLOGIE MICROBIENNE

Composition du jury :

Rapporteurs : Françoise BINET (CR CNRS, EBE, UMR 6553, Université de Rennes 1)
Robert DURAN (Pr., UMR IPREM 5254, Université de Pau et des pays de l'Adour, Pau)
Pascal SIMONET (DR CNRS, Laboratoire Ampère, UMR 5005, Université Claude Bernard, Lyon)

Examinateurs : Aurélie CEBRON (CR CNRS, LIMOS, UMR 7137, Université de Lorraine, Nancy)
Diego Pablo MORGAVI (DR INRA, Unité de Recherche Herbivores, INRA Clermont-Ferrand/Theix)
Eric PEYRETAILLADE (Dr., EA-4678 CIDAM, Université d'Auvergne, Clermont-Ferrand)

Directeur : Pierre PEYRET (Pr., EA-4678 CIDAM, Université d'Auvergne, Clermont-Ferrand)

Laboratoire « Microorganismes : Génome et Environnement »
Unité mixte de recherche CNRS 6023

Equipe d'Accueil 4678 « Conception, Ingénierie et Développement de l'Aliment et du Médicament

Détermination de sondes oligonucléotidiques pour outils moléculaires à haut-débit : application pour le développement d'une nouvelle approche de capture de gènes pour l'écologie microbienne

Résumé :

Les microorganismes, par leurs fascinantes capacités d'adaptation liées à l'extraordinaire diversité de leurs capacités métaboliques, jouent un rôle fondamental dans les tous les processus biologiques. Ils interviennent notamment au niveau des changements globaux, comme le réchauffement climatique, en partie occasionné par les émissions croissantes de méthane dans l'atmosphère, mais également par les pollutions résultant de la dispersion de molécules comme les Hydrocarbures Aromatiques Polycycliques. Ainsi, les communautés microbiennes vont participer à réduire ou à augmenter les effets délétères de l'anthropisation des écosystèmes. La régulation des changements globaux passe donc par une meilleure connaissance de ces communautés qui doivent être explorées dans leur globalité au sein des environnements. Néanmoins en raison de leur forte complexité, une telle exploration n'est possible qu'en utilisant des outils d'analyse haut-débit. Cependant, l'emploi d'outils moléculaires à haut-débit comme les biopuces à ADN passe par la détermination de sondes combinant à la fois une forte sensibilité, une très bonne spécificité et un caractère exploratoire. Pour concevoir de telles sondes un nouveau logiciel KASpOD a donc été développé. De même, en utilisant des sondes présentant les mêmes caractéristiques, le développement d'une nouvelle approche innovante en écologie microbienne de capture de gènes en solution été entrepris. Cette nouvelle méthode d'enrichissement de gènes d'intérêt couplée à du séquençage haut-débit a été appliquée pour l'exploration des communautés méthanogènes du lac Pavin. Les résultats obtenus montrent la pertinence de l'approche qui assure une meilleure évaluation de diversité de l'écosystème avec notamment l'identification de populations appartenant à la biosphère rare. L'autre ajout majeur de cette approche est qu'elle autorise l'identification de grandes régions d'ADN génomique exploitable pour caractériser de nouveaux gènes ou de nouveaux processus adaptatifs.

Mots clés : *changement global, métagénomique, détermination de sondes, capture de gènes*

Selection of oligonucleotide probes for high-throughput molecular tools : application for a new gene capture method's development for microbial ecology

Abstract :

Microorganisms play a crucial role in all biological processes related to their huge metabolic potentialities. They are involved in global changes such as global warming partially caused by the growing methane emissions in the atmosphere, but also by the release of pollutants such as Polycyclic Aromatic Hydrocarbons. Thus, microbial communities will contribute to reduce or increase the negative effects of human impacts on ecosystems. The regulation of global changes needs a better knowledge of the microbial communities involved in complex environments functioning. Nevertheless, a complete exploration of such environments requires the use of high-throughput tools, due to the extraordinary diversity of microorganisms within the ecosystems. The use of DNA microarrays requires a probe design step allowing the selection of highly sensitive, specific and explorative oligonucleotides. For this purpose, we have developed KASpOD, a new software, allowing the generation of efficient probes dedicated to environmental applications. Using high quality probe sets, an innovative in solution-based gene capture method combined with Next Generation Sequencing, was developed and applied for the exploration of the methanogen communities in lake Pavin, Results showed the relevance of this approach that allows a better evaluation of the methanogen diversity with an efficient detection of populations belonging to the rare biosphere. The other main advantage of this approach is the identification of large regions of genomic DNA, useful for the characterization of new genes or adaptive processes.

Keywords: *global change, metagenomics, probe design, gene capture*

Remerciements

Voilà, le temps est venu pour moi de refermer cette thèse en y mettant le point final. Mais avant ceci, je voulais adresser mes plus sincères remerciements à bon nombre de personnes qui m'ont accompagné, soutenu, épaulé, guidé au cours de ces trois années. Même si une thèse est un challenge scientifique, c'est avant tout une folle aventure semée quelques fois d'embuches où l'entraide, le soutien moral et la bonne humeur sont les armes indispensables pour mener à bien cette quête du « Graal » !

Mes remerciements vont tout d'abord à mes différents directeurs d'unités, Christian Amblard et Télésphore Sime-Ngando du LMGE aux Cézeaux, et Monique Alric de l'EA 4678-CIDAM au CBRV. Merci à vous pour votre accueil et pour m'avoir donné les moyens de mener à bien cette thèse.

Sincères remerciements à Françoise Binet, Robert Duran et Pascal Simonet d'avoir accepté d'être rapporteurs de cette thèse, ainsi qu'à Aurélie Cébron et Diego Morgavi pour avoir accepté d'examiner mon travail. Un grand merci également au CNRS et à la Région Auvergne pour le financement de ces 3 années de doctorat.

Des remerciements chaleureux vont ensuite à mon directeur de thèse, Pierre Peyret pour m'avoir conseillé, guidé dans mes travaux, et m'avoir montré les voies de la rigueur et de la clarté du message scientifique. Merci à vous pour votre confiance et cette grande liberté dont j'ai pu bénéficier pendant ces 3 années, associées à des moyens financiers de qualité, pour mener à bien ces travaux de thèse. Je n'oublie pas non plus nos déplacements en congrès à Genève et à Copenhague et à notre recherche des spécialités locales !

Viens ensuite Corinne Biderre-Petit à qui je dois beaucoup. Merci pour ton aide, tes conseils, ton soutien et ton humour à toute épreuve. Merci pour cette générosité que tu transmets, au travail ou en compagnie de ta famille. J'adresse par la même occasion un grand merci à ton mari Mathias, pour m'avoir rééduqué et remis sur de bons pieds suite à mon accident.

Les remerciements vont ensuite à Eric Peyretaillade, Monsieur « Pierretaillade », à qui je dois également dire un grand merci. Ta rigueur et ton avis critique ont été des atouts précieux pendant ces 3 années, de même que ton investissement sur mes travaux, et sur la correction du manuscrit. Je n'oublie pas ta bonne humeur légendaire et notre formidable escapade aux Karellis en janvier 2011. Que de bons souvenirs ! J'en profite également pour m'excuser de ce plongeon mémorable du télésiège à l'insu de ton plein gré…tu ne m'en veux pas ?

C'est à ton tour Delphine Boucher ! Cette année passée au CBRV nous aura rapprochés, alors tout simplement un grand merci à toi, ta gentillesse, merci pour tout.

Mon cher Eric Dugat-Bony tu es le suivant ! Comment puis-je te dire merci pour tout ce que tu as fait pour moi mon ami? à travers nos discussions, nos délires, nos soirées ?!!

Quelle belle aventure nous avons partagé ensemble pendant plus de deux ans ! Cela a été un réel plaisir de travailler avec toi, d'avoir été mon compagnon de route et d'offrir généreusement ta passion de la recherche, ta bonne humeur et ta simplicité. Je te souhaite pleins de bonnes choses du côté de Paris, et je sais que cette carrière à l'INRA sera bien remplie ! Un grand merci à Elodie, à tes parents, et à la « bande de jeunes » pour reprendre la chanson de Renaud Séchan, qui officie à Bourg Lastic et que je salue chaleureusement !!

Je continue en remerciant mon acolyte musical, mon partenaire de nos soirées culture G, mon équipier des tournois pétanques, oui c'est toi mon cher Nicolas Parisot ! Un grand merci à toi pour ton aide, ton amitié et les développements bioinformatiques sans qui, et bien, tous ces travaux n'auraient pas abouti ! Cela aura été un réel plaisir de travailler ensemble, à naviguer sous la même brise et je te souhaite plein de bonnes pour la suite. Que cette thèse que tu nous prépares soit bien remplie !

Je souhaite également remercier tous les anciens membres de l'équipe GIIM du LMGE, qui m'ont accueilli, aidé et témoigné une très grande amitié et sympathie au cours de cette thèse. Merci à Brigitte Chebance, Anne Moné et Isabelle Pinto pour votre gentillesse ! Merci à Olivier Gonçalves pour son humour unique et sa passion pour Günther, et Sébastien Terrat pour tous les bons moments passés au cours de cette thèse, mais aussi pour ta soutenance ! Quelle bonne soirée, en espérant que le canard t'as bien fait plaisir ! Je n'oublie pas non plus Sophie Comtet et Faouzi Jaziri à qui je souhaite une bonne fin de thèse. Enfin j'adresse mes remerciements à Ourdia Bouzid, Emilie Dumas et Mohieddine Missaoui, présents à mes débuts de thèse.

La suite de mes remerciements va chaleureusement à mes amis et collègue du LMGE. Je commencerai par Mylène Hugoni qui a beaucoup compté durant ces trois années. Ton amitié et ton soutien ont été précieux, je garde le merveilleux souvenir de nos soirées passées ensemble, nos séances Royat Tonic ou encore notre lambada ! C'est quand tu veux pour la prochaine ! Un immense merci à Benjamin Misson, pour ton amitié, ta bonne humeur mais aussi pour m'avoir fait partager le plus beau jour de votre vie avec Anne ! Quel mariage ! J'en garde un merveilleux souvenir ! Je n'oublie pas non plus Stéphanie Palesse, Emilie Duffaut, Julie Aufauvre, Mathieu Roudel, Cyril Vidau, Guillaume Borrel, Marion Sabart, Olivier Brouard, Eléonore Attard, Najwa Taib et bien d'autres !

Je tiens également à remercier les stagiaires passés dans l'équipe que j'ai eu le plaisir de côtoyer ou d'encadrer durant mon doctorat : Sarah Orlhac, Laurianne Roux, Nicolas Gallois, Stella Baret, Emilie Girard, Valérie Georges, Anne-Sophie Yvroud, Thomas Douellou, Stéphane Freitas, Laura Dumas, Pierre Marijon, Mickaël Mege…en vous souhaitant pleins de bonnes pour la suite !

Viens ensuite le plaisir pour moi de remercier tous les membres de l'EA CIDAM qui m'ont accueilli dans les locaux du CBRV pour cette fin de troisième année de thèse.

Les premiers remerciements vont naturellement et chaleureusement vers Lucie Etienne-Mesmin, ma compagne de rédaction, avec qui nous avons partagé beaucoup de choses même dans les moments un peu plus difficiles. Je voulais te dire un grand merci Lucie

pour ton amitié, nos discussions, ton aide précieuse et ta générosité. Que ton avenir prochain en post-Doc soit radieux et comblé !

J'adresse ensuite mille mercis à Charlotte Cordonnier qui représente par sa simplicité, sa sincérité et sa générosité, une belle et prometteuse relève ! Merci pour tout ce que tu as fait pour moi en cette fin de rédaction Charlotte, pour ton soutien, ta bonne humeur et merci pour les Schokobons !

Et M. Jarrige ! Nous étions devenus à partir de 19h30 les seuls habitants du 5ème étage ! Je souhaitais chaleureusement vous remercier pour votre humour, votre ouverture d'esprit, nos discussions de couloir tardives, qui en cette fin de rédaction m'ont fait extrêmement plaisir.

Je n'oublie pas non plus Jonathan Thévenot, je voulais te remercier de ta gentillesse et de tes visites nocturnes dans le bureau pour échanger quelques mots et savoir si le moral était bon !

Et merci enfin à Céline Ribière pour ton aide précieuse pour la relecture du manuscrit.

J'en profite également pour remercier tous les autres acteurs de la bonne humeur au 5ème étage : William Tottey, Christelle Guyard, Jean-François Brugère, Sylvain Denis, Sandrine Chalancon, Carine Mazal, Stéphanie Blanquet-Diot, Pierre Charles Romond, Marie Cousseau, Aurélie Guerra, Nadia Gassi. Un grand merci également aux habitants du rez-de-chaussée du CBRV, avec qui j'ai eu le plaisir d'effectuer mon monitorat : Ghislain Garrait, Eric Beyssac, Valérie Hoffart, Jean-Michel Cardot, Xie Xiaoyu et Emmanuel Lainé.

Mes remerciements se poursuivent jusqu'à d'autres personnes du CBRV et de l'INRA avec qui j'ai eu le plaisir de partager de nombreux bons moments : Bruno Lamas, Benoit Chassaing, Jennifer Raisch, Nicolas Barnich, Jeremy Denizot, Amélie De Vallée, Laureen Crouzet et Priscilla Branchu.

Je n'oublie pas non plus Justine Dauzat, avec qui j'ai partagé de très bons moments ! Merci à toi Justine d'avoir été présente et avoir été une oreille très attentive durant cette dernière année de thèse.

Cette thèse n'aurait pas abouti également sans l'aide de ma famille, en particulier mes parents qui ont toujours été présents et témoigné un grand intérêt à mes travaux, mes grands-mères Yvonne et Odette à qui je pense très fort, mon frère et ma belle-sœur qui m'ont offert la joie immense d'être un parrain et un tonton comblé de leur petit Emmanuel. Je tenais également à remercier mes Ardennes d'adoption, mon lieu privilégié de détente au grand air de la ferme d'Ambly-Fleury, avec mon beau-frère, fidèle et inséparable ami Julien, et mes beaux-parents Béatrice et Denis. Merci également à mes amis les « Apôtres », anciens piliers de l'IUT, avec Patron, Lenny, Fred, Nico, Damien, mes amis de Polytech avec Ralfouf, Pierrot, Perrine… et tous les autres qui me le plaisir d'être présent à ma soutenance !

Pour conclure, j'ai une pensée très émue pour mes défunts grands-pères André et Gaston, qui je sais auraient été extrêmement ravis et fiers d'être parmi nous pour venir

encourager leur petit-fils. De même, que pour toi Marion, qui ne pourra être présente le jour de ma soutenance…merci pour cet amour fort que je partage et construit avec toi depuis ces 3 belles années, merci d'avoir été présente au quotidien et d'avoir toujours cru en moi, même depuis la région Champagne-Ardenne.

A mes grands-pères, à Marion, je vous dédie cette thèse…

Table des matières

Table des figures

Liste des tableaux

Liste des abbréviations

%GC	Pourcentage en Guanine et Cytosine
16S	16 Sverdberg
Adf	alcool déshydrogénase-coenzyme F420 dépendante
ADN	Acide DésoxyriboNucléique
ADNc	ADN Complémentaire
ADNg	ADN Génomique
ADNr	ADN Ribosomique
AGV	Acide Gras Volatil
AK-PTA	Acétate Kinase PhosphoTransAcétylase
ANME	Anaerobic Metahnotroph
ANR	Agence Nationale de la Recherche
ARDRA	Amplified Ribosomal DNA Restriction Analysis
A-RISA	Automated-Ribosomal Intergenic Spacer Analysis
ARN	Acide RiboNucléique
ARNm	ARN messager
ARNr	ARN ribosomique
ATP	Adénosine TriPhospahte
BLAST	Basic Local Alignment Search Tool
BSR	Bactéries Sulfato-Réductrices
BTEX	Benzène, Toluène, Éthylbenzène et Xylènes
CARD-FISH	CAtalysed Reporter Deposition-FISH
CDD	Charge Coupled Device
CNRS	Centre National de la Recherche Scientifique
CoA	Coenzyme A
CODH	CO Désydrogénase
CoM	Coenzyme M
CPU	Central Processing Unit
CRRI	Centre Régional des Ressources Informatiques
CRT	Cyclic reversible Termination
DDBJ	DNA Data Bank of Japan
ddNTP	Didésoxyribonucléotide triphosphate
DGGE	Denaturing Gradient Gel Electrophoresis
dNTP	Désoxyribonucléotides triphosphate
EGEE	Enabling Grid for E-sciencE
EMBL	European Molecular Biology Laboratory
emPCR	PCR en émulsion
EPA	Enviornmental Protection Agency
EPO	ErythroPOïétine
Fdh	Formate Déshydrogénase
FGA	Functional Gene Array
FISH	Fluorescence In Situ Hybridization
GPGPU	General-purpose Processing on Graphics Processing Units
HAP	Hydrocarbure Aromatique Polycyclique
HITChip	Human Intestinal Tract Chip
HOMIM	Human Oral Microbe Identification Microarray
kDa	Kilo Dalton
Kow	Coefficient de partage eau/octanol
kpb	Kilo paire de bases
LBBE	Laboratoire de Biométrie et Biologie Evolutive
MAR-FISH	MicroAutoRadiography-FISH
MCR	Méthyl-Coenzyme M Reductase
McrA	sous unité α de la Méthyl-Coenzyme M Reductase
McrB	sous unité β de la Méthyl-Coenzyme M Reductase
McrC	sous unité C de la Méthyl-Coenzyme M Reductase
McrD	sous unité D de la Méthyl-Coenzyme M Reductase
McrG	sous unité γ de la Méthyl-Coenzyme M Reductase
MeNPOC	MethylNitroPoperonyl OxyCarbonyl
MFR	Méthanofurane
MIP	Molecular Inversion Probes
MMO	Méthane Monooxygénase
MMOH	Composante hydorxylase de la MMO
MPI	Message Passing Interface
MRT	Isoforme II de la Méthyl-Coenzyme M Reductase
MspA	Porine A de *Mycobacterium smegmatis*
NADP	Nicotinamide Adénine Dinucléotide Phosphate
NanoSIMS	Spectromètre de masse à ionisation secondaire à l'échelle nanométrique
NGS	Next Generation Sequencing
NO	Oxyde Nitrique
OMS	Organisation Mondiale de la Santé
ORF	Open Reading Frame (ou cadre de lecture ouvert)
OTU	Operational Taxonomic Unit
PCBs	PolyChloroByphényls
PCR	Polymerase Chain reaction
pH	Potentiel d'Hydrogène
pMMO	Forme Particlaire de la MMO
POA	Phylogenetic Oligonucleotide Array
POP	Polluant Organique Persistant
ppb	partie par billions
PPi	Pyrophospahte inorganique
ppm	partie par millions
PTP	PicoTiterPlate
qPCR	PCR quantitative
RDP	Ribosomal Database Project
RHD	Ring Hydroxylating Dioxygenase
SIP	Stable Isotope Probing
sMMO	Forme soluble de la MMO
SMRT	Single Molecule Real Time Technology
SNP	Single Nucleotide Polymorphism
SSCP	Single Strand Conformation Polymorphism
TGGE	Temperature Gradient Gel Electrophoresis
Tm	Melting temperature (ou temperature de fusion)
TOR	Trimethylamine N-Oxide Reductase
T-RFLP	Terminal-Restriction Length Polymorphism
USB	Universal Serial Bus
UTP	Uridine TriPhosphate
WGA	Whole Genome Array
YAC	Yeast Artificial Chromosome

Introduction générale

De par la grande diversité de leurs métabolismes et leur incroyable capacité d'adaptation, les microorganismes sont retrouvés dans de nombreux écosystèmes même les plus extrêmes, et interviennent dans tous les processus globaux parmi lesquels le cycle du carbone, dont la production de méthane constitue un point clé. Depuis la révolution industrielle, la composition chimique de l'atmosphère a été profondément modifiée du fait des activités humaines qui sont à l'origine d'émission croissante de nombreux gaz dont le méthane, absorbant le rayonnement infrarouge et participant à l'effet de serre. Ce dernier participe directement aux changements globaux actuellement observés et plus particulièrement dans le réchauffement climatique. En effet, l'impact du méthane a été évalué comme étant environ vingt-trois fois supérieur à celui d'autres gaz à effet de serre tel que le dioxyde de carbone. La grande majorité de la production de méthane étant d'origine microbienne (Conrad 2009), il est important de pouvoir explorer ces populations afin de mieux comprendre les mécanismes à la base de la production et de l'utilisation de ce composé. En outre, ces connaissances pourraient être à la base d'approches conduisant à limiter l'émission de méthane.

Outre une présence dans une grande diversité d'environnements, ces populations pourraient également jouer un rôle dans la dégradation anaérobie des hydrocarbures aromatiques polycycliques (HAP). En France, la base de données BASOL (http://basol.environnement.gouv.fr/) recense actuellement 4411 sites pollués, avec près de 17% de ces sites contaminés par des HAP. Ces pollutions touchent différents compartiments environnementaux comme les sols, les sédiments lacustres ou marins. Les HAP sont classés comme des polluants organiques persistants (POP) en raison de leur temps de rétention très important dans l'environnement lié à leur faible solubilité dans l'eau. De même, ces molécules ont démontré un caractère cancérigène et mutagène, avec certains HAP classés comme polluants prioritaires par des agences internationales comme l' « US Environmental Protection Agency » (EPA) ou encore l'Organisation Mondiale de la Santé (OMS).

Dans ce contexte, une bonne connaissance des communautés microbiennes et plus particulièrement des méthanogènes, s'avère indispensable pour évaluer leurs capacités d'adaptation et leurs implications dans les changements globaux ou des actions à mener pour

limiter les effets délétères de l'anthropisation. Cependant, l'acquisition de ces données représente un encore défi majeur du fait de l'extraordinaire diversité et complexité des communautés microbiennes présentes au sein des différents écosystèmes (Torsvik et al 1990, Whitman et al 1998). Toutefois, grâce à l'émergence des outils dits à haut-débit, une caractérisation structurale et fonctionnelle des communautés microbiennes devrait être facilitée. En effet, l'essor de la métagénomique, lié à l'évolution du séquençage massif, a laissé entrevoir de nouvelles perspectives. Cependant, l'utilisation directe des nouvelles approches de séquençage sur des environnements complexes reste encore délicate du fait des difficultés d'interprétation des masses de données générées et des coûts restant élevés malgré une réduction significative du prix des séquences de deuxième génération. La réduction de la complexité semble être une bonne alternative notamment pour explorer les populations peu abondantes comme les méthanogènes. Avec l'objectif de réduire cette complexité, l'enrichissement préalable des gènes d'intérêt représente donc une approche innovante. Parallèlement, pour appréhender les communautés microbiennes des environnements complexes, les biopuces à ADN (phylogénétiques ou fonctionnelles) apparaissent également être des outils de choix du fait de leur simplicité d'utilisation, de leur capacité de multiplexage assurant la gestion simultanée d'un grand nombre d'échantillons et de la facilité d'interprétation des résultats. Cependant, le point clef du développement de ces deux types d'approches porte sur la détermination *in silico* de sondes à la fois spécifiques et sensibles, mais également par la mise en place de nouvelles stratégies moléculaires d'enrichissement. Il sera alors possible de pouvoir caractériser la diversité encore inconnue non répertoriée dans les bases de données, mais également d'identifier les populations peu abondantes jouant un rôle prépondérant dans les processus globaux étudiés. Les objectifs de ces travaux de thèse s'inscrivent dans ces problématiques autant biologiques (place des méthanogènes au sein des changements globaux) que méthodologiques (développement de nouvelles stratégies haut débit d'étude de la diversité microbienne).

Ainsi, le mémoire de thèse sera structuré en quatre parties, dont la première fera état des connaissances bibliographiques des méthanogènes au sein du cycle du carbone et de l'anthropisation des écosystèmes, mais également des outils haut débit récemment développés pour caractériser cette diversité. Cette première partie se terminera par la description des nouvelles approches envisagées pour dépasser les limites rencontrées pour l'exploration d'environnements complexes. Les parties suivantes présenteront sous forme de publications ou de chapitres d'ouvrage les travaux de cette thèse.

La deuxième partie portera ainsi sur la présentation d'un chapitre d'ouvrage décrivant les stratégies bioinformatiques utilisables pour la sélection de sondes oligonucléotidiques, permettant l'évaluation de la diversité microbienne connue ou encore inconnue. Les troisième et quatrième parties présenteront quant à elles, sous forme de deux articles, les résultats ayant conduit d'une part à l'élaboration d'un nouveau logiciel (KASpOD) de détermination de sondes oligonucléotidiques, et d'autre part au développement d'une nouvelle méthode de capture de gènes couplée au séquençage massif de seconde génération. Cette dernière approche a d'ailleurs permis la mise en évidence d'une grande diversité de méthanogènes et de nouvelles organisations géniques au sein d'un environnement lacustre, le lac Pavin. Elles contribuent ainsi en une meilleure connaissance des adaptations des méthanogènes participants au bon fonctionnement de cet écosystème.

Une conclusion générale fera le bilan des avancées apportées par ces travaux de thèse et des perspectives pour l'étude des formidables capacités d'adaptation des microorganismes.

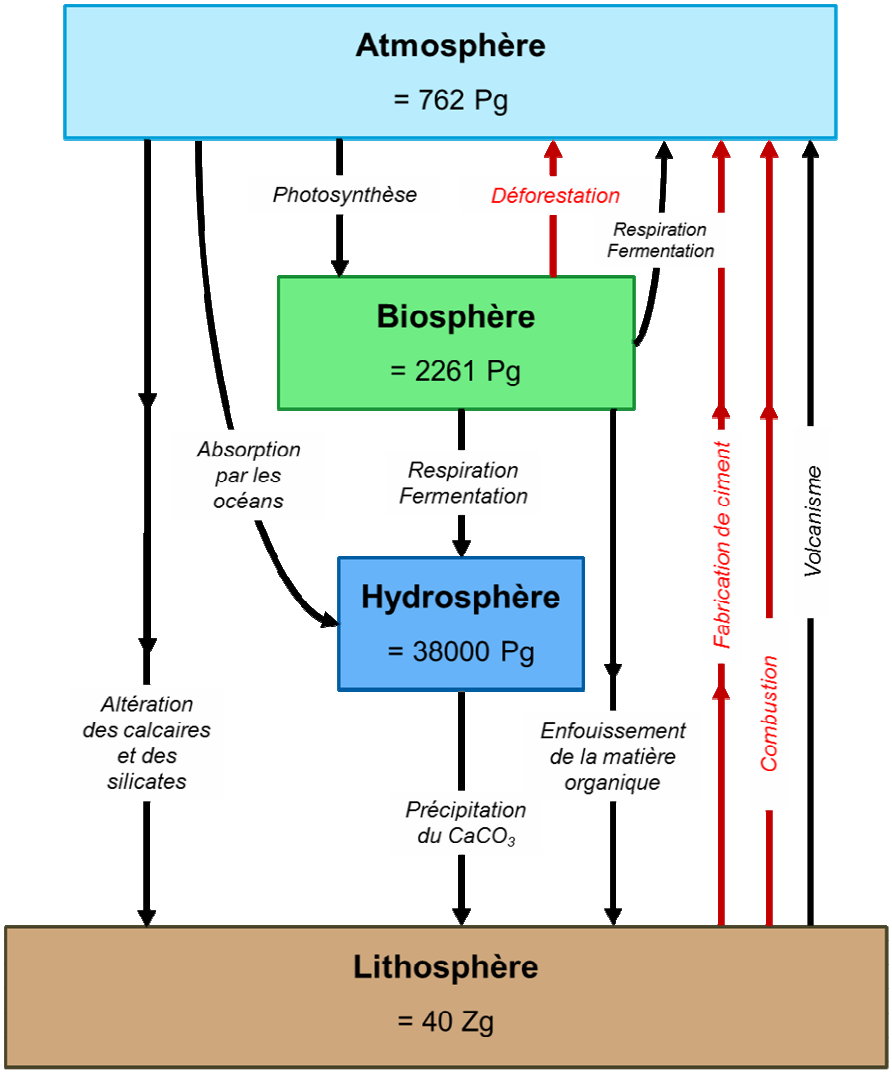

Figure 1. Le cycle global du carbone et ses principaux réservoirs. Les réservoirs sont en gras et les flux en italique. Les flux précédant la perturbation anthropique sont indiqués en noir et ceux résultant des activités humaines en rouge. 1 Pg (un pétagramme) = 10^{15}g ; 1 Zg (un zétagramme) = 10^{21}g. (Adapté d'après Berner et Berner, 1996 ; Kump, Kasting et Crane, 1999).

PARTIE 1 : Synthèse bibliographique

1. Diversité microbienne et changement global

A l'heure actuelle, une attention toute particulière de la communauté scientifique est portée sur les modifications majeures engendrées par l'activité humaine et les répercussions sur les écosystèmes. Ces modifications connues sous le terme de changement global, regroupent notamment les changements du climat et la modification des écosystèmes. Par leur biomasse élevée (4 à 6 × 10^{30} cellules microbiennes au niveau terrestre), leur diversité (nombre d'espèces bactériennes estimées par certains chercheurs à plus d'une dizaine de millions) (Eisen 2007, Whitman et al 1998) et leurs fascinantes capacités métaboliques et d'adaptation, les microorganismes ont un rôle essentiel dans le fonctionnement et l'évolution des cycles biogéochimiques et par conséquent celui des écosystèmes. De ce fait, les microorganismes sont des acteurs clefs au centre des changements globaux pouvant amplifier les effets néfastes liés à ces changements, ou au contraire participer à réduire ces effets. Les activités humaines impactant particulièrement le cycle biogéochimique du carbone, une attention toute particulière devra être portée sur l'étude des changements des communautés microbiennes au sein de ce cycle.

1.1 Le cycle du carbone

Sur Terre l'élément carbone (C) subit un ensemble de transformations caractérisées par des processus d'oxydo-réduction allant de sa forme la plus oxydée (exemple le dioxyde de carbone CO_2, nombre d'oxydation +IV) jusqu'à sa forme la plus réduite (le méthane CH_4, nombre d'oxydation –IV). Le carbone dit organique (associé à des éléments tel que l'oxygène, l'hydrogène, l'azote, le soufre ou encore le phosphore) possède toujours un nombre d'oxydation compris entre 0 (glucose, acétate) et –IV, tandis que celui du carbone dit minéral est toujours compris entre 0 et +IV (exemple du CO_2 [+IV], carbonate de calcium $CaCO_3$ [+IV], monoxyde de carbone CO [+II]). L'ensemble des transformations biotiques et abiotiques des différentes formes du carbone constitue le cycle biogéochimique du carbone.

1.1.1 Les principaux réservoirs

Le cycle du carbone se décline au travers de flux entre ses principaux réservoirs que sont l'atmosphère, la biosphère continentale, l'hydrosphère, et la lithosphère (**Figure 1**). L'atmosphère contient du carbone sous sa forme inorganique à savoir le CO_2 (à hauteur de

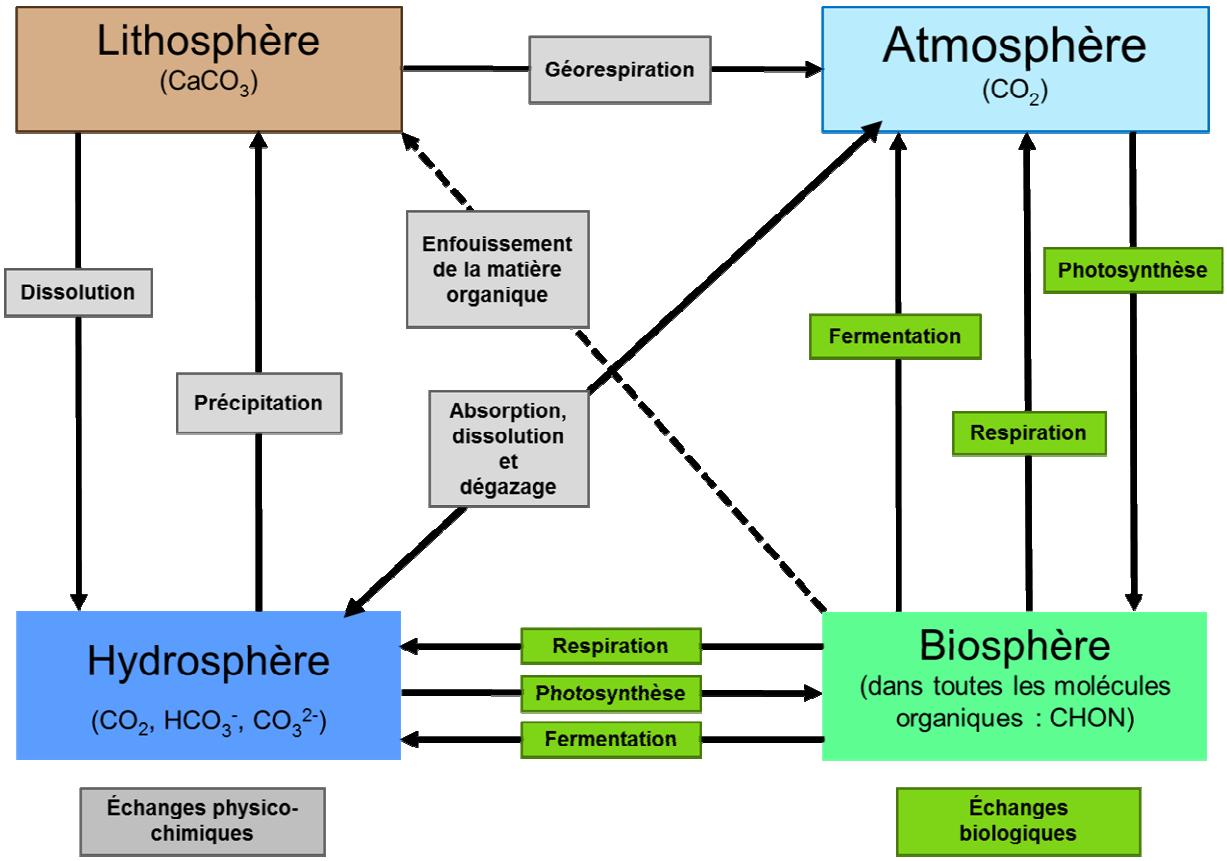

Figure 2. Les différents mécanismes de transfert de carbone. Les échanges physico-chimiques sont indiqués en gris et les échanges biologiques sont indiqués en vert.

0,0368% ou 368 ppm) et le CO (0,1 ppm), de même que du méthane CH_4 (environ 1,7 ppm). Le stockage du carbone dans ce réservoir est estimé à environ 762 Pg (1 Pg correspondant à 10^{15} g). La biosphère continentale stocke une quantité de carbone estimée à 2261 Pg se répartissant au niveau de la végétation (550 Pg), des sols et de la matière en décomposition (1711 Pg). Il est intéressant de noter que la végétation au niveau de la biosphère continentale contient moins de carbone que l'atmosphère (550 contre 762 Pg), alors que le premier mètre des sols en contient deux voire trois fois plus (1500 à 2000 Pg). L'hydrosphère contient quant à elle une quantité estimée à 37 000 Pg de carbone inorganique et 1000 Pg de carbone organique, ce qui représente une quantité 50 fois plus importante que pour l'atmosphère et 70 fois plus que pour la végétation terrestre. Il est également intéressant de constater que les zones océaniques profondes (mésopélagique et bathypélagique, entre 200 et 4000 mètres de profondeur) contiennent la grande majorité (environ 97%) du carbone de l'hydrosphère. Enfin, la lithosphère représente le plus gros réservoir de carbone, qu'elle stocke au niveau des roches carbonatées (environ 30 Zg correspondant à 30×10^{21} g) et des roches carbonées (environ 10 Zg). Ces dernières correspondent aux résidus de la matière organique accumulée à l'échelle géologique et enfouis dans des couches sédimentaires (gaz, pétrole, charbon…) (Berner and Berner 1996, Bertrand et al 2012, Kump et al 1999).

1.1.2 Mécanismes de transfert de carbone

Le cycle du carbone est caractérisé par des flux entre les réservoirs impliquant différents processus s'étalant sur différentes périodes. Il est possible de distinguer deux types d'échanges : ceux impliquant des processus d'échanges physico-chimiques et ceux impliquant des échanges biologiques (**Figure 2**).

1.1.2.a Echanges physico-chimiques

Les échanges physico-chimiques consistent en (1) la dissolution des roches silicatées de la lithosphère correspondant à un phénomène d'érosion par l'eau ($2CO_2 + 3H_2O + CaSiO_3 \rightarrow Ca^{2+} + 2HCO_3^- + H_4SiO_4$) où les produits dissouts sont transportés vers l'hydrosphère et précipitent sous forme de carbonate de calcium ($CaCO_3$) et de dioxyde de silicium (SiO_2), tous deux stockés dans les sédiments; (2) la pompe chimique de l'hydrosphère basée sur l'absorption du carbone inorganique de l'atmosphère du fait de la solubilité du CO_2 dans l'eau et grâce au système carbonate assurant le stockage du CO_2 sous forme de bicarbonates et de carbonates ($CO_2 + H_2O \leftrightarrow HCO_3^- + H^+ \leftrightarrow CO_3^{2-} + 2H^+$); (3) le stockage net de carbone organique dans les sédiments de l'hydrosphère par sédimentation et enfouissement,

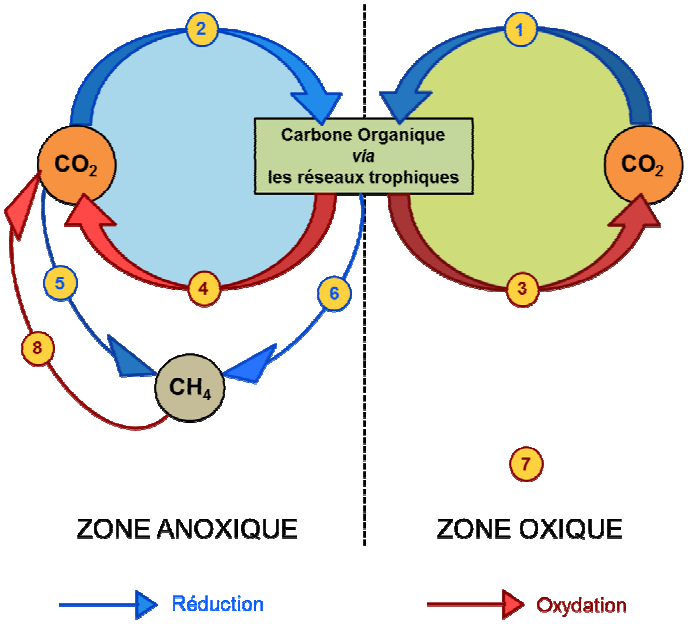

Figure 3. Rôle des microorganismes dans le cycle du carbone. (1) Production de carbone organique (biomasse) à partir du CO_2 (autotrophie) en conditions oxiques : bactéries et archées chimiolithotrophes, micro-eucaryotes photosynthétiques ; **(2)** Production de carbone organique à partir de la réduction du CO_2 (autotrophie) en conditions anoxiques : bactéries et archées chimiolithotrophes ou photolithotrophes ; **(3)** et **(4)** Utilisation du carbone organique comme source d'énergie et oxydation en CO_2 (minéralisation) en conditions oxiques (**3**, respiration aérobie) et en conditions anoxiques (**4**, fermentation) ; **(5)** production de méthane par des archées méthanogènes à partir de la réduction du CO_2 (respiration du CO_2) ; **(6)** production de méthane par des archées méthanogènes à partir de la réduction de composés organiques simples à 1, 2 ou 3 carbones ; **(7)** et **(8)** Oxydation du méthane en CO_2 (méthanotrophie) par des bactéries méthanotrophes en conditions oxiques **(7)** ou par des archées méthanotrophes en conditions anoxiques **(8)**. Les flèches en bleu correspondent à des processus de réduction, les flèches rouges à des processus d'oxydation. (Redessiné d'après Bertrand et al., 2012)

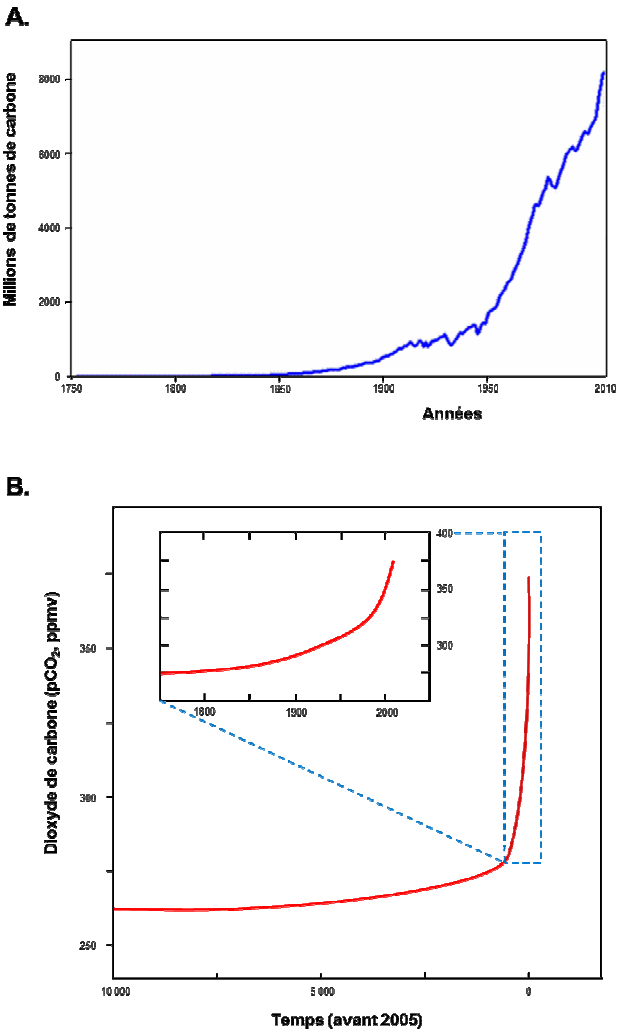

Figure 4. Evolution de la concentration en CO_2 dans l'atmosphère. (A) Emissions annuelles de carbone par la combustion des énergies fossiles (charbon, pétrole et gaz). La production de CO_2 lors de la production de ciment correspondant à une décomposition thermique de calcaire de chaux est également incluse. **(B)** pression partielle en CO_2 atmosphérique depuis 10 000 ans. L'encadré correspond à un agrandissement de la période 1751-2005. (Modifié et redessiné d'après Bertand et al., 2012).

représentant un processus 1000 fois plus faible que les processus biologiques de photosynthèse et de respiration ; (4) la géorespiration assurant le retour du CO_2 dans l'atmosphère résultant de plusieurs mécanismes tels que l'oxydation de la matière organique fossile des roches de la lithosphère (dissolution oxydative, dégradation microbienne ou thermique de la matière organique) mais également de l'activité de la pompe des carbonates de l'hydrosphère ($Ca^{2+} + 2HCO_3^- \rightarrow CaCO_3 + CO_2 + H_2O$) et de la décarbonatation des roches silicatées de la lithosphère ($CaCO_3 + SiO_2 \rightarrow CO_2 + CaSiO_3$) par le volcanisme, le métamorphisme (modifications minéralogiques et texturales sous l'effet de la pression et de la température) et la diagénèse (transformation des sédiments en roches sédimentaires) (Bertrand et al 2012).

1.1.2.b Echanges biologiques

Au niveau de ces échanges, trois réactions biologiques principales sont à considérer avec (1) la photosynthèse ($6CO_2 + 12H_2O + h\nu$ [lumière] $\rightarrow C_6H_{12}O_6 + 6O_2 + 6H_2O$) réalisée par les végétaux de la biosphère et certains microorganismes photosynthétiques comme les microalgues et les cyanobactéries; (2) la respiration ($C_6H_{12}O_6 + 6O_2 + 36ADP + 36Pi \rightarrow 6CO_2 + 6H_2O + 36ATP$) étant un processus inverse correspondant à l'oxydation de la matière organique permettant de convertir l'énergie stockée dans les liaisons chimiques des molécules en une énergie assimilable par les cellules, (3) la fermentation ($2CH_2O \rightarrow CO_2 + CH_4$) correspondant à une utilisation du carbone organique de la biosphère et de l'hydrosphère en anaérobiose par des microorganismes afin de récupérer de l'énergie, aboutissant à la production de CO_2 et de méthane CH_4 (Bertrand et al 2012) (**Figure 3**).

1.1.3 Modifications anthropiques du cycle du carbone

Les échanges biologiques et physico-chimiques évoqués précédemment ne modifient pas la quantité totale de carbone qui circule dans le cycle, mais ils provoquent sa redistribution entre les différents réservoirs. Ainsi, l'oxydation des combustibles fossiles représente un gain net de carbone au niveau atmosphérique. En effet, jusque vers les années 1800, le réservoir de carbone fossile n'influençait quasiment pas ce cycle, ce qui n'est plus le cas depuis la révolution industrielle (moitié du XIXe siècle) à partir de laquelle les émissions de CO_2 par les activités humaines ont augmenté de manière exponentielle (**Figure 4A**). Cette hausse de la quantité de CO_2 émise dans l'atmosphère a atteint 6,4 Pg de carbone par an entre 1990 et 2000, et 7,2 Pg de carbone par an entre 2000 et 2005. Même si ces émissions sont relativement faibles par rapport aux échanges biologiques de la photosynthèse et de la

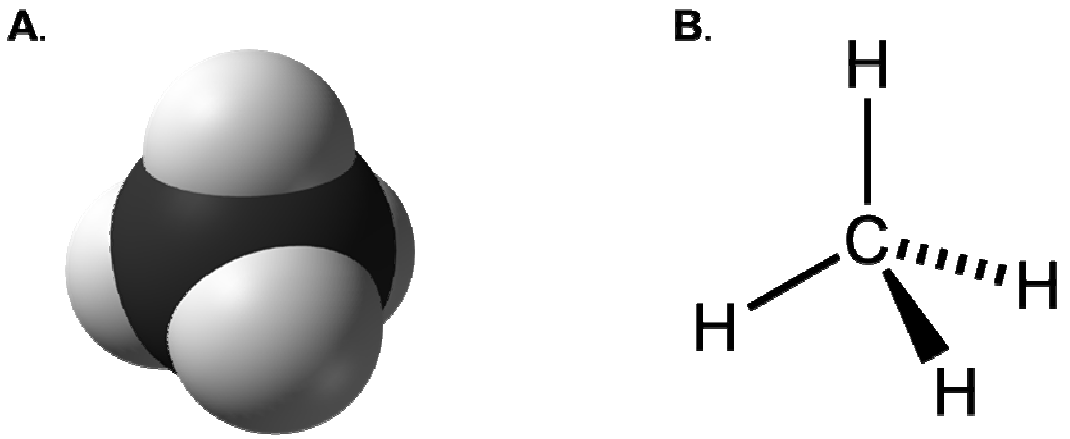

Figure 5. Représentations d'une molécule de méthane (CH_4). (A) Représentation en trois dimensions; **(B)** Représentation de Cram.

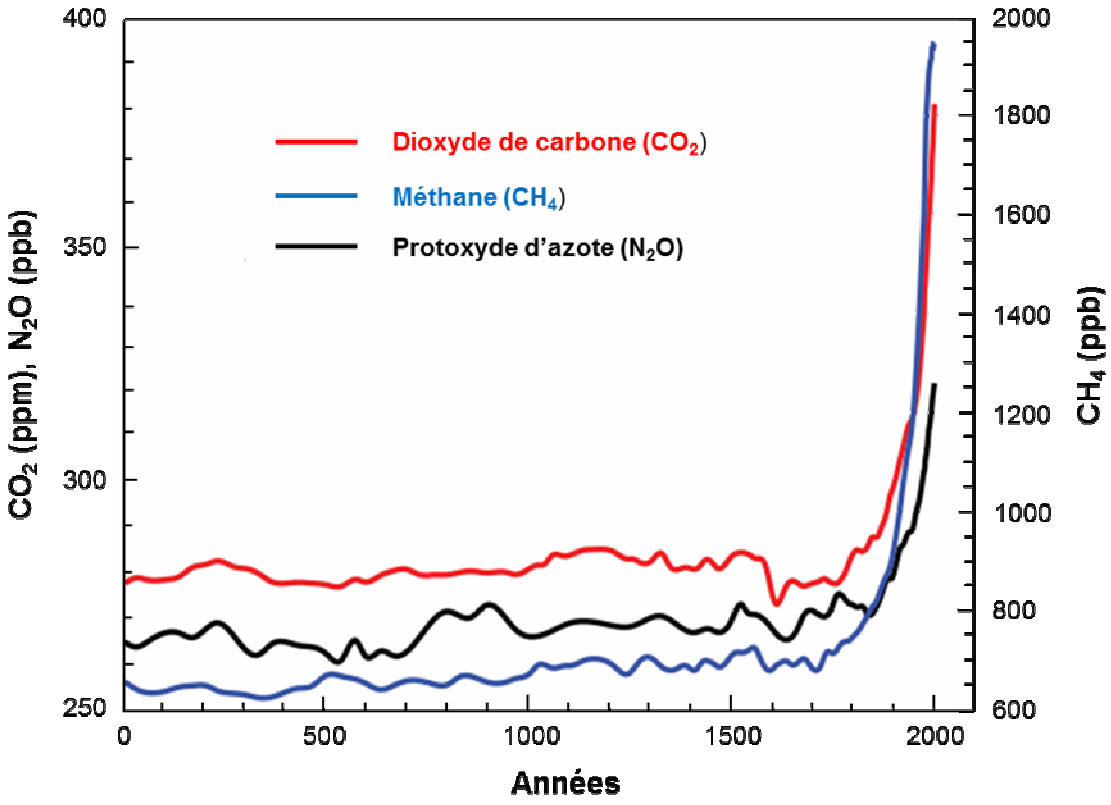

Figure 6. Concentrations atmosphériques des principaux gaz à effet de serre de l'an 0 à l'an 2000. Les concentrations du dioxyde de carbone (CO_2), du méthane (CH_4) et du protoxyde d'azote (N_2O) sont indiquées en partie par million (ppm) ou par milliard (ppb). (Modifié et redessiné d'après Solomon, 2007).

respiration (environ 120 Pg de carbone par an) et aux échanges physico-chimiques entre l'océan et l'atmosphère (environ 90 Pg de carbone par an), elles sont responsables de l'augmentation de la quantité de carbone dans l'atmosphère. La concentration du CO_2 atmosphérique a augmenté significativement depuis l'ère pré-industrielle où cette dernière était estimée à 280 ppm, pour atteindre 379 ppm en 2005 (Bertrand et al 2012) (**Figure 4B**). Ces données illustrent bien la conséquence des activités anthropiques par utilisation des combustibles fossiles, sur le relargage dans l'atmosphère d'une quantité de CO_2 non négligeable, et participant activement au phénomène dit de l'effet de serre et donc au réchauffement climatique.

Mais l'action de l'Homme sur les environnements ne s'arrête pas là. Les activités agricoles et industrielles, de plus en plus intenses, aboutissent également à la production de molécules comme les hydrocarbures ayant un rôle clé dans le cycle du carbone. Ces derniers comme le méthane (CH_4), participent de manière encore plus intense que le CO_2 au réchauffement climatique. De même, le raffinement des combustibles fossiles provoque la dissémination de composés responsables de nuisances et de pollutions, ayant une répercussion importante sur la santé des individus et sur le bon fonctionnement des écosystèmes.

1.2 Le méthane

Le méthane (CH_4) a été découvert et isolé par Alessandro Volta en 1776 en étudiant les émanations gazeuses des berges du lac Majeur (Italie), où il a pu mettre en évidence le caractère inflammable de ce « gaz des marais ». Le méthane est le plus simple des hydrocarbures, composé d'une molécule de carbone et de quatre molécules d'hydrogène (**Figure 5**). C'est un gaz inodore et incolore, principal constituant du gaz naturel qui participe activement, tout comme le CO_2, au processus du réchauffement climatique. En effet, le forçage radiatif se définit comme l'équilibre entre le rayonnement solaire entrant et les émissions de rayonnements infrarouges sortant de l'atmosphère. Le pouvoir radiatif associé au CH_4 est le deuxième plus important au niveau des gaz à effet de serre (+0,48 $W.m^2$) après le CO_2 (+1,66 $W.m^2$) (Forster et al 2007, Ramaswamy et al 2001). Le méthane contribuerait à environ 30% du forçage radiatif total (tenant compte des autres forçages positifs et négatifs) estimé à +1,6 $W.m^2$ (Conrad 2009, Forster et al 2007). Depuis le début de l'ère industrielle, la concentration atmosphérique en CH_4 a augmenté de 150% (**Figure 6**) pour être estimée à l'heure actuelle à 1800 ppb (Dlugokencky et al 2009) et son potentiel de réchauffement global (équivalent CO_2) est estimé à 25, c'est-à-dire qu'une tonne de CH_4 émise aujourd'hui aurait

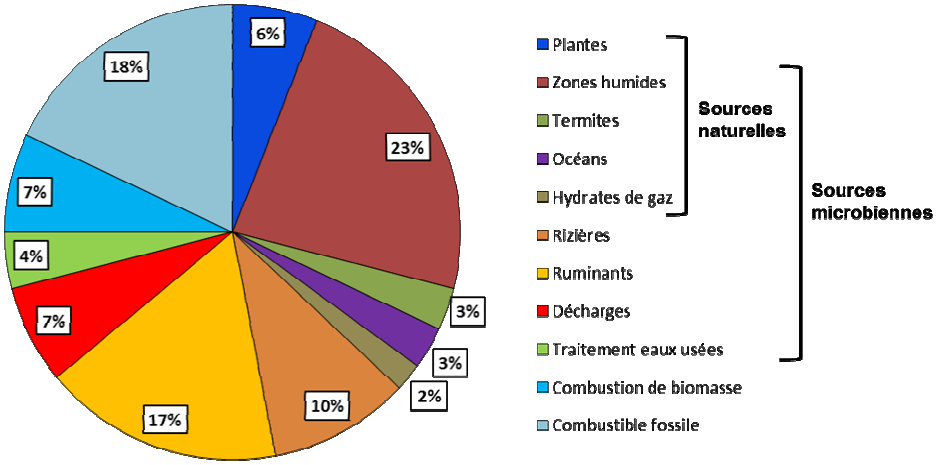

Figure 7. Contribution des sources (naturelles et anthropiques) de méthane au bilan atmosphérique total (500-600 Tg de CH_4 par an). 1 Tg (un Téragramme) = 10^{12}g. (Redessiné d'après Conrad, 2009)

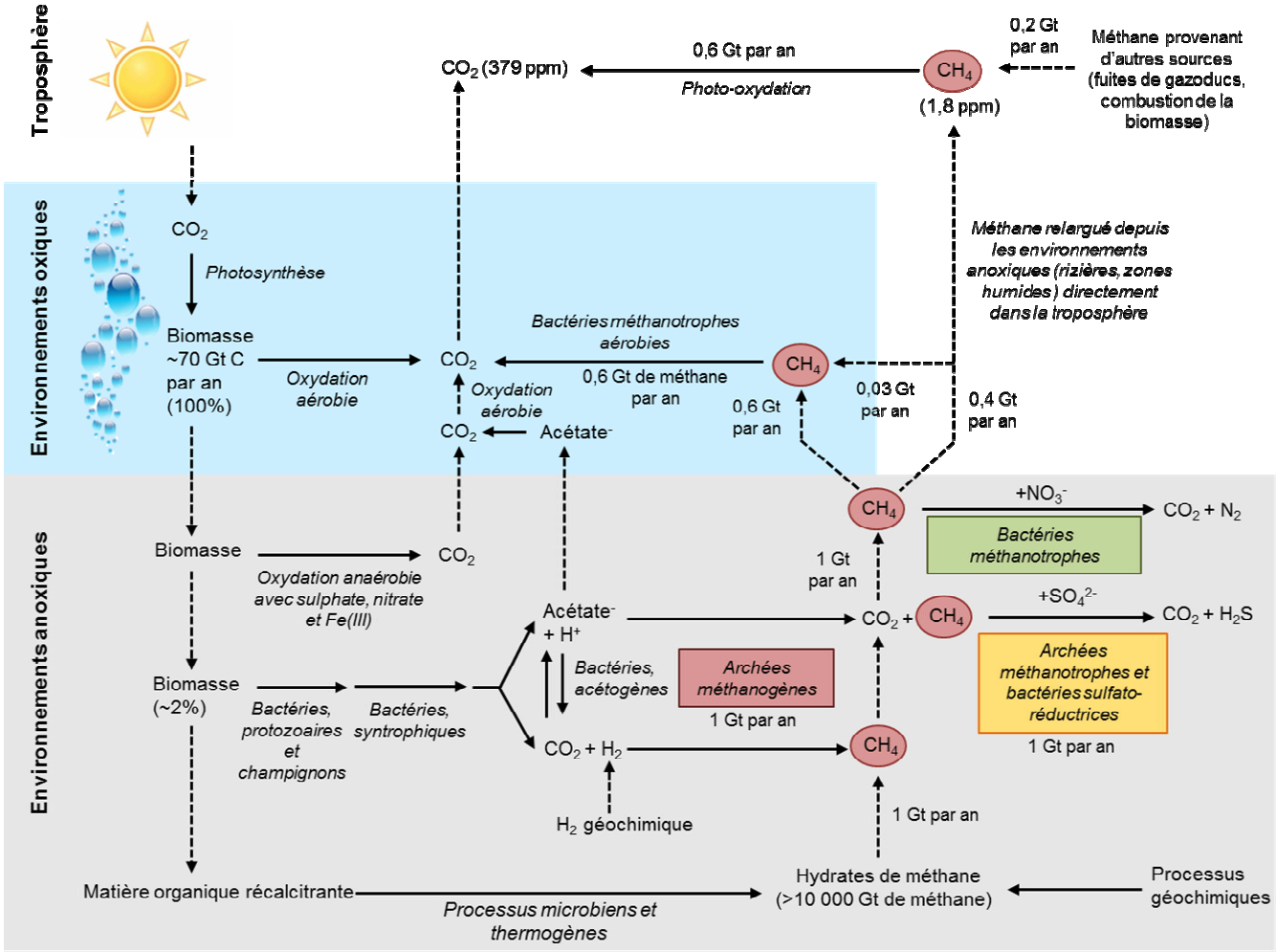

Figure 8. Le méthane comme intermédiaire du cycle du carbone. Les flèches continues indiquent une transformation chimique alors que les flèches en pointillés indiquent une diffusion et/ou une convection. (Redessiné d'après Thauer et al., 2008).

dans 100 ans un effet sur le réchauffement équivalent à 25 tonnes de CO_2. Toutefois, même si le temps de résidence du CH_4 dans l'atmosphère est beaucoup moins important que le CO_2 (une douzaine d'année pour le méthane contre une centaine pour le dioxyde de carbone), son action sur le réchauffement climatique demeure préoccupante. En effet, selon l'IPCC (Intergouvernemental Panel on Climate Change), la température au cours de ces cent dernières années a augmenté en moyenne d'environ 0,75°C dans le monde. Ces 25 dernières années, le rythme s'est accéléré avec un réchauffement estimé à 0,18°C par décennie.

1.2.1 Origine du méthane atmosphérique

Le flux de méthane atmosphérique est estimé à 500-600 Tg de CH_4 par an (Conrad 2009, Lelieveld et al 1998), avec une grande majorité de la production (environ 75%) d'origine biologique (**Figure 7**). Les 25% restant sont représentés par une production provenant de la combustion de la matière organique fossile, des feux de forêts ou encore des sources géologiques comme les écoulements froids des fonds océaniques. De même, il a été montré que les plantes pouvaient participer de manière importante à la production de méthane atmosphérique (environ 5%), mais cette production demeure controversée puisque les mécanismes biochimiques à la base de cette formation du méthane restent encore inconnus (Keppler et al 2006, Wang et al 2011). Ainsi, la majorité du méthane produit est d'origine microbienne (environ 70%) et résulte de l'activité de microorganismes hautement spécialisés et appartenant au domaine des *Archaea*.

1.2.2 Production biologique du méthane

La production biologique du méthane est assurée par l'activité de microorganismes méthanogènes anaérobies stricts appartenant au domaine des *Archaea*. Ces archées méthanogènes possèdent un métabolisme énergétique restreint à la formation de méthane à partir du CO_2 et du dihydrogène (H_2), du formate, du méthanol, des méthylamines ou encore de l'acétate (Thauer et al 2008), jouant un rôle important dans le cycle du carbone (**Figure 8**). Ces microorganismes sont des producteurs obligatoires de CH_4 et ubiquistes au sein d'environnements dépourvus d'oxygène comme les sédiments d'eau douce, les marais, les tourbières, les rizières, les décharges ou encore le tractus intestinal des ruminants et des termites (Thauer 1998).

1.2.2.a Les archées méthanogènes

Tableau 1. Caractéristiques des différents ordres d'archées méthanogènes.

Ordre	Morphologie cellulaire	Composition de la paroi	Composition en lipides	Substrats de la méthanogénèse	Présence de cytochrome	Température de croissance (°C)	Habitat
Methanobacteriales	Bâtonnets ou filaments	Pseudomurèine	Archaeol, caldarchaeol, Hydroxyarchaeol	H_2+CO_2, H_2+methanol, formate	non	37-65	Sédiments marins et d'eau douce, sols, tractus gastro-intestinal, environnements géothermaux
Methanocellales	Bâtonnets	ND	ND	H_2+CO_2, formate	ND	25-40	Rizières, sédiments marins et eau douce, sols
Methanococcales	Coques irréguliers	Protéine (couche S)	Archaeol, Caldarchaeol, Hydroxyarchaeol, Phospholipides archéobactériens	H_2+CO_2, formate	non	35-88	Environnements marins
Methanomicrobiales	Bâtonnets, coques	Protéine ou glycoprotéine	Archaeol. Caldarchaeol	H_2+CO_2, formate	non	15-55	Sédiments marins et d'eau douce. digesteurs et tractus gastrointestinal
Methanopyrales	Bâtonnets	Pseudomurèine	Archaeol	H_2+CO_2	non	84-110	Cheminées hydrothermales marines
Methanosarcinales	Coques irréguliers, amas, bâtonnets	Protéine ou hétéropolysaccharides	Archaeol, Hydroxyarchaeol, caldarchaeol	H_2+CO_2, $MeNH_2$, Acetate	oui	20-60	Sédiments marins et d'eau douce, tractus gastro-intestinal des animaux, digesteurs

ND: non déterminé

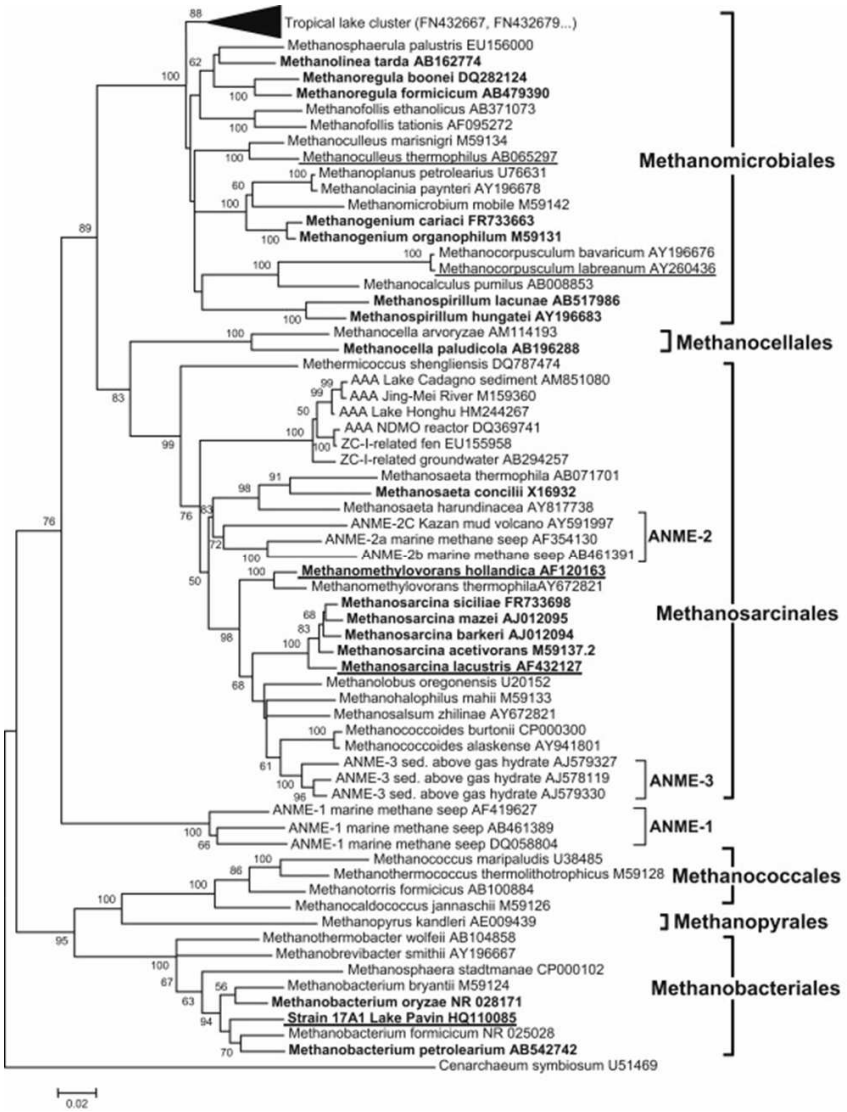

Figure 9. Arbre phylogénétique basé sur les séquences d'ADNr 16S illustrant les relations entre les six ordres méthanogènes ainsi que les trois groupes d'ANME « Anerobic Methanotrophs » appartenant au domaine des Archaea. (D'après Borrel et al., 2011)

Toutes les archées méthanogènes sont affiliées au phylum des *Euryarchaeota* et représentent environ 50% de la totalité des espèces d'archées cultivées (Plasencia et al 2010). Des analyses phylogénétiques, basées sur 53 protéines ribosomales, ont montré que les méthanogènes constituent un groupe paraphylétique (Bapteste et al 2005). De nombreuses études suggèrent que les archées méthanogènes non encore cultivées appartiennent potentiellement à de nouveaux ordres (Castro et al 2004, Juottonen et al 2005, Mihajlovski et al 2010, Nettmann et al 2008). Les archées méthanogènes actuellement connues et cultivées sont divisées en six ordres : les *Methanomicrobiales*, les *Methanosarcinales*, les *Methanocellales* (anciennement appelé « Rice Cluster I », RC-I), les *Methanobacteriales*, les *Methanococcales* et les *Methanopyrales* (Conrad et al 2006, Garcia et al 2000, Liu and Whitman 2008, Sakai et al 2010). Cette classification se base sur des critères phénotypiques et métaboliques (forme, structure de la paroi, composition en lipides, gamme de substrats, présence ou non de cytochromes) (**Tableau 1**), et sur leurs séquences d'ADNr 16S (**Figure 9**). Les archées méthanogènes sont largement distribuées dans les environnements riches en matière organique, mais leur croissance n'est possible qu'à des potentiels d'oxydoréductions très bas (inférieurs à -300 mV). Elles colonisent ainsi uniquement des environnements très réducteurs, strictement anaérobies où la présence d'accepteurs d'électrons tels que l'O_2, NO_3^-, Fe^{3+} ou encore SO_4^{2-} est limitée (Liu and Whitman 2008). Lorsque des accepteurs d'électrons autres que le CO_2 sont présents, les méthanogènes sont en compétition avec des bactéries qui les utilisent telles que les bactéries sulfato-réductrices (BSR), les dénitrifiantes ou encore les bactéries réductrices du fer. Ces composés étant de meilleurs accepteurs d'électrons, et leur réduction étant thermodynamiquement plus favorable que celle du CO_2 en CH_4, le développement de ces bactéries est favorisé par rapport à celui des méthanogènes (Liu and Whitman 2008).

Tout comme les *Archaea*, les méthanogènes sont distribuées dans de nombreux environnements, y compris dans des milieux extrêmes (Chaban et al 2006). Ils peuvent se développer à des températures de 4°C à 110°C (Franzmann et al 1997) (Kurr et al 1991), à des pH de 5 à 9 (Mathrani et al 1988, Patel et al 1990) ou encore en présence de fortes concentrations en NaCl (Proctor et al 1997).

1.2.2.b Métabolisme et production de CH_4 : la méthanogénèse

Malgré leur grande diversité phylogénétique, les archées méthanogènes constituent un groupe métabolique de microorganismes hautement spécialisés utilisant un nombre limité de

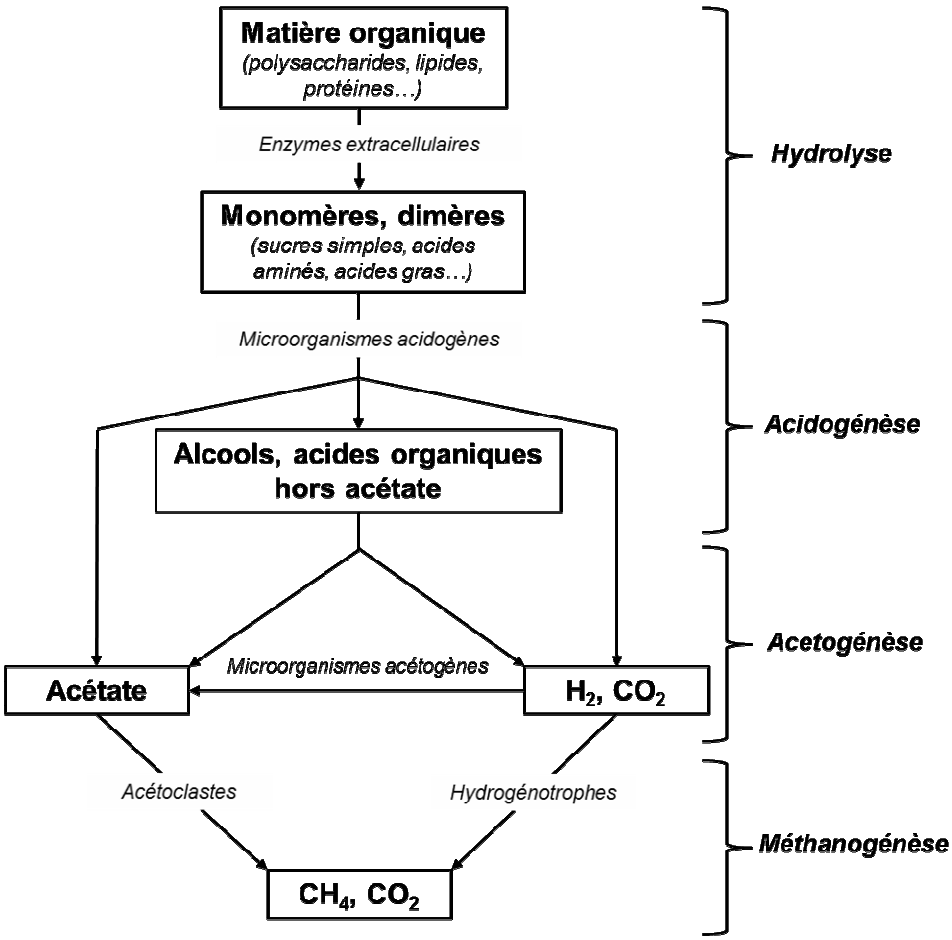

Figure 10. Voie de dégradation anaérobie de la matière organique. (Modifiée et redessinée d'après Moletta, 2002)

substrat pour la production de méthane (méthanogénèse), représentant le produit final de leur respiration anaérobie et fournissant une grande partie de leur énergie. Les substrats utilisés pour la méthanogénèse sont généralement les produits finaux issus de la dégradation anaérobie de la matière organique. Ils sont obtenus par l'association syntrophique de bactéries hydrolytiques et fermentaires, des bactéries acidogènes et acétogènes, des protozoaires et des champignons anaérobies stricts. Ainsi, les archées méthanogènes sont incapables d'utiliser directement des substrats naturellement abondants dans l'environnement tels que les hydrates de carbone, ou encore les acides gras à longue chaîne et se retrouvent donc dépendantes de l'activité métabolique d'autres microorganismes.

i. Dégradation anaérobie de la matière organique

La dégradation de la matière organique comprend une première phase d'hydrolyse où les macromolécules complexes sont clivées sous l'action d'enzymes extracellulaires microbiennes pour être transformées en molécules plus simple facilement assimilables (**Figure 10**) (Moletta 2002). La seconde étape correspond à l'acidogénèse où les monomères issus de l'hydrolyse, ainsi que les composés dissous, servent de substrats à des microorganismes fermentaires. Ceux-ci les dégradent principalement en acides de faibles poids moléculaires comme les acides gras volatils (AGV) tels que le propionate, le butyrate, le valérate, mais également en acides organiques comme le pyruvate, le lactate, ou en alcools tels que le méthanol ou encore l'éthanol (**Figure 10**) (Moletta 2002). Les microorganismes réalisant cette étape peuvent aussi bien être anaérobies facultatifs que strictement anaérobies. La troisième étape est l'acétogénèse où les produits de l'acidogénèse, mais aussi certains résultant directement de l'étape d'hydrolyse, sont réduits en acétate, hydrogène et dioxyde de carbone (**Figure 10**). Cette étape est réalisée par un groupe hétérogène de trois populations bactériennes : (1) les acétogènes syntrophes productrices d'hydrogène (*e.g.* espèces appartenant aux genres *Syntrophomonas*, *Syntrophobacter*, *Syntrophus*…) ; (2) les bactéries acétogènes non-syntrophes parmi lesquelles on distingue les bactéries fermentatives acétogènes (*e.g.* espèces appartenant aux genres *Selomonas*, *Clostridium*, *Ruminococcus*) et (3) les acétogènes hydrogénotrophes ou homoacétogènes (réduisent l'H_2 et le CO_2 en acétate : avec par exemple des espèces appartenant aux genres *Acetogenium*, *Acetobacterium*, *Clostridium*) (Moletta 2002). La dernière étape de la dégradation de la matière organique correspond à la méthanogenèse (**Figure 10**). Elle est réalisée par les *Archaea* et aboutit à la production de CH_4 en utilisant les différents produits issus de la dégradation de la matière organique : le CO_2 / H_2 ou le formate (méthanogénèse hydrogénotrophe), l'acétate

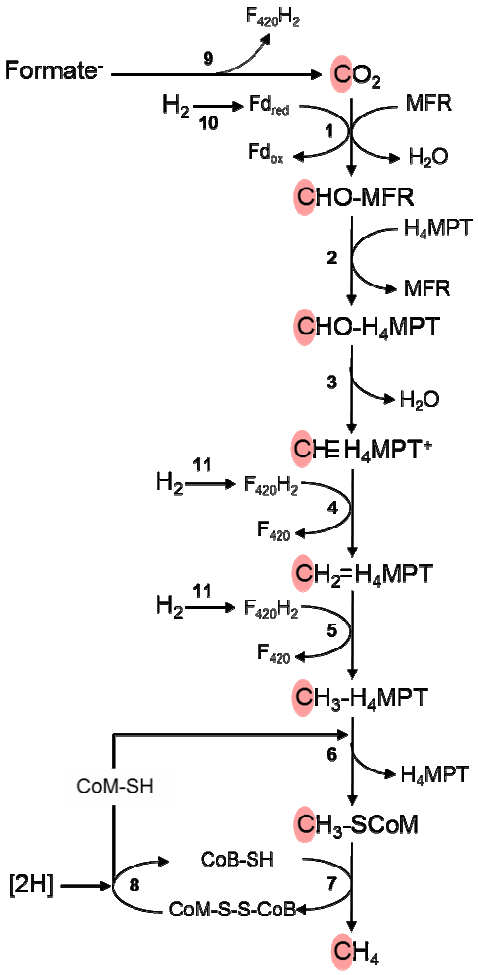

Figure 11. Voie métabolique de la méthanogénèse hydrogénotrophe. Fd_{red} = forme réduite de la ferrédoxine ; Fd_{ox} = forme oxydée de la ferrédoxine ; $F_{420}H_2$ = forme réduite du coenzyme F420 ; MFR = méthanofurane; H_4MPT = tétrahydrométhanopterine; CoM-SH = Coenzyme M; CoB-SH = coenzyme B; CoM-S-S-CoB = hétérodisulfide du CoM et CoB. Enzymes: **(1)** Formyl-MFR deshydrogénase (Fmd); **(2)** Formyl-MFR:H_4MPT formyltransférase (Ftr); **(3)** Méthenyl-H_4MPT cyclohydrolase (Mch); **(4)** Méthylène-H_4MPT deshydrogénase (Hmd); **(5)** Méthylène-H_4MPT réductase (Mer); **(6)** Méthyl-H_4MPT:HS-CoM méthyltransférase (Mtr); **(7)** Méthyl-CoM réductase (Mcr); **(8)** Héterodisulfide réductase (Hdr); **(9)** Formate deshydrogénase (Fdh); **(10)** Energyconserving hydrogénase (Ech); **(11)** F_{420}-reducing hydrogénase. (Redessinée d'après Lui et Whitman, 2008).

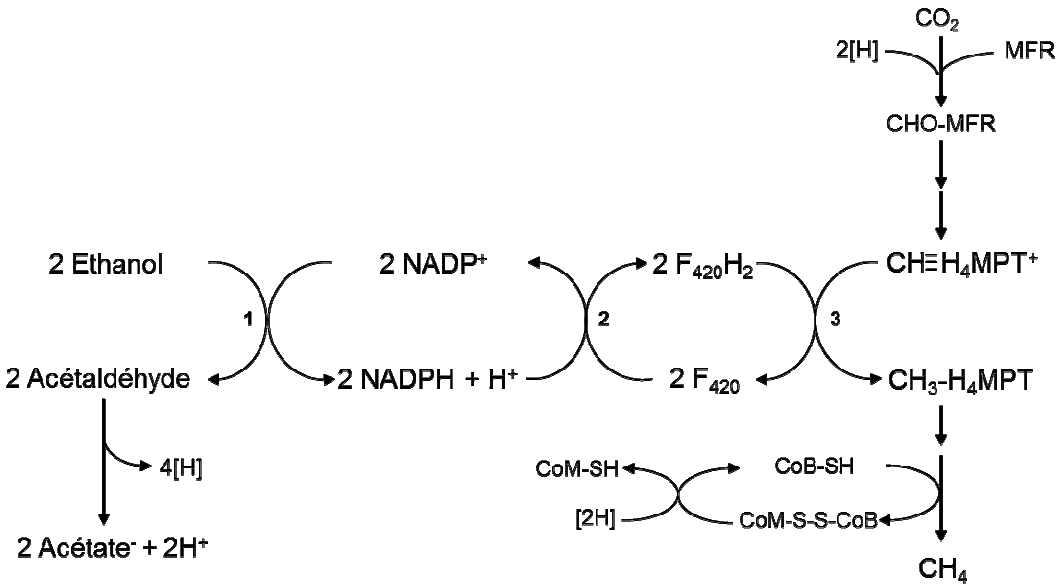

Figure 12. Voie métabolique de la réduction du CO_2 en CH_4 avec des alcools comme donneurs d'électrons chez des méthanogènes contenant une NADP-alcool dépendante deshydrogénase. MFR = Méthanofurane ; CH-OMFR = Formylméthanofurane ; H_4MPT = Tétrahydrométhanoptérine ; $CH \equiv H_4MPT^+$ = N^5, N^{10} – Méthenyltétrahydrométhanoptérine ; CH_3-H_4MPT N^5 = Méthyltétrahydrométhanoptérine ; CoM-SH = Coenzyme M ; CoB-SH = Coenzyme B ; CoM-S-S-CoB = hétérodisulfide du CoM et CoB. (Redessinée d'après Berk et Thauer, 1997).

(méthanogénèse acétoclaste) ou encore les composés méthylés comme le méthanol, les méthylamines et les méthylsulfides (méthanogénèse méthylotrophe) (Liu and Whitman 2008).

ii. La méthanogenèse hydrogénotrophe

La majorité des archées méthanogènes sont hydrogénotrophes et réduisent le CO_2 en CH_4 en utilisant le H_2 comme donneur d'électrons. Au niveau de la méthanogénèse hydrogénotrophe, le CO_2 est donc réduit successivement jusqu'au méthane à travers la formation de groupements formyle (-CHO), méthylène (=CH_2) et méthyle (-CH_3) où le groupement carbone est pris en charge par différents coenzymes à savoir le méthanofurane (MFR), le tetrahydrométhanopterine (H_4MPT) et le coenzyme M (CoM) (**Figure 11**). La méthanogénèse hydrogénotrophe constitue la principale voie de consommation du H_2 produit au cours de la dégradation anaérobie de la matière organique par les bactéries fermentaires. Ces dernières sont organisées avec les archées méthanogènes au sein de *consortia* syntrophiques, permettant de maintenir une pression partielle en H_2 basse, condition nécessaire pour le processus de fermentation (Stams and Plugge 2009).

Certaines espèces de méthanogènes hydrogénotrophes peuvent utiliser le formate pour générer les molécules d'H_2 utilisées comme donneurs d'électrons. Dans ce cas, quatre molécules de formate sont oxydées en CO_2 et permettent la production de quatre molécules d'H_2 par l'intermédiaire de la formate déshydrogénase (Fdh), assurant ainsi la réduction d'une molécule de CO_2 en CH_4 (**Figure 11**). D'autres méthanogènes hydrogénotrophes peuvent utiliser des alcools secondaires (2-propanol, 2-butanol, cyclopentanol) pour synthétiser le donneur d'électrons (H_2) au cours de leur oxydation en cétones via l'alcool secondaire déshydrogénase-coenzyme F_{420} dépendante (Adf) (Aufhammer et al 2004, Widdel and Wolfe 1989). Un plus petit nombre de ces méthanogènes peut également utiliser l'éthanol (Bleicher et al 1989, Widdel 1986) par son oxydation en acétate via la nicotinamide adénine dinucléotide phosphate (NADP)-alcool déshydrogénase dépendante (Berk and Thauer 1997). L'oxydation de ces différents alcools primaires et secondaires aboutit donc à la production de H_2 utilisé directement pour réduire le CO_2 en CH_4 (**Figure 12**). Deux espèces ont également été montrées comme utilisant le monoxyde de carbone CO comme réducteur (Daniels et al 1977, O'Brien et al 1984), où quatre molécules de CO sont oxydées en CO_2 via la CO déshydrogénase (CODH), avant qu'une molécule de CO_2 soit réduite en méthane (Daniels et al 1977). Comme précédemment, le dihydrogène H_2 produit est un intermédiaire au niveau de cette réaction et sert de donneur d'électrons pour la réduction du CO_2 en CH_4.

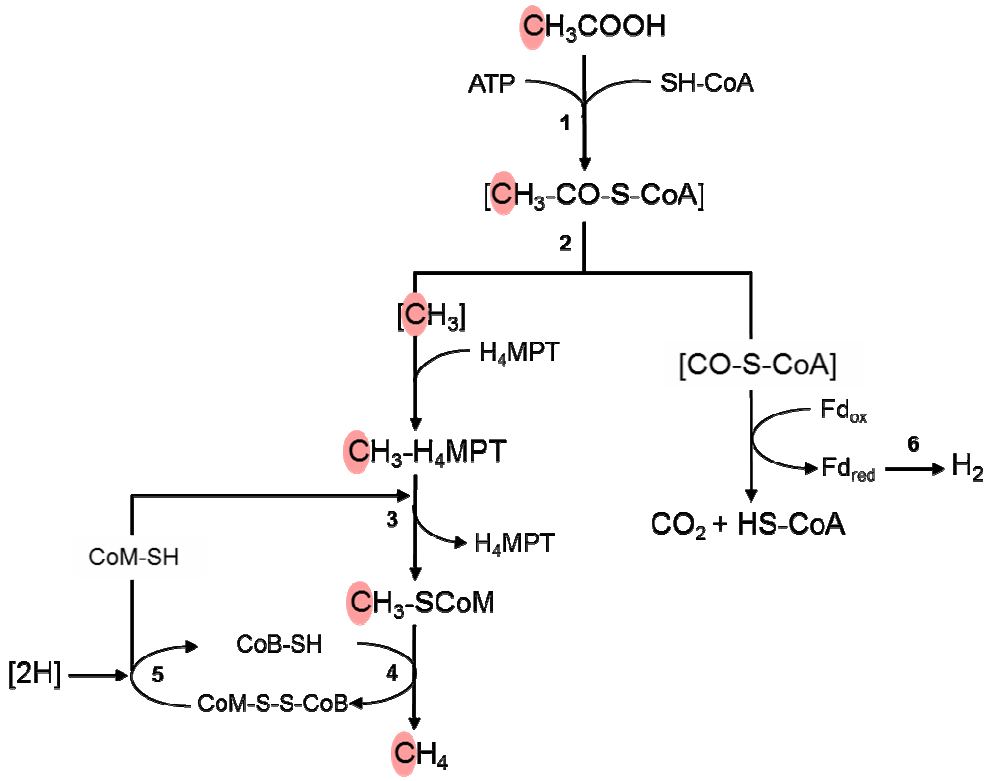

Figure 13. Voie métabolique de la méthanogénèse acétoclaste. Fd_{red} = forme réduite de la ferrédoxine ; Fd_{ox} = forme oxydée de la ferrédoxine ; H_4MPT = tétrahydrométhanopterine; CoM-SH = Coenzyme M; CoB-SH = coenzyme B; CoM-S-S-CoB = hétérodisulfide du CoM et CoB. Enzymes: **(1)** Système acétate kinase (AK)-phosphotransacétylase (PTA) chez *Methanosarcina* ; AMP-forming acétyl-CoA synthétase chez *Methanosaeta*; **(2)** CO deshydrogénase/acetyl-CoA synthase (CODH/ACS) ; **(3)** Méthyl-H_4MPT:HS-CoM méthyltransférase (Mtr); **(4)** Méthyl-CoM réductase (Mcr); **(5)** Héterodisulfide réductase (Hdr); **(6)** Energyconserving hydrogénase (Ech). (Redessiné d'après Liu et Whitman, 2008)

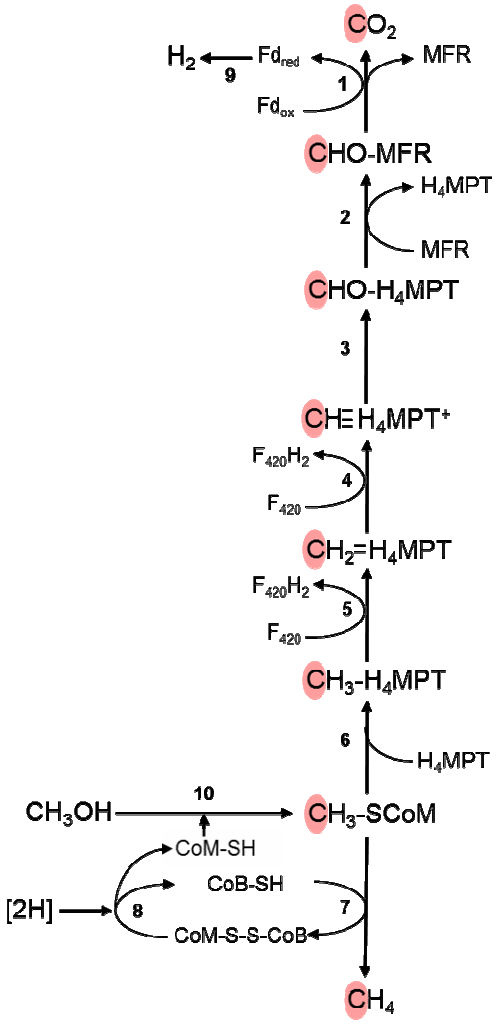

Figure 14. Voie métabolique de la méthanogénèse méthylotrophe. Fd_{red} = forme réduite de la ferrédoxine ; Fd_{ox} = forme oxydée de la ferrédoxine ; $F_{420}H_2$ = forme réduite du coenzyme F420 ; MFR = méthanofurane; H_4MPT = tetrahydrométhanopterine; CoM-SH = Coenzyme M; CoB-SH = coenzyme B; CoM-S-S-CoB = hétérodisulfide du CoM et CoB. Enzymes: **(1)** Formyl-MFR deshydrogénase (Fmd); **(2)** Formyl-MFR:H_4MPT formyltransférase (Ftr); **(3)** Méthenyl-H_4MPT cyclohydrolase (Mch); **(4)** Méthylène-H_4MPT deshydrogénase (Hmd); **(5)** Methylène-H_4MPT réductase (Mer); **(6)** Méthyl-H_4MPT:HS-CoM méthyltransférase (Mtr); **(7)** Méthyl-CoM réductase (Mcr); **(8)** Héterodisulfide réductase (Hdr); **(9)** Energyconserving hydrogénase (Ech); **(10)** Méthyltransférase. (Redessiné d'après Lui et Whitman, 2008).

iii. La méthanogénèse acétoclaste

L'acétate est un intermédiaire majeur au niveau de la dégradation anaérobie de la matière organique où ce dernier est responsable de deux tiers environ de la production biologique du CH_4 (Conrad 1999, Liu and Whitman 2008). Paradoxalement, seuls les deux genres *Methanosarcina* et *Methanosaeta* appartenant à l'ordre des *Methanosarcinales*, sont capables d'utiliser l'acétate pour la méthanogénèse. Au niveau de cette voie de production, la molécule d'acétate est scindée en deux, le groupement carboxyle (-COOH) est oxydé en CO_2 et le groupement méthyle réduit en CH_4 (**Figure 13**). D'une manière générale, les espèces affiliées au genre *Methanosarcina* utilisent préférentiellement le méthanol et les méthylamines par rapport à l'acétate, et certaines espèces utilisent également le dihydrogène H_2. Les espèces affiliées au genre *Methanosaeta* utilisent quant à elles uniquement l'acétate à des concentrations très basses allant de 5 à 20 µM, tandis que les espèces du genre *Methanosarcina* requièrent une concentration minimale de 1 mM (Jetten et al 1992). Cette différence d'affinité pour l'acétate est potentiellement due à des différences dans les premières étapes du métabolisme acétoclaste. Le genre *Methanosarcina* utilise une acétate kinase-phosphotransacétylase (AK-PTA), de faible affinité pour l'acétate permettant son activation en acétyl-CoA, tandis que le genre *Methanosaeta* utilisent une adénosine monophosphate-forming acetyl-CoA synthétase de haute affinité pour l'acétate (Jetten et al 1992, Smith and Ingram-Smith 2007).

iv. La méthanogénèse méthylotrophe

Les archées au niveau de cette voie utilisent le groupement méthyle ($-CH_3$) porté par différents composés incluant le méthanol, les méthylamines (mono-, di-, triméthylamine et tetraméthylammonium) ainsi que les sulfures de méthyle (méthanethiol et sulfure de diméthyle). La méthanogénèse méthylotrophe est réalisée par des archées appartenant à l'ordre des *Methanosarcinales*, à l'exception du genre *Methanosphaera* et des espèces du genre *Methanomicrobium* appartenant à l'ordre des *Methanobacteriales*. Au cours de cette voie de production du méthane, les groupements méthyles sont transférés à une « cognate corrinoïd protein » puis au coenzyme M (CoM), où le méthylcoenzyme M ainsi formé est réduit en méthane (**Figure 14**). Chez la plupart des méthanogènes méthylotrophes, les électrons requis pour la réduction des groupements méthyles en méthane sont obtenus par l'oxydation de groupements méthyles additionnels en CO_2 via une méthanogénèse hydrogénotrophe réalisée de manière inversée (**Figure 14**). Par cette voie de synthèse du méthane, trois groupements méthyles sont réduits en méthane pour chaque molécule de CO_2

formée (Liu and Whitman 2008). Ce processus est une dismutation puisque l'oxydation d'une partie du substrat est utilisée pour la réduction de l'autre partie. A la différence de cette méthanogénèse méthylotrophe basée sur un principe de dismutation, celle de *Methanomicrococcus blatticola* et des espèces affiliées au genre *Methanosphaera* est H_2 dépendante (Miller and Wolin 1985, Sprenger et al 2000). Ce sont des espèces méthylotrophes et hydrogénotrophes obligatoires spécialisées dans la réduction des groupements méthyles (issus du méthanol uniquement pour *Methanosphaera*, et à la fois du méthanol et des méthylamines pour *M. blatticola*) avec du H_2.

Contrairement à la méthanogénèse hydrogénotrophe et acétoclaste, la voie méthylotrophe empruntée par les méthanogènes n'entraîne pas de compétition avec d'autres microorganismes comme les BSR qui n'utilisent pas le méthanol et les méthylamines comme substrat (Oremland et al 1982, Oremland and Polcin 1982). Les méthanogènes méthylotrophes peuvent donc coloniser des environnements où il est possible également de retrouver d'autres accepteurs d'électrons (comme SO_4^{2-}) et donc la présence de BSR.

1.2.2.c La méthyl-coenzyme M réductase (MCR) : enzyme clé de la méthanogénèse

Les trois voies de la méthanogénèse impliquent plusieurs enzymes et coenzymes différentes, mais toutes ces voies de synthèse du méthane partagent une étape commune catalysée par un complexe enzymatique, la méthyl-coenzyme M réductase (MCR), appelée également coenzyme-B sulfoéthylthiotransférase, souvent considérée comme l'enzyme clé de la méthanogénèse (Ermler 1997). Elle catalyse la réduction d'un groupement méthyle lié au coenzyme M avec une libération concomitante de CH_4 (Thauer 1998). Cette enzyme possède une masse moléculaire d'environ 300 kDa et elle est composée de trois différentes sous unités, α (McrA), β (McrB) et γ (McrG) arrangées dans une conformation $\alpha_2\beta_2\gamma_2$. Cette enzyme est retrouvée chez tous les méthanogènes à l'exception de *Methanosphaera stadtmanae* (ordre des *Methanobacteriales*) qui possède uniquement une isoforme de MCR, appelée MRT. Cette isoforme est également retrouvée en plus de MCR chez les méthanogènes affiliées à l'ordre des *Methanococcales* et des *Methanobacteriales* (Brenner et al 1993, Lehmacher and Klenk 1994, Nölling et al 1996, Rospert et al 1990, Thauer 1998). Il a été montré que l'expression de ces deux isoenzymes est différentiellement régulée suivant les conditions de croissance (Bonacker et al 1992, Pennings et al 1997, Pihl et al 1994, Reeve et al 1997)

Les gènes codants pour les trois sous unités de l'isoforme I (MCR) sont organisés en opéron (*mcrBDCGA*), contenant également deux gènes (*mcrC* et *mcrD*) codant pour deux protéines, McrD et McrC de fonction inconnue (Reeve et al 1997). La protéine McrC, pourrait être impliquée dans des modifications post-traductionnelles de la sous-unité α (Ermler 1997). Chez les archées méthanogènes affiliées aux *Methanobacteriales* et *Methanococcales*, les gènes codant pour les trois sous-unités de l'isoforme II (MRT) sont aussi organisés en opéron (*mrtBDGA*) (Lehmacher and Klenk 1994, Pihl et al 1994), contenant également un gène supplémentaire (*mrtD*) codant pour une protéine dont la séquence est relativement proche de McrD (Nölling et al 1996). Chez *Methanococcus jannaschii* (ordre des *Méthanococcales*), les gènes sont organisés sous la forme *mrtBGA* avec le gène *mrtD* distant d'environ 37kpb de l'opéron *mrt*.

Les gènes *mcrA* et *mrtA* apparaissent être très conservés (Lehmacher and Klenk 1994, Nölling et al 1996, Springer et al 1995), permettant de les utiliser comme biomarqueur phylogénétique des archées méthanogènes. En effet, leur évolution relativement lente permet de disposer de séquences montrant suffisamment de similarités pour pouvoir étudier les liens de parenté d'espèces phylogénétiquement distantes, et ceci de manière congruente aux phylogénies basées sur les séquences codantes pour l'ARNr 16S (Lueders et al 2001, Luton et al 2002, Springer et al 1995). Les gènes *mcrA* et *mrtA* ont été utilisés dans de nombreuses études moléculaires pour caractériser la diversité, la structure et la distribution des communautés méthanogènes, et ceci dans de nombreux environnements (Biderre-Petit et al 2011, Castro et al 2004, Chin et al 2004, Earl et al 2003, Galand et al 2002, Hales et al 1996, Juottonen et al 2006, Lueders et al 2001, Mihajlovski et al 2008, Milferstedt et al 2010).

1.2.3 Consommation microbienne du méthane

La majorité du CH_4 produit n'est pas libérée dans l'atmosphère mais est oxydée à proximité de ses zones de production. Par exemple pour le CH_4 produit par les rizières, le pourcentage de méthane libéré dans l'atmosphère par rapport à la production totale ne dépasserait pas 17% (Cheng et al 2005). Il en est de même pour le CH_4 provenant de sédiments marins ou lacustres dont la fraction émise est très faible, du fait d'une consommation anaérobie dans les sédiments et une consommation aérobie dans la colonne d'eau. Cette consommation du méthane est assurée par des bactéries dites méthanotrophes et pour lesquelles le CH_4 est une source de carbone et d'énergie. Celles-ci jouent donc un rôle primordial dans la régulation des émissions de méthane (Hanson and Hanson 1996).

Tableau 2. Caractéristiques phylogénétiques et physiologiques des bactéries méthanotrophes aérobies. Les méthanotrophes facultatifs sont indiqués en gras.

	γ-Protéobactéries (Type I)	α-Protéobactéries (Type II)		Verrucomicrobia
FAMILLE	*Methylococcaceae*	*Methylocystaceae*	*Beijerinckiaceae*	*Methylacidiphilaceae*
GENRE	*Methylomonas* *Methylobacter* *Methylomicrobium* *Methylosarcina* *Methylosphaera* *Methylosoma* *Methylococcus* *Methylocaldum* *Methylothermus* *Methylohalobium* ***Crenothrix*** *Clonothrix*	*Methylosinus* ***Methylocystis***	***Methylocapsa*** ***Methylocella***	*Methylacidiphilum*
ACIDES GRAS DES PHOSPHOLIPIDES	C16:1ω7c, C16:1ω8c, C16:0, C14:0	C18:1ω8c, C18:1ω7c, C18:2ω7c,12c	C18:1ω7c	C18:0, C16:0, aC15:0. C14:0
Activité				
sMMO	Oui/Non	Oui/Non	Oui/Non	Non
pMMO	Oui	Oui	Oui/Non	Oui
VOIE D'ASSIMILATION DU CARBONE	Ribulose monophosphate	Voie de la sérine	Voie de la sérine	Voie alternative de la sérine

Actuellement, deux voies biologiques principales sont impliquées dans l'oxydation du méthane avec (1) l'oxydation du méthane en condition aérobie et (2) l'oxydation anaérobie du méthane réalisée par des microorganismes spécialisés et récemment identifiés au cours de ces quinze dernières années (Hinrichs et al 1999, Raghoebarsing et al 2006).

1.2.3.a Oxydation aérobie du méthane

Le processus d'oxydation du méthane, réalisé par des bactéries dites méthanotrophes, est connu depuis longtemps, avec des premiers isolats caractérisés au début du XX^{ème} siècle (Kaserer 1905, Sohngen 1906). Ces dernières années, un nouvel intérêt est apparu pour leur rôle primordial au niveau du changement global.

i. Les bactéries méthanotrophes

Actuellement, les bactéries méthanotrophes appartiennent à 16 genres bactériens répartis au sein des α- et γ- protéobactéries (**Tableau 2**). Plus récemment, trois méthanotrophes obligatoires appartenant au phylum des *Verrucomicrobia* ont été mis en évidence (Dunfield et al 2007, Islam et al 2008, Pol et al 2007). Ces dernières semblent restreintes à des environnements extrêmes, avec des conditions de croissance à des pH d'environ 1 et des températures supérieures à 50°C. Le séquençage complet du génome d'un des isolats a montré l'acquisition des gènes de la méthanotrophie par transfert horizontal (Hou et al 2008). Les méthanotrophes appartenant aux protéobactéries peuvent être classées au niveau de la famille (**Tableau 2**) avec (1) les *Methylococcaceae* connus également comme méthanotrophes de type I, qui utilisent la voie du ribulose monophosphate pour la fixation du carbone et ont une composition de leurs phospholipides avec des acides gras à longue chaîne (16 carbones) ; (2) les *Methylocystaceae* (ou méthanotrophes de type II) utilisant la voie de la sérine pour la fixation du carbone avec une composition des phospholipides contenant des acides gras à 18 carbones, excepté pour l'espèce *Methylocystis heyeri* contenant des acides gras à 16 carbones (Dedysh et al 2007) ; (3) les *Beijerinckiaceae* possédant des métabolismes diversifiés avec des méthanotrophes obligatoires, non obligatoires et facultatifs (Dunfield et al 2010). A l'image des méthanotrophes appartenant aux *Verrucomicrobia*, des espèces appartenant aux Protéobactéries (e.g. celles appartenant aux genres *Methylococcus*, *Methylocaldum* et *Methylothermus*) sont capables de s'adapter à des environnements extrêmes avec une température de croissance pouvant atteindre 65°C (Trotsenko et al 2009). De même, certaines espèces sont adaptées à des environnements présentant des températures oscillant entre 0 et 30°C (Trotsenko and Khmelenina 2005, Wartiainen 2006), de fortes concentrations

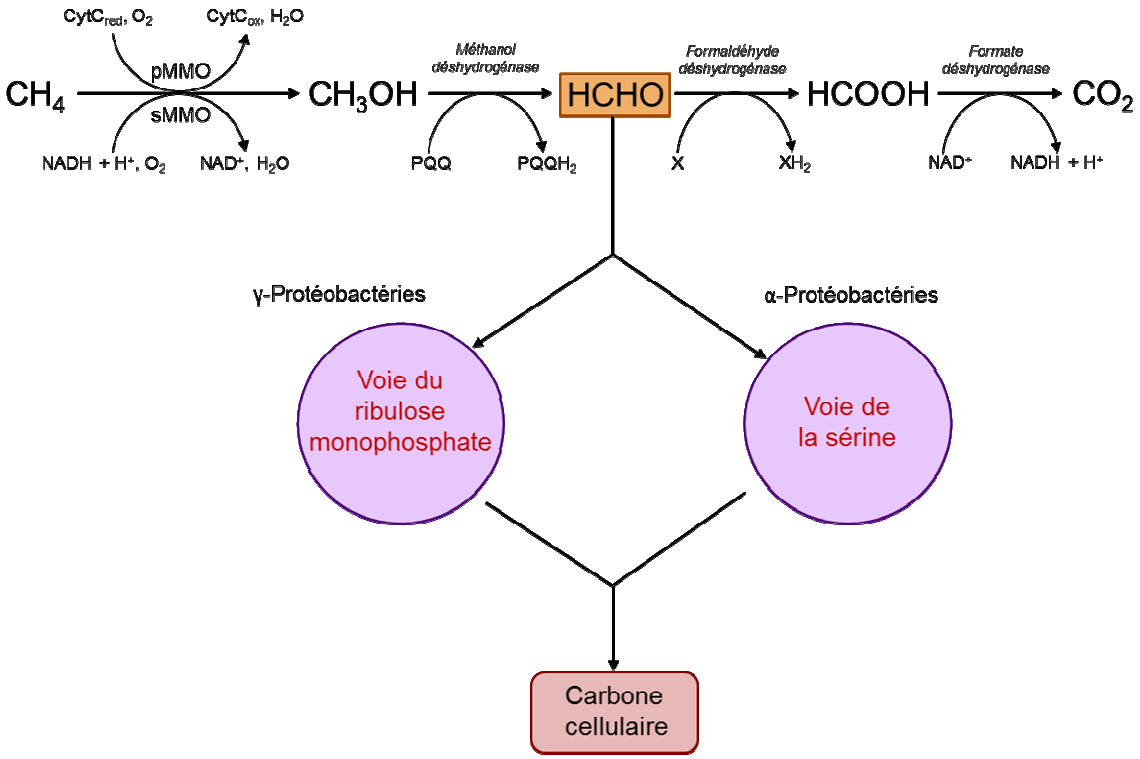

Figure 15. Voie métabolique de l'oxydation aérobie du méthane et assimilation du formaldéhyde. CytC = Cytochrome C ; PQQ = pyrroloquinoline quinine; X = $NADP^+$ ou cytochrome lié. (Redessiné d'après Hanson et Hanson, 1996).

en sels entre 0.2 et 2.5 M (Heyer 2005) et des pH relativement bas (Dedysh et al 2000, Dedysh et al 2002).

ii. Métabolisme aérobie

Les méthanotrophes aérobies utilisent des composés à un carbone pour leur métabolisme à savoir le méthane (CH_4) et le méthanol (CH_3OH). Ce n'est pas le cas des méthanotrophes facultatifs, récemment isolés et appartenant notamment aux genres *Methylocella* (Dedysh et al 2005), *Crenothrix (Stoecker 2006)*, *Methylocystis* (Belova et al 2011, Im et al 2011) ou encore *Methylocapsa* (Dunfield et al 2010), qui eux peuvent utiliser en plus d'autres substrats initiaux à un ou plusieurs carbones. La voie métabolique d'oxydation aérobie du méthane comprend quatre étapes, où le CH_4 est oxydé en CO_2 via la formation d'intermédiaires comme le méthanol, le formaldéhyde et le formate (**Figure 15**). La première étape (oxydation du méthane en méthanol) est catalysée par la méthane monooxygénase (MMO). Cette dernière est constituée de trois composantes : une hydroxylase composée de trois sous unités arrangées dans une conformation $\alpha_2\beta_2\gamma_2$ (MMOH), une composante B et une réductase (Lipscomb 1994). Deux formes de cette enzyme ont été identifiées à des localisations cellulaires différentes, l'une dite soluble et à localisation cytoplasmique (sMMO) et l'autre dite particulaire et liée à la membrane intracytoplasmique (pMMO) (Lipscomb 1994). Tous les méthanotrophes possèdent la forme pMMO excepté celles appartenant aux genres *Methylocella* (Dedysh et al 2000) et *Methyloferula* (Vorobev et al 2010) qui possèdent par contre la forme sMMO. Cette forme soluble et cytoplasmique de la méthane monooxygénase n'est cependant pas retrouvée chez tous les méthanotrophes (Murrell et al 2000). La forme sMMO est capable d'utiliser un large spectre de substrats incluant des alcanes, alcènes et des composés aromatiques tandis que la forme pMMO est seulement capable d'oxyder le méthane, des alcanes et des alcènes à chaînes constituées au plus de 5 carbones (Burrows et al 1984, Colby et al 1977).

A l'image de la méthyl coenzyme M réductase (MCR), la méthane monoxygénase est ubiquiste chez les méthanotrophes. Le gène *pmoA* codant pour la sous unité α de la pMMO, est donc utilisé comme marqueur fonctionnel de la consommation aérobie du méthane (McDonald et al 2008). Pour la plupart des méthanotrophes appartenant aux α- et γ-Protéobactéries, la phylogénie basée sur *pmoA* est congruente avec celle basée sur l'ARNr 16S puisque qu'aucun transfert horizontal de ce gène n'a été mis en évidence chez ces espèces (Bourne et al 2001, Costello and Lidstrom 1999, Horz et al 2005, Murrell et al 1998). Il faut

cependant noter la présence d'un gène *pmoA* particulier chez les membres du genre *Crenothrix* (Stoecker 2006).

Bien que des analyses phylogénétiques basées sur le gène *pmoA* aient permis de conclure à une divergence ancestrale entre les méthanotrophes affiliés aux *Verrucomicrobia* et aux Protéobactéries (Dunfield et al 2007, Op den Camp et al 2009), le séquençage complet du génome d'un isolat appartenant au phylum des *Verrucomicrobia* a cependant montré l'acquisition des gènes de la méthanotrophie par transfert horizontal à partir des Protéobactéries (Hou et al 2008).

1.2.3.b Oxydation anaérobie du méthane

Pendant longtemps, la consommation aérobie du méthane a été considérée comme la voie unique d'utilisation du CH_4. Cependant, depuis environ trente ans, suite aux premières études géochimiques une oxydation anaérobie du méthane dans des sédiments marins et des colonnes d'eau anoxiques a été démontrée (Barnes and Goldberg 1976, Martens and Berner 1974, Reeburgh 1976). L'implication de microorganismes dans ce processus n'a été révélée que récemment. L'oxydation anaérobie du méthane fait intervenir trois types de métabolisme avec (1) une oxydation couplée à la sulfato-réduction ; (2) couplée à la dénitrification et (3) couplée à la réduction du fer et du manganèse.

i. Oxydation anaérobie couplée à la sulfato-réduction

Les premières observations montrant des concentrations en méthane et sulfate variant de manière inversée entre les sédiments et les colonnes d'eau, ont permis de proposer que le méthane était oxydé en anaérobiose par un consortium d'*Archaea* et de bactéries sulfato-réductrices (Barnes and Goldberg 1976, Martens and Berner 1974, Reeburgh 1976). Dans cette association de type syntrophique, les BSR utiliseraient le dihydrogène produit lors de l'oxydation du méthane par des archées méthanogènes, qui réaliseraient une méthanogénèse dite inverse de la réduction du CO_2 en CH_4 (Hoehler et al 1994, Zehnder and Brock 1979). Des analyses phylogénétiques basées sur des séquences codantes pour l'ARNr 16S montrent qu'il existe au moins trois groupes d'*Archaea* associées à l'oxydation anaérobie du méthane (« Anerobic Methanotrophs » ANME-1, -2 ou -3). Le groupe ANME-2 affilié aux *Methanosarcinales* et le groupe ANME-1 se positionnant entre les méthanogènes acétoclastes et non acétoclastes sont les plus fréquemment rencontrés. Le groupe ANME-3 est essentiellement apparenté aux *Methanococcoides* spp (Hinrichs et al 1999, Knittel et al 2005). Ces trois différents groupes d'ANME ne sont pas monophylétiques et montrent une similarité

de séquences de 75-92%. Quant aux BSR partenaires des *Archaea* ANME, ces dernières sont majoritairement affiliées aux deltaprotéobactéries et plus particulièrement à l'ordre des *Desulfosarcinales* et *Desulfobacterales* (Orphan 2001, Orphan et al 2009, Strous and Jetten 2004).

Le fonctionnement exact de la « méthanogénèse inverse » n'est pas encore totalement connu. Une des questions porte sur la nature du composé échangé entre les archées méthanogènes et les BSR. Certaines hypothèses placent l'acétate et le formate comme métabolites potentiels pour plusieurs populations de BSR. D'autres placent le CO_2 (en plus du H_2 produit par les *Archaea*) présent dans l'environnement comme source de carbone pour les BSR ou encore le sulfure de méthyle (CH_3SH) produit par les *Archaea* comme donneur d'électrons (Hoehler et al 1994, Moran et al 2007, Valentine 2002). En plus de l'étroite relation phylogénétique entre les ANME et les archées méthanogènes, d'autres similitudes d'ordre génomique ont été mises en évidence. Le gène *mcrA* codant pour la sous-unité α de la méthyl-coenzyme M réductase a été montré comme étant associé à la communauté des ANME (Hallam et al 2003), et une étude a montré que tous les gènes de la méthanogénèse étaient présents au sein du génome d'*Archaea* ANME-1 provenant de sédiments marins (Hallam 2004). La présence des gènes de la méthanogénèse chez les ANME suggère une oxydation anaérobie du méthane réalisée par méthanogénèse inverse (Knittel and Boetius 2009, Krüger et al 2003, Meyerdierks et al 2010).

ii. Oxydation anaérobie couplée à la dénitrification, réduction du fer et du manganèse

Actuellement, il est établi que l'oxydation anaérobie du méthane peut impliquer des accepteurs terminaux d'électrons différents des sulfates tels que par exemple les nitrates, le fer et le manganèse. Un couplage entre l'oxydation anaérobie du méthane et la dénitrification a été premièrement mis en évidence au niveau d'un aquifère contaminé par des nitrates (Smith et al 1991), puis confirmé plus tard expérimentalement sur une étude en réacteurs (Islas-Lima et al 2004). Plus récemment une étude a mis en évidence un *consortium* microbien couplant la dénitrification avec l'oxydation anaérobie du méthane au niveau de sédiments lacustres ayant de fortes concentrations en nitrates pouvant atteindre 1 mM (Raghoebarsing et al 2006). Ce *consortium* (contenant également des *Archaea* proches des *Methanosarcinales* et des ANME-2) était dominé par une division candidate bactérienne nommée NC10 et plus précisément par la bactérie *Candidatus Methylomirabilis oxyfera*, dont le séquençage du génome a révélé une

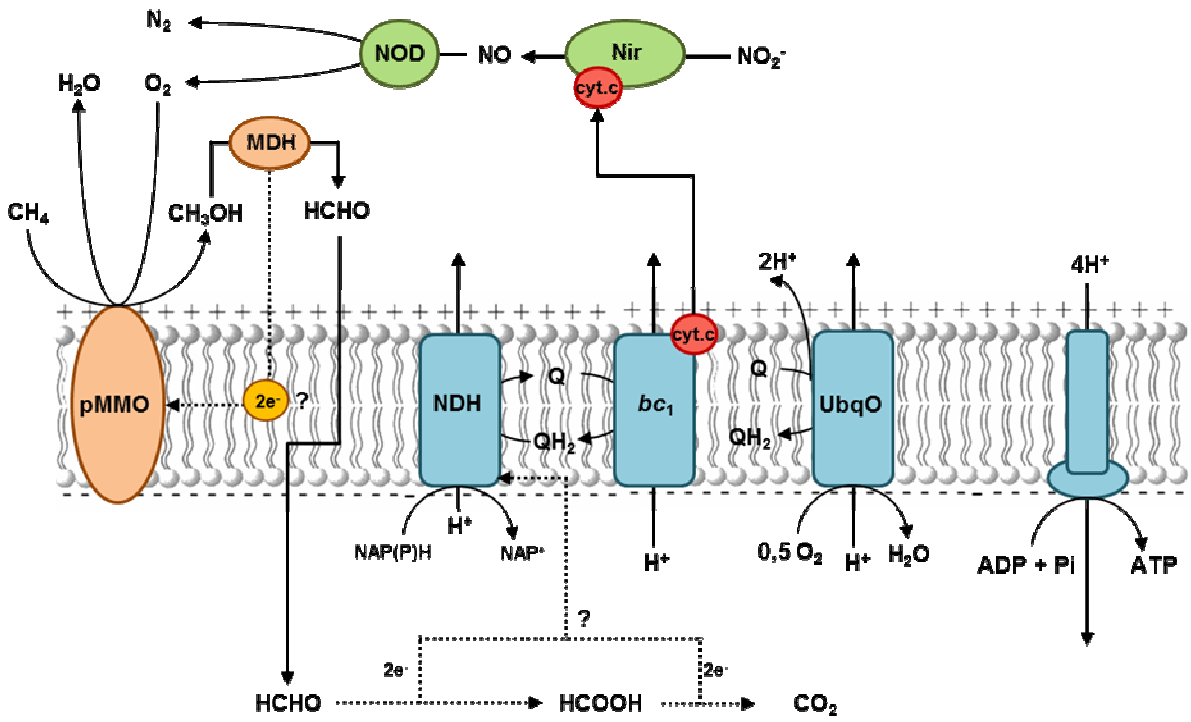

Figure 16. Métabolisme énergétique et catabolisme central chez *Candidatus Methylomirabilis oxyfera*. Les flèches en pointillés indiquent des voies métaboliques incomplètes ; en orange voie d'oxydation du méthane ; en vert voie de réduction des nitrites. Cyt.c = cytochrome C; FDH = formate deshydrogénase; MtdB, MDH = complexe NAD(P)H deshydrogénase; Nir = nitrite réductase; NOD = NO dismutase; pMMO = particulate méthane mono-oxygénase; UbqO = ubiquinol oxidase. (Redessiné et modifié d'après Wu et al., 2011).

nouvelle voie métabolique de dénitrification dite « intra-aérobique » (**Figure 16**) (Ettwig et al 2010, Raghoebarsing et al 2006). De manière très intéressante, ce processus se comporte comme une oxydation aérobie classique du méthane en l'absence totale d'O_2 apporté de manière externe (Ettwig et al 2010). Ce paradoxe biochimique est expliqué par la capacité unique de *Candidatus Methylomirabilis oxyfera* à produire du dioxygène intracellulaire via une voie alternative de dénitrification n'impliquant pas l'oxyde nitreux (N_2O) comme intermédiaire. Dans cette voie particulière, l'oxyde nitrique (NO) généré lors de la réduction des nitrites (NO_2^-) est dismuté en diazote (N_2) et dioxygène (O_2) (**Figure 16**) (Ettwig et al 2010, Wu et al 2011). La majorité de l'O_2 produit est utilisé pour activer et oxyder le CH_4 dans une réaction strictement aérobie et catalysée par la pMMO. Le dioxygène restant est utilisé pour d'autres processus biochimiques comme une respiration utilisant des terminales O_2 réductases (TOR) suggérant une co-respiration chez *Candidatus Methylomirabilis oxyfera* utilisant les nitrites et le dioxygène (Luesken et al 2012, Wu et al 2010). Ainsi l'utilisation du gène *pmoA* peut constituer également un marqueur pour l'étude des communautés microbiennes oxydatrices du méthane en anaérobiose couplée à la dénitrification (Luesken et al 2011).

L'oxydation anaérobie du méthane a de même été mise en évidence couplée à la réduction du fer et du manganèse dans des sédiments marins profonds impliquant des communautés d'*Archaea* non cultivées et des ANME-1 et -2 (Beal et al 2009). Cependant les processus biochimiques mis en jeu de même que les communautés bactériennes impliquées, restent encore peu connus.

1.3 La biodégradation des hydrocarbures aromatiques polycycliques en conditions méthanogènes

L'exploitation intensive du carbone fossile par l'Homme aboutit à la contamination des différents compartiments (biosphère, lithosphère, hydrosphère et atmosphère) par des composés toxiques comme les hydrocarbures aromatiques polycycliques (HAP). Ces derniers peuvent affecter les compartiments anoxiques largement répandus dans l'environnement comme dans les sédiments, les eaux ou encore les sols. L'étude de la participation des communautés méthanogènes pour l'étude de la biodégradation anaérobie de ces composés néfastes apparaît être primordiale pour appréhender les différents processus métaboliques mis en jeu.

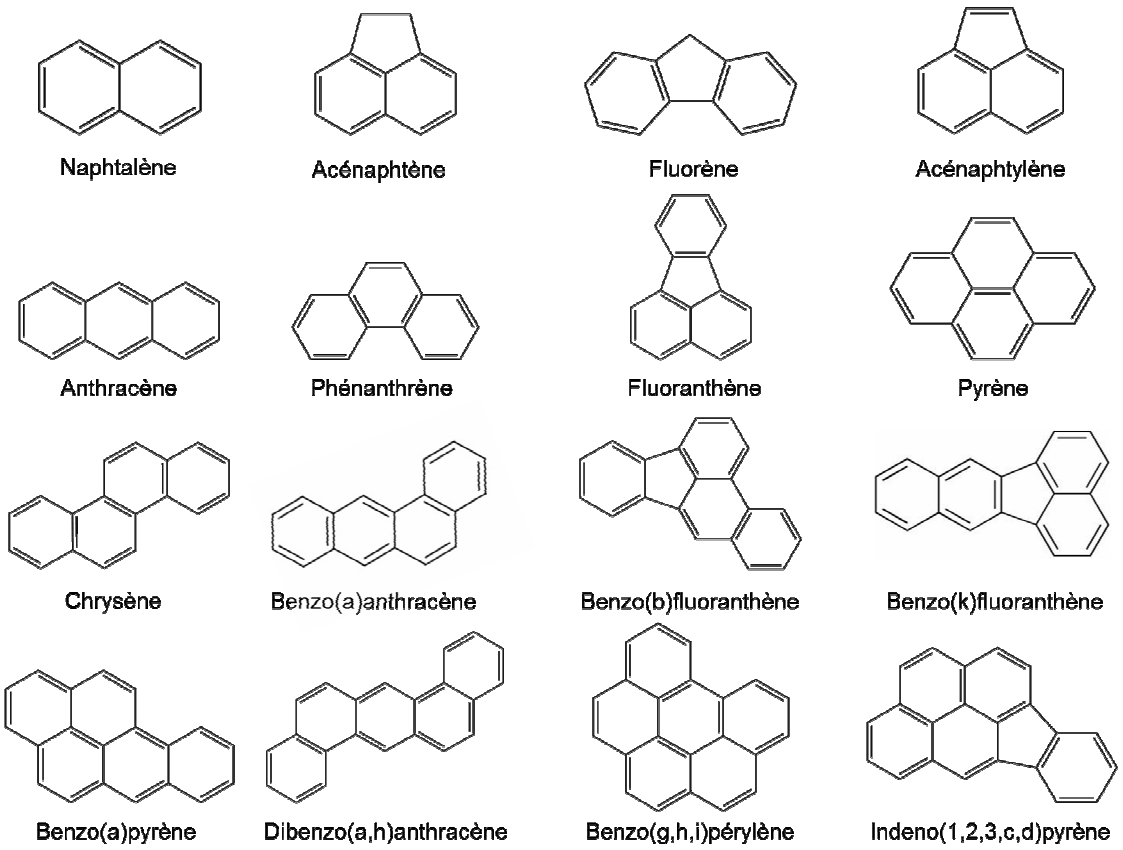

Figure 17. Liste et formule topologique des 16 HAP classés prioritaires selon l'Agence Américaine de Protection de l'Environnement (US Environmental Protection Agency, EPA).

Tableau 3. Propriétés physico-chimique des 16 HAP classés prioritaires par l'US-EPA (d'après ATSDR 2005: Toxicology profile for polyaromatic hydrocarbons. ATSDR's Toxicological Profiles on CD-ROM, CRC Press, Boca Raton, FL).

	Nombre de cycles	Formule chimique	Masse molaire ($g.mole^{-1}$)	Solubilité dans l'eau ($mg.L^{-1}$)	Pression de vapeur (mm Hg)	Log K_{ow}
Naphtalène	2	$C_{10}H_8$	128,17	31	$8,89.10^{-2}$	3,37
Acénaphtène	3	$C_{12}H_{10}$	154,21	3,8	$3,75.10^{-3}$	3,92
Fluorène	3	$C_{13}H_{10}$	166,22	1,9	$3,24.10^{-3}$	4,18
Acénaphtylène	3	$C_{12}H_{18}$	152,20	16,1	$2,90.10^{-2}$	4,00
Anthracène	3	$C_{14}H_{10}$	178,23	0,045	$2,55.10^{-5}$	4,54
Phénanthrène	3	$C_{14}H_{10}$	178,23	1,1	$6,80.10^{-4}$	4,57
Fluoranthène	4	$C_{16}H_{10}$	202,26	0,26	$8,13.10^{-6}$	5,22
Pyrène	4	$C_{16}H_{10}$	202,26	0,132	$4,25.10^{-6}$	5,18
Chrysène	4	$C_{18}H_{12}$	228,29	0,0015	$7,80.10^{-9}$	5,91
Benzo(a)anthracène	4	$C_{18}H_{12}$	228,29	0,011	$1,54.10^{-7}$	5,91
Benzo(b)fluoranthène	5	$C_{20}H_{12}$	252,32	0,0015	$8,06.10^{-8}$	5,80
Benzo(k)fluoranthène	5	$C_{20}H_{12}$	252,32	0,0008	$9,59.10^{-11}$	6,00
Benzo(a)pyrène	5	$C_{20}H_{12}$	252,32	0,0038	$4,89.10^{-9}$	5,91
Dibenzo(a,h)anthracène	6	$C_{22}H_{14}$	278,35	0,0005	$2,10.10^{-11}$	6,75
Benzo(g,h,i)pérylène	6	$C_{22}H_{12}$	276,34	0,00026	$1,00.10^{-10}$	6,50
Indeno(1,2,3,c,d)pyrène	6	$C_{22}H_{12}$	276,34	0,062	$1,40.10^{-10}$	6,50

1.3.1 Les hydrocarbures aromatiques polycycliques

Les HAP font partie des contaminants persistants auxquels leurs origines multiples (naturelles ou anthropiques), leur confèrent un caractère ubiquiste dans l'environnement. Actuellement, 16 HAP sont répertoriés (**Figure 17**) et figurent sur la liste des polluants prioritaires de l'Agence Américaine de Protection de l'Environnement (US Environmental Protection Agency, EPA) (Keith and Telliard 1976). Ils sont également inventoriés comme des composés néfastes pour l'Homme et l'environnement par l'Organisation Mondiale de la Santé (OMS) et la Communauté Européenne.

1.3.1.a Structure et propriétés physico-chimiques

Les HAP sont constitués d'atomes de carbone et d'hydrogène dont la structure des molécules comprend au moins deux cycles aromatiques assemblés de manière linéaire (anthracène), angulaire (phénanthrène) ou condensée (pyrène) (**Figure 17**). Suivant le nombre de cycles composant les HAP, ils peuvent être qualifiés de légers (trois cycles au plus) ou de lourds (plus de quatre cycles) (Doyle et al 2008). L'énergie de résonance créée par le nuage dense d'électrons π stabilise les HAP les rendant résistants aux attaques nucléophiles. Ces molécules sont très hydrophobes, peu volatiles, très peu hydrosolubles avec des solubilités se situant entre 1 et 30 $mg.L^{-1}$ pour les HAP les plus légers et à moins de 1 $\mu g.L^{-1}$ pour les plus lourds (**Tableau 3**). A l'inverse leur solubilité est grande dans les phases organiques (solvants apolaires, lipides) avec des valeurs de coefficient de partage eau/octanol élevées (Log Kow, permettant d'appréhender le caractère hydrophobe d'une molécule) se situant entre 3 et 8 (Miller et al 1985). En outre, le caractère hydrophobe des HAP augmente avec le nombre de cycles aromatiques et leurs différentes propriétés physico-chimiques les rendent récalcitrants à la dégradation naturelle et donc persistants dans l'environnement.

1.3.1.b Toxicité et effets biologiques

Le caractère hydrophobe et lipophile des HAP entraîne leur bioconcentration dans les tissus lipidiques et donc au niveau de la chaîne alimentaire (Marsili et al 2001). Les HAP ne sont pas directement actifs, mais suite à leur pénétration dans l'organisme par inhalation, ingestion ou contact cutané, ces derniers se retrouvent au niveau de tissus cibles où ils subiront des réactions de détoxication aboutissant à la formation de métabolites génotoxiques (Kleinow et al 1998, Slater 1984). Au niveau hépatique, les monooxygénases microsomales de type cytochrome P450 transforment les HAP par oxydation en époxydes, dérivés phénoliques, quinones ou en radicaux très réactifs, capables de se lier de manière covalente au

niveau des acides nucléiques et des protéines pour former des adduits. Ces derniers se lient et perturbent ainsi les phénomènes de réplication et de transcription au niveau de l'ADN, pouvant entraîner un dérèglement du processus de division cellulaire et donc aboutir à la formation de tumeurs (Lin et al 2001, Szeliga and Dipple 1998).

1.3.1.c Origines des HAP et distribution dans l'environnement

Les hydrocarbures aromatiques polycycliques peuvent avoir une origine naturelle ou anthropique. Les HAP tirent leur origine de la matière organique et sont principalement formés par combustion incomplète de cette dernière, naturellement au cours des éruptions volcaniques ou encore des feux de forêt (Baumard et al 1999, Mandalakis et al 2005), ou bien de manière anthropique par combustion du carbone fossile tel que le charbon ou les produits pétroliers (Samanta et al 2002, Venkataraman et al 2002). Depuis le début de l'ère industrielle, les HAP émis dans l'environnement sont majoritairement d'origine anthropique (Wild and Jones 1995). Le transport, le chauffage résidentiel, le traitement des déchets par incinération émettent dans l'atmosphère des particules riches en HAP qui se dispersent avant de se déposer sur les sols par le biais des précipitations. En effet, du fait de leurs propriétés physico-chimiques, les HAP se répartissent dans l'atmosphère au niveau des phases gazeuses et particulaires (Gundel et al 1995, Ligocki and Pankow 1989). Cette répartition est régie majoritairement par leur pression de vapeur saturante où les HAP les plus légers (ayant une pression de vapeur saturante élevée) sont présents en phase gazeuse, alors que les HAP plus lourds (ayant une pression de vapeur saturante faible) sont associés à la phase particulaire (Van Jaarsveld et al 1997). De plus, l'exploitation, le raffinage et les rejets accidentels ou incontrôlés des produits pétroliers et du charbon sont à l'origine de la pollution des sols et des milieux marins par les HAP. Ainsi, les HAP peuvent aussi bien être issus de la combustion incomplète de produits pétroliers, du charbon ou encore du gaz naturel (Ravindra et al 2001), que de la combustion de biomasse ou encore d'ordures ménagères (Besombesa et al 2001). Enfin, les HAP peuvent aussi être synthétisés lors de processus physico-chimiques et naturels comme la diagénèse en même temps que les énergies fossiles (charbon, pétrole). Ces formations se produisent à des températures relativement basses (50-150°C) au cours de l'enfouissement de la matière organique dans les bassins sédimentaires (Soclo et al 2000). Les HAP sont donc présents naturellement dans les produits pétroliers (Cahnmann 1955).

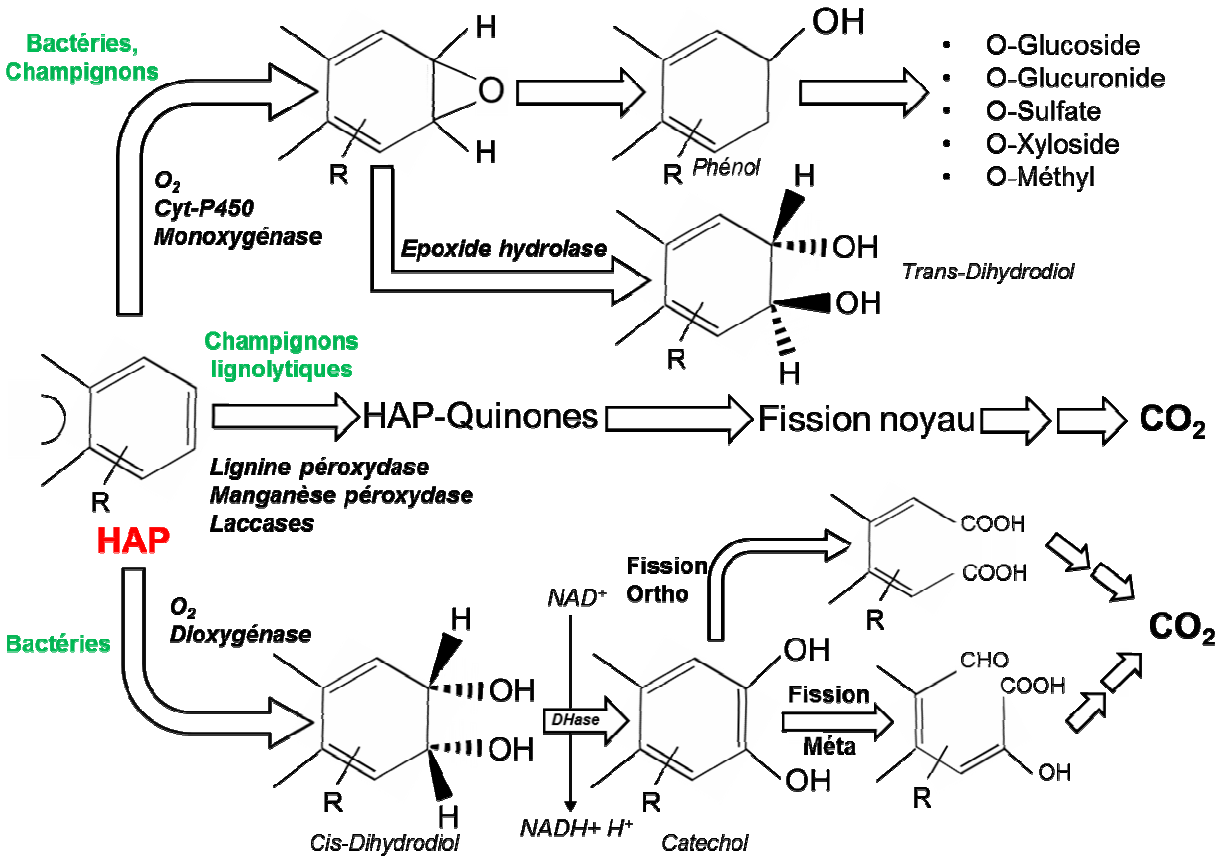

Figure 18. Voies de dégradation aérobies des HAP par les microorganismes. (Modifié et redessiné d'après Cerniglia, 1992)

1.3.2 Biodégradation microbienne aérobie des HAP

La dégradation microbienne aérobie reste actuellement la voie de dégradation la plus étudiée et la plus documentée en relation avec la capacité des microorganismes à utiliser les HAP comme source de carbone et d'énergie (Seo et al 2009). De nombreux organismes sont capables de dégrader les HAP comme les bactéries, les algues ou encore les champignons (Cerniglia 1992) (**Figure 18**). La minéralisation complète des HAP en CO_2 et H_2O est majoritairement observée chez les bactéries où elle fait intervenir notamment des dioxygénases, enzymes multimériques permettant l'incorporation de deux atomes d'oxygène au sein d'un cycle aromatique. Les dioxygénases (appelées également arène dioxygénases ou ring hydroxylating dioxygenase, RHD) sont des métallo-enzymes très répandues chez les bactéries (Cerniglia 1992). Elles catalysent (1) l'oxydation des HAP par hydroxylation simultanée de deux carbones adjacents au niveau d'un cycle aromatique, aboutissant à la formation de *cis*-dihydrodiols, et (2) elles permettent la fission du noyau aromatique avec les intradiol dioxygénases qui clivent le noyau aromatique en position *ortho* (entre les carbones porteurs des groupements hydroxyles – OH) et les extradiol dioxygénases en position *meta* (clivage entre les deux carbones adjacents au diol).

La dégradation bactérienne aérobie des HAP est composée de deux phases distinctes. Une première voie de dégradation, dite haute, est spécifique du HAP à dégrader permettant son attaque initiale pour aboutir à la formation de métabolites comme le catéchol ou ses précurseurs comme le protocatéchuate et le salycilate (Díaz 2004, Seo et al 2009). La deuxième voie de dégradation est dite basse et est commune à plusieurs composés autres que les HAP (benzène, toluène, ethylbenzène, xylène, BTEX), permettant de produire les substrats du métabolisme cellulaire central par la fission *ortho* ou *meta* des cycles aromatiques. Chez certaines bactéries, la voie haute de dégradation des HAP n'aboutit pas à la formation du catéchol, mais à la formation du gentisate dont le précurseur est l'acide salicylique, et permet également la production d'énergie via une voie basse annexe de dégradation dite du gentisate (Lee et al 2011, Zhou et al 2001). Les gènes de dégradation des HAP sont majoritairement organisés en opérons sur le chromosome bactérien ou plus fréquemment sur des éléments génétiques mobiles comme les plasmides ou encadrés de transposons. Cependant, les gènes de dégradation des HAP peuvent présenter des organisations génétiques différentes (Khan et al 2001, Kim et al 2006a, Saito et al 2000).

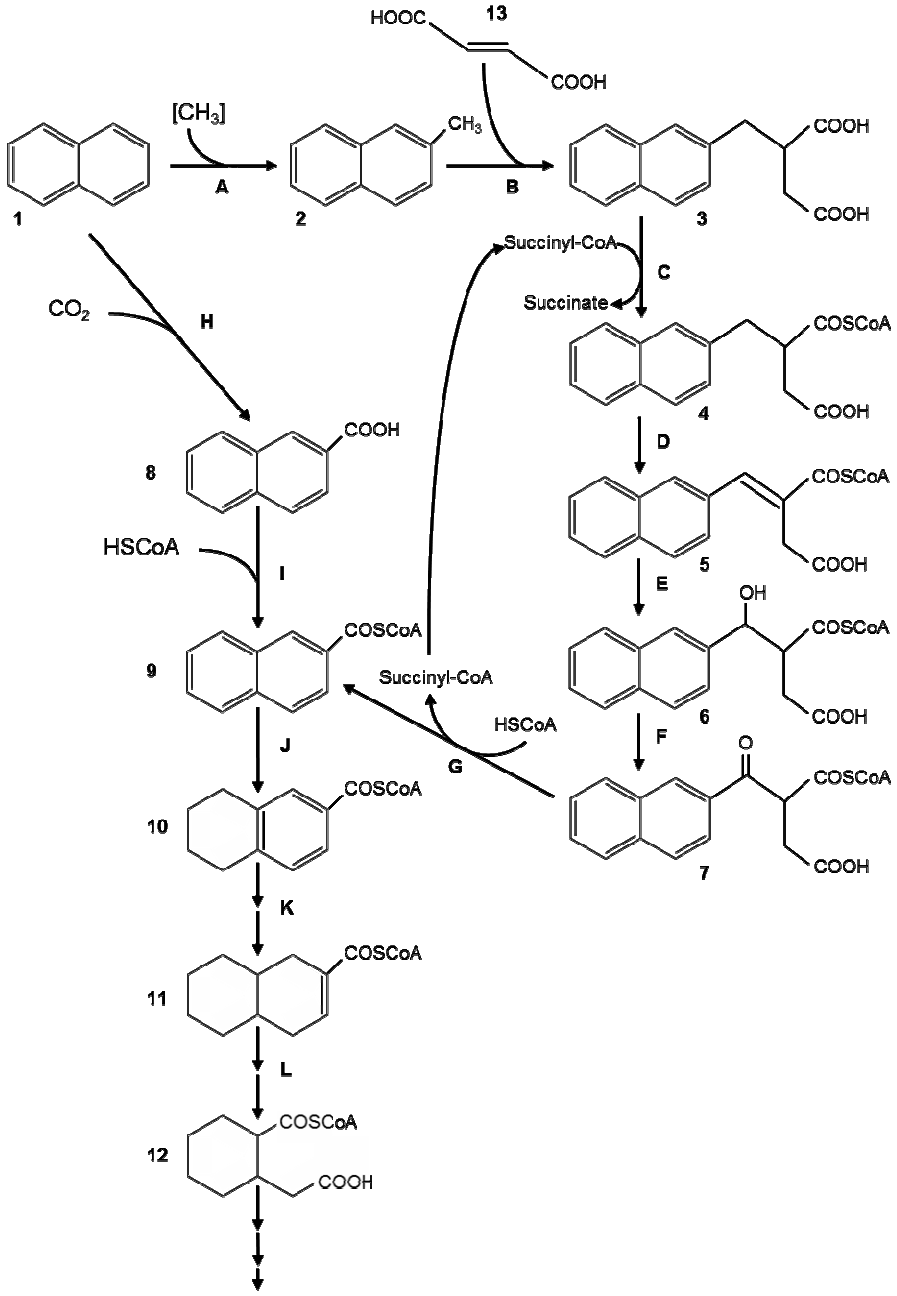

Figure 19. Voie de dégradation anaérobie du naphtalène. Les deux voies d'activation possibles pour la dégradation du naphtalène sont la carboxylation ou la méthylation. Intermédiaires: **(1)** naphtalène; **(2)** 2-méthylnaphthalene; **(3)** acide naphthyl-2-méthyl-succinique; **(4)** naphthyl-2-méthyl-succinyl-CoA; **(5)** naphthyl-2-méthylène-succinyl-CoA; **(6)** naphthyl-2-hydroxyméthyl-succinyl-CoA; **(7)** naphthyl-2-oxométhyl-succinyl-CoA; **(8)** acide naphthoïque; **(9)** 2-naphthoyl-CoA; **(10)** 5,6,7,8-tétrahydro-2-naphthoyl-CoA; **(11)** octahydro-2-naphthoyl-CoA; **(12)** cis-2-carboxycyclo-hexylacetyl-CoA; **(13)** fumarate. Enzymes : **(A)** naphtalène méthyl-transférase; **(B)** napthyl-2-méthylsuccinyl synthase; **(C)** napthyl-2-méthyl-succinyl-CoA transférase; **(D)** naphthyl-2-méthyl-succinyl-CoA déshydrogénase; **(E)** naphthyl-2-méthylene succinyl-CoA hydratase; **(F)** naphthyl-2-hydroxyméthyl-succinyl-CoA déshydrogénase; **(G)** napthyl-2-oxométhyl-succinyl-CoA thiolase; **(H)** naphthoate carboxylase; **(I)** naphthoyl-CoA ligase; **(J et K)** 2-napthtoyl-CoA réductase; **(L)** enoyl-CoA hydratase. (Modifié et redessiné d'après Meckenstock et Mouttaki, 2011).

Les voies de dégradation aérobie des HAP légers notamment celle du naphtalène (2 cycles aromatiques) (Davies and Evans 1964, Treccani et al 1954) et du phénanthrène (3 cycles aromatiques) (Doyle et al 2008, Haritash and Kaushik 2009, Peng et al 2008, Seo et al 2009) ont été très étudiées depuis une soixantaine d'années, et ces dernières servent de base pour la compréhension des mécanismes de dégradation par les bactéries pour d'autres types de HAP. Alors que les HAP à deux ou trois noyaux aromatiques sont facilement dégradés par les bactéries, les HAP de hauts poids moléculaires sont beaucoup plus récalcitrants à la biodégradation (Cerniglia 1992, Peng et al 2008). Certaines étapes sont encore mal caractérisées et certains mécanismes de même que les enzymes impliqués restent partiellement connus (Kim et al 2006b, Kweon et al 2007, Peng et al 2008).

D'autres voies de dégradation sont connues chez les champignons et les algues, mais ces voies sont incomplètes. Ainsi la dégradation est réalisée majoritairement par cométabolisme aboutissant à la production de métabolites qui seront pris en charge ultérieurement par d'autres communautés microbiennes (Doyle et al 2008, Peng et al 2008).

1.3.3 Biodégradation microbienne anaérobie

En raison de leurs propriétés physico-chimiques particulières, et notamment leur faible solubilité dans l'eau, les HAP se retrouvent préférentiellement dans les phases particulaires et donc dans différents environnements anoxiques comme les sédiments ou les sols (Karthikeyan and Bhandari 2001). Les bactéries sont également capables de dégrader les HAP en anaérobiose mais ceci de manière plus lente qu'en présence d'oxygène (Haritash and Kaushik 2009, Meckenstock and Mouttaki 2011). Actuellement, relativement peu d'études décrivent les voies métaboliques impliquées au niveau de cette biodégradation en absence d'oxygène, mais il a été montré que les bactéries utilisent majoritairement le nitrate, le sulfate, le fer ferrique, le chlorate, le manganèse ou encore le CO_2 comme accepteurs d'électrons (Gibson and S. Harwood 2002, Meckenstock et al 2004, Meckenstock and Mouttaki 2011).

1.3.3.a Voies de dégradation anaérobie

La dégradation du naphtalène est actuellement la plus documentée. Cette dégradation débute soit par une méthylation suivie d'une addition du fumarate, ou soit par une carboxylation (Meckenstock and Mouttaki 2011, Safinowski and Meckenstock 2006) (**Figure 19**). La voie initiée par une méthylation donnant le méthylnaphtalène est relativement bien caractérisée et elle est associée notamment à la réduction du sulfate en sulfure ou du nitrate en azote gazeux. L'attaque initiale de la molécule de méthylnaphtalène est une addition d'une

molécule de fumarate sur le groupement méthyle pour donner du naphtyl-2-méthyl-succinate. Une molécule de coenzyme A (CoA) est ensuite transférée du succinyl-CoA au naphtyl-2-méthyl-succinate et le carbone méthylé de ce dernier est ensuite beta-oxydé. Par la suite, une molécule d'H_2O est ajoutée, suivi d'une nouvelle étape d'oxydation et d'une réaction de clivage de la chaîne latérale. Ce clivage aboutit à la formation de deux intermédiaires, une molécule de succinyl-CoA (réutilisée au niveau de l'étape d'ajout du coenzyme A sur le naphtyl-2-méthyl-succinate) et une molécule de 2-naphtoyl-CoA (Meckenstock et al 2004, Meckenstock and Mouttaki 2011) (**Figure 19**). Le naphtalène empruntant la voie par carboxylation aboutit également à la production de 2-naphtoyl-CoA via l'ajout préalable d'une molécule de CO_2 produisant de l'acide 2-naphtoïque, puis d'une molécule de coenzyme A (**Figure 19**). Le 2-naphtoyl-CoA est ensuite réduit par étapes successives pour former du *cis*-2-carboxycyclo-hexylacétyl-CoA. Par la suite, l'ouverture du noyau aromatique aboutira à la formation d'acétyl-CoA et de CO_2 (Meckenstock and Mouttaki 2011). Concernant la dégradation du phénanthrène, cette dernière a été montrée associée à la réduction du sulfate où la molécule à trois cycles aromatiques est activée par carboxylation directe et est dégradée de manière similaire au naphtalène (Davidova et al 2007, Meckenstock and Mouttaki 2011, Zhang and LY. 1997).

1.3.3.b Implication des communautés méthanogènes

Les hydrocarbures aromatiques polycycliques peuvent contaminer les environnements anoxiques et notamment ceux propices à la méthanogénèse comme par exemple les sédiments, où les HAP peuvent potentiellement servir de source de carbone pour des communautés méthanogènes (Berdugo-Clavijo et al 2012). Même si le métabolisme syntrophique de dégradation des HAP couplé à la production de méthane est considéré comme un processus peu énergétique (Schink 1997), il apparaît thermodynamiquement faisable (Dolfing et al 2009) et a été analysé dans quelques études en laboratoire. Il a été montré que des enrichissements de communautés méthanogènes obtenus à partir de boues de station d'épuration étaient capables d'utiliser des HAP de plus de trois noyaux aromatiques (Christensen et al 2004, Trably et al 2003). De même, il a pu être observé une baisse de 50% de la concentration en naphtalène et en phénanthrène au niveau de sédiments portuaires incubés en l'absence d'accepteurs d'électrons (Chang et al 2006). De le même manière, il a pu être montré une dégradation de cinq différents types de HAP en utilisant des enrichissements obtenus à partir de sédiments de rivières (Yuan and Chang 2007). Par ailleurs l'addition d'acétate, de lactate et de pyruvate dans ces mêmes incubations a permis d'augmenter la

production de méthane et le taux de dégradation des HAP. Plus récemment, certaines études ont montré une production modérée de méthane au niveau de sédiments supplémentés avec du naphtalène (Siegert et al 2011), de sédiments supplémentés avec de l'anthracène (Zhang et al 2011) ou encore une consommation de certains HAP et alcanes par un *consortium* thermophile de méthanogènes au sein d'échantillons provenant de champs pétrolifères (Gieg et al 2010). Enfin, sur un même type d'échantillons issus de champs pétrolifères, il a pu être montré une consommation de méthyl-naphtalène corrélée à la production de méthane avec la mise en évidence des acides naphtoiques comme étant des métabolites clés dans la dégradation des HAP en conditions méthanogènes (Berdugo-Clavijo et al 2012). De plus, au niveau de cette étude, certaines communautés méthanogènes comme celles appartenant aux genres *Methanosaeta* et *Methanoculleus* ont été mises en évidence, suggérant leur importance au niveau de ce métabolisme particulier.

En outre, de récentes études biogéochimiques suggèrent que les HAP peuvent être dégradés en étant couplé à la méthanogénèse et ceci de manière *in situ*. En effet, plusieurs métabolites de dégradation anaérobie des HAP ont été détectés au niveau du pétrole brut échantillonné à différents endroits du globe et suggérant ainsi que les HAP servent de substrat dans ces environnements profonds et anoxiques (Aitken et al 2004). Une telle découverte couplée à des études isotopiques a montré qu'une proportion du CH_4 et du CO_2 présents dans les environnements pétrolifères est d'origine biogénique, et a suggéré que la biodégradation de composants du pétrole incluant les HAP était couplée à une production de méthane (Gray et al 2010). Plusieurs autres types d'environnements comme des aquifères contaminés par des essences (Essaid et al 2011, Gieg and Suflita 2002, Griebler et al 2004, Kleikemper et al 2005, Morasch et al 2011) ont montré une dégradation des HAP *in situ* en conditions méthanogènes. Il est intéressant de noter que certains de ces sites avaient été caractérisés au préalable comme étant capable de dégrader les HAP en conditions de réduction des sulfates et des nitrates, plutôt qu'en conditions méthanogènes. Enfin une autre étude a montré la minéralisation *in situ* du ^{14}C-naphtalène et du $^{14}CO_2$ au niveau d'aquifères contaminés par du créosote en conditions réductrices du fer et en conditions méthanogènes (Bianchin et al 2006).

Cependant, même si des études démontrent la biodégradation des HAP en conditions méthanogènes, les voies métaboliques ainsi que les microorganismes impliqués ne sont pas encore clairement identifiés (Berdugo-Clavijo et al 2012).

1.4 Conclusion

Les microorganismes jouent un rôle central au niveau des cycles biogéochimiques, notamment celui du carbone, en colonisant les différents réservoirs au travers de communautés microbiennes complexes. Ces mêmes communautés sont les acteurs du bon fonctionnement des écosystèmes, mais ils peuvent aussi amplifier les perturbations engendrées par l'activité humaine contribuant au changement global. Les informations disponibles sur les microorganismes et les enzymes impliqués dans les voies de production et de consommation du méthane ou encore les voies de dégradation des HAP en anaérobiose restent à l'heure actuelle très incomplètes. En effet, ces données sont accessibles soit à partir d'études de microorganismes cultivables alors que la grande majorité des microorganismes reste encore non cultivée, soit au travers de données moléculaires parcellaires après amplifications de biomarqueurs phylogénétiques et fonctionnels. C'est pourquoi il est indispensable d'étudier les écosystèmes dans leur globalité en mettant en place de nouvelles approches permettant d'explorer la diversité microbienne au sein des environnements complexes.

2. Etude de la diversité des communautés méthanogènes

L'exploration des environnements complexes demeure actuellement l'un des défis majeurs en écologie microbienne du fait de l'importante diversité des microorganismes qu'ils hébergent. Pour une bonne compréhension du fonctionnement des écosystèmes il est important (1) d'identifier les microorganismes (structure des communautés), (2) de caractériser leurs fonctions métaboliques, (3) de relier la structure à la fonction. Un grand nombre de méthodes culturales, moléculaires et biochimiques ont été dès lors appliquées pour répondre à ces questions et permettant notamment l'exploration des communautés méthanogènes.

2.1 Approches basées sur la culture

Les méthodes classiques de microbiologie basées sur la culture impliquent l'inoculation d'un échantillon environnemental sur des milieux de culture (solides ou liquides) dont la composition doit favoriser l'isolement des microorganismes d'intérêts (Hugenholtz 2002). Les paramètres de croissance tels que la température, le pH, le temps d'incubation, l'aération, la présence ou l'absence de lumière diffèrent selon les populations à caractériser. Les approches dépendantes de la culture sont mises en œuvre soit pour dénombrer les cellules viables (Sait et al 2002), soit pour sélectionner des microorganismes présentant un caractère particulier. Elles ont été couramment utilisées pour la caractérisation phénotypique et l'isolement de nouvelles souches d'archées méthanogènes à partir d'échantillons environnementaux (Borrel et al 2012, Brauer et al 2010, Chong et al 2002, Doerfert et al 2009, Imachi et al 2008, Kitamura et al 2011, Lomans et al 1999, Sakai et al 2007, Sakai et al 2008, Sakai et al 2011, Tian et al 2009, Uchiyama et al 2010, Yashiro et al 2009). Cependant, ces méthodes présentent certaines limites comme celle de favoriser la sélection des microorganismes à croissance rapide (Sait et al 2002). De plus, ces méthodes ne permettent pas de reproduire exactement les conditions de l'environnement à la base de l'isolement. Par exemple, les Acidobactéries représentent en moyenne 20% des communautés bactériennes du sol, mais ces organismes sont difficilement cultivables sur la base des connaissances actuelles (Schloss and Handelsman 2004). De nouvelles méthodes de culture utilisent des chambres de diffusion (Nichols et al 2010) ou encore l'encapsulation de cellules dans des microgouttelettes de gel combinée à de la cytométrie en flux, pour permettre une culture à grande échelle des microorganismes (Zengler 2002, Zengler et al 2005). De plus, des techniques de séparation physique, telles que la filtration, la centrifugation sur gradient de densité ou encore la

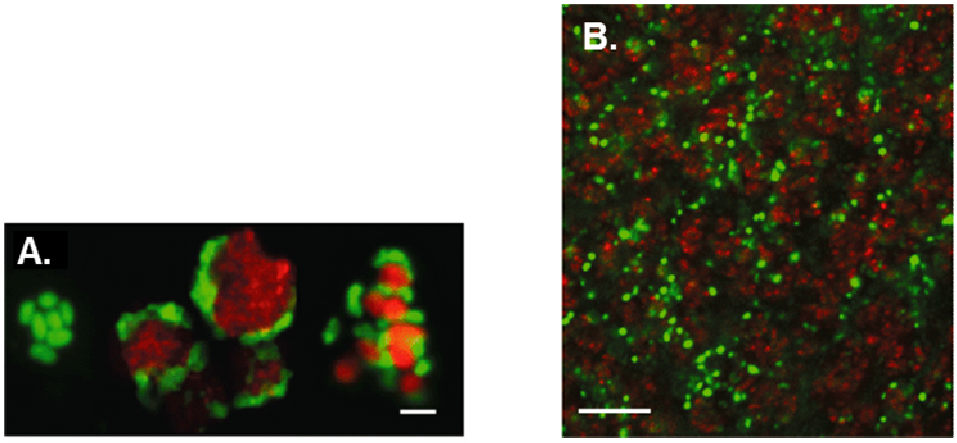

Figure 20. Caractérisation de l'association des bactéries sulfato-réductrices et des archées ANME-2 par la technique CARD-FISH au niveau de sédiments marins contenant des hydrocarbures. (A) utilisation d'une sonde SEEP2-658 (vert) ciblant les BSR et d'une sonde ANME2-538 ciblant le groupe des ANME-2. **(B)** Visualisation à l'aide des sondes SEEP2-658 et ANME2-538 d'un échantillon de sédiments où ce dernier apparaît dominé par les ANME-2 (rouge). (D'après Kleindienst et al., 2012).

Tableau 4. Liste des différents phyla bactériens connus et référencés dans les bases de données. Les phyla possédant un représentant cultivé sont indiqués en rouge (d'après la base de données SILVA, http://www.arb-silva.de/browser/, septembre 2012)

Phyla bactériens	Représentant cultivé	Phyla Bactériens	Représentant cultivé
Acidobacteria	oui	Fusobacteria	oui
Actinobacteria	oui	GAL08	non
Aquificae	oui	Gemmatimonadetes	oui
Armatimonadetes	oui	GOUTA4	non
Bacteroidetes	oui	HDB-SIOH1705	non
BD1-5	non	Hyd24-12	non
BHI80-139	non	JL-ETNP-Z39	non
Caldiserica	oui	Kazan-3B-28	non
Candidate division BRC1	non	LD1-PA38	non
Candidate division KB1	non	Lentisphaerae	oui
Candidate division OD1	non	MVP-21	non
Candidate division OP11	non	Nitrospirae	oui
Candidate division OP3	non	NPL-UPA2	non
Candidate division OP9	non	OC31	non
Candidate division SR1	non	Planctomycetes	oui
Candidate division TM7	non	Proteobacteria	oui
Candidate division WS3	non	RF3	non
Candidate division WS6	non	RsaHF231	non
Chlamydiae	oui	S2R-29	non
Chlorobi	oui	SM2F11	non
Chloroflexi	oui	Spirochaetes	oui
Chrysiogenetes	oui	Synergistetes	oui
CK-1C4-19	non	TA06	non
Cyanobacteria	oui	Tenericutes	oui
Deferribacteres	oui	Thermodesulfobacteria	oui
Deinococcus-Thermus	oui	Thermotogae	oui
Dictyoglomi	oui	TM6	non
Elusimicrobia	oui	Verrucomicrobia	oui
Fibrobacteres	oui	WCHB1-60	non
Firmicutes	oui		

cytométrie en flux (Hugenholtz 2002), ont été mises en place pour permettre un tri préalable des microorganismes. Ces nouvelles approches ont permis certes d'augmenter considérablement la possibilité d'isoler de nouveaux microorganismes, mais elles ne donnent toujours accès qu'à une infime partie de l'immense diversité microbienne.

De manière à obtenir des données complémentaires aux approches culturales certaines méthodes d'observation ont été mises au point afin de visualiser directement l'abondance, la répartition et les interactions des communautés d'intérêt, et ce de manière *in situ*. Ces approches sont basées sur l'utilisation de sondes fluorescentes ciblant spécifiquement des séquences d'acides nucléiques particulières **(Figure 20)**. Les plus utilisées actuellement sont le Fluorescent In Situ Hybridization (FISH) et le CARD-FISH (Amann et al 1990, Bottari et al 2006, DeLong et al 1989, Schönhuber et al 1997, Valm et al 2011, Valm et al 2012). Ces dernières, notamment le FISH, ont été utilisées pour la détection des méthanogènes au sein d'échantillons environnementaux grâce à l'utilisation de sondes spécifiques de ces communautés (Narihiro and Sekiguchi 2011). Ces techniques génèrent des données quantitatives intéressantes mais renseignent difficilement sur les capacités métaboliques des microorganismes. Ainsi d'autres stratégies, couplant le FISH à des techniques isotopiques (MAR-FISH, FISH-NanoSIMS) (Lee et al 1999, Li et al 2008), ou encore la SIP (Stable Isotope Probing) incorporant des isotopes stables issus de substrats marqués au niveau des molécules d'ADN (Lueders et al 2004), ont été développées de manière à pouvoir relier l'identification des microorganismes à leurs fonctions métaboliques. Toutefois, ces méthodes restent limitées quant à leur application du fait de leur difficulté de mise en œuvre mais surtout par leur faible débit.

Des études moléculaires basées sur la caractérisation des séquences d'ADNr 16S ont révélé des divisions candidates bactériennes telles que BRC, OP10, OP11, SC3, TM7, WS2 et WS3 qui pour certaines n'ont toujours pas de représentants cultivés à ce jour. Les études moléculaires ont révélé la présence de 59 phyla bactériens (Ludwig et al 2004, Pruesse et al 2007) dont la moitié d'entre eux n'ayant pas encore de représentants cultivés (**Tableau 4**) (Harris et al 2004, Hugenholtz et al 1998, Rappé and Giovannoni 2003). Le même constat est observé chez les méthanogènes où des séquences d'ADNr 16S ont été assignées à des groupes candidats, comme par exemple le groupe WSA2 (ou ArcI) fréquemment retrouvé dans des systèmes de traitement des eaux usées (Chouari et al 2005). Ce groupe est considéré comme étant un taxa archéen ne possédant actuellement aucun représentant cultivé (Hugenholtz

2002). Un autre exemple est le Rice Cluster II assimilé aux méthanogènes puisque les séquences ADNr 16S affiliées à ce groupe sont retrouvées dans des cultures enrichies contenant de l'éthanol comme donneur d'électrons, et ce groupe forme un clade phylogénétique distinct au sein des *Methanomicrobiales* et *Methanosarcinales* (Lehmann-Richter et al 1999).

Ces observations mettent en avant le potentiel des techniques moléculaires qui permettent de contourner les limites des approches culturales, et ainsi permettre de caractériser finement les communautés microbiennes.

2.2 Méthodes moléculaires

Le développement des techniques moléculaires ces vingt-cinq dernières années permet aujourd'hui d'aborder les problématiques d'écologie microbienne simplement au travers de l'analyse des molécules d'acides nucléiques (ADN et/ou ARN) en contournant les limites des méthodes culturales ou d'observation au microscope des microorganismes (Amann et al 1995, Pace 1997). Ces approches indépendantes de la culture reposent sur l'utilisation de génomes entiers ou de biomarqueurs capables de renseigner sur l'identité (biomarqueurs phylogénétiques) ou le rôle fonctionnel (biomarqueurs fonctionnels) d'un grand nombre de microorganismes.

2.2.1 Utilisation des biomarqueurs

Le biomarqueur phylogénétique le plus utilisé en écologie microbienne est le gène codant pour la petite sous-unité de l'ARN ribosomique (ARNr 16S chez les procaryotes et ARNr 18S chez les eucaryotes) (Woese et al 1990). Le biomarqueur phylogénétique ADNr 16S (Woese 1987) a été largement utilisé en écologie microbienne pour la description des communautés bactériennes et archées de l'environnement puisque (1) il est ubiquiste c'est-à-dire retrouvé chez tous les procaryotes où il joue un rôle clé dans la traduction de l'ARNm en protéine, (2) il possède une structure en mosaïque incluant des régions conservées (permettant son isolement) mais aussi variables et hypervariables (à la base des comparaisons phylogénétiques), (3) il ne subit pas ou peu de transfert horizontal et de recombinaisons (Hugenholtz 2002). De plus sa taille adaptée (~1500 pb) ainsi que le nombre croissant de séquences codantes pour l'ARNr 16S présentes dans les bases de données, font de lui un marqueur de choix. Certaines signatures de ce biomarqueur peuvent en outre être caractéristiques de groupes fonctionnels comme par exemple les microorganismes

déhalorespirants ou encore les sulfato-réducteurs. Inversement, des biomarqueurs fonctionnels peuvent renseigner sur l'identité des microorganismes comme par exemple le gène *mcrA* codant pour la sous-unité α de la méthyl coenzyme M réductase chez les archées ou encore le gène *pmo*A codant pour la méthane monooxygénase et retrouvé uniquement chez les bactéries méthanotrophes. Toutefois, les biomarqueurs fonctionnels sont généralement des gènes codant pour des enzymes impliquées dans des métabolismes d'intérêt. D'autres biomarqueurs tels que les gènes codant pour la sous-unité β de l'ARN polymérase (*rpoB*), la sous-unité β de l'ADN gyrase (*gyrB*), la recombinase A (*recA*) ou encore la « heat shock protein 60 » (Hsp60), ont été utilisés en écologie microbienne pour l'étude des communautés microbiennes ou encore pour différencier certaines espèces bactériennes (Ghebremedhin et al 2008).

Le biomarqueur ARNr 16S a été fréquemment utilisé pour la détection des communautés méthanogènes au sein d'environnements complexes. Cependant, son utilisation est limitée au niveau des méthanogènes, puisque ces dernières se retrouvent au sein d'un groupe d'organismes paraphylétique englobant des groupes d'archées non méthanogènes comme les Archaeoglobales, les Thermoplasmatales et les Halobacteriales (Bapteste et al 2005). Pour s'affranchir de cette contrainte, il est couramment utilisé des biomarqueurs fonctionnels spécifiques et conservés chez les méthanogènes comme le gène *mcrA* codant pour la méthyl-coenzyme M réductase (MCR) (Friedrich 2005) ou encore celui codant pour son isoenzyme (*mrtA*). En outre, les phylogénies basées sur les séquences nucléiques des gènes *mcrA*/*mrtA* ou sur les séquences protéiques déduites de ces gènes, sont congruentes avec celles déterminées par utilisation des séquences ADNr 16S (Friedrich 2005, Luton et al 2002, Springer et al 1995).

2.2.2 Analyse partielle des communautés basée sur l'amplification PCR

Ces stratégies utilisent la technique de réaction de polymérisation en chaîne (PCR) (Saiki et al 1985) pour amplifier une région d'ADN cible grâce à un couple d'amorce. Cette méthode a révolutionné l'étude des communautés microbiennes présentes dans les environnements complexes en étant capable de cibler spécifiquement n'importe quelle population ou groupe de microorganismes pour lesquels des informations de séquences sont disponibles.

2.2.2.a Détermination et utilisation des couples d'amorces

De nombreux couples d'amorces spécifiques du gène ADNr 16S ont été déterminés pour étudier de manière globale les communautés méthanogènes au sein d'environnements complexes (Marchesi et al 2001, Wright and Pimm 2003). De plus, ceux-ci ont été déterminés à des niveaux taxonomiques inférieurs (classe, ordre, famille voire même espèce) permettant la détection spécifique des communautés méthanogènes dans des boues de traitement des eaux usées (Ariesyady et al 2007, Franke-Whittle et al 2009a, Hori et al 2006, Narihiro et al 2009a, Narihiro et al 2009b, Rocheleau et al 1999, Zheng and Raskin 2000), au niveau du rumen (Skillman et al 2004, Yanagita et al 2000), dans des sédiments marins profonds (Boetius et al 2000, Nercessian et al 2004), des sédiments (Zepp et al 1999), du microbiote intestinal humain (Ratner et al 2009) ou encore au niveau de zones humides (Bräuer et al 2006, Zhang et al 2008a, Zhang et al 2008b).

Concernant l'utilisation du biomarqueur fonctionnel *mcrA*/*mrtA*, des couples d'amorces généralistes ont été déterminés (Luton et al 2002, Mihajlovski et al 2008). Néanmoins, d'autres amorces ont été définies pour identifier certaines familles comme les *Methanosarcinaceae* (Springer et al 1995), mais également pour étudier les populations spécifiquement retrouvées au sein d'un environnement donné comme les zones humides (Hales et al 1996).

2.2.2.b Les différentes méthodes d'analyse

L'utilisation de la PCR a permis de disposer d'une gamme de techniques permettant d'obtenir des informations sur la structure et la fonction des communautés microbiennes. Des méthodes de clonage/séquençage permettent d'obtenir des données de séquences phylogénétiques qui seront ultérieurement comparées à des bases de données généralistes comme Genbank ou plus spécifiques comme ARB (Ludwig et al 2004), Ribosomal Database Project (RDP) (Cole et al 2009) ou encore Greengenes (DeSantis et al 2006). Ainsi, les séquences obtenues sont affiliées à différents rangs taxonomiques (allant du phylum jusqu'à l'espèce) en identifiant différents seuils de similarité nucléique (DeSantis et al 2007). Bien que les banques de clones construites à partir de séquences d'ADNr 16S permettent d'explorer initialement la diversité et d'identifier de nouveaux taxa bactériens et archéens, des études ont montré que des échantillons environnementaux, tels que les sols, requièrent plus de 40 000 clones pour décrire 50% de la diversité (Dunbar et al 2002). Cependant les banques de clones d'ADNr 16S construites pour les études environnementales contiennent généralement un peu moins de 1000 séquences, et proposent donc une vision très réduite de la diversité

microbienne présente dans un échantillon. Malgré leurs limites (incluant également un coût en temps et en réactifs), les banques de clones sont toujours considérées comme un outil de base en écologie microbienne pour explorer de manière préliminaire la diversité microbienne (DeSantis et al 2007). De telles méthodes appliquées aux communautés méthanogènes et ciblant l'ADNr 16S ont pu montrer une dominance de certains phylotypes dans différents environnements complexes. Par exemple, dans des sédiments profonds de dépôts d'hydrates de gaz, les espèces des genres *Methanosarcina* et *Methanobrevibacter* sont surreprésentées (Marchesi et al 2001). L'application de ces méthodes a aussi permis la mise en évidence des phylotypes particuliers comme les *Methanomicrobia* dans des sédiments lacustres (Banning et al 2005). Des résultats similaires ont été obtenus en ciblant les gènes *mcrA*/*mrtA*, pour mettre en évidence la diversité et la dominance de communautés méthanogènes dans différents environnements, comme la colonne d'eau et les sédiments de lac d'eau douce (Biderre-Petit et al 2011), des sources hydrothermales (Dhillon et al 2005) ou des volcans de boues (Alain et al 2006).

D'autres méthodes basées sur l'amplification PCR, comme les techniques d'empreintes génétiques, donnent un profil des communautés microbiennes basé sur l'analyse directe des amplicons obtenus à partir d'ADN environnemental (Ramette 2009). Au cours de ces 25 dernières années, un large éventail de techniques a été développé pour la description des communautés microbiennes en produisant des empreintes moléculaires basées sur des polymorphismes de séquences ou de longueurs des gènes biomarqueurs (Kirk et al 2004). Parmi ces différentes techniques, il est possible de citer la Denaturing Gradient gel Electrophoresis / Temperature Gradient gel Electrophoresis (DGGE / TGGE) (Gelsomino et al 1999), la SSCP (Single Strand Conformation Polymorphism) (Lee et al 1996), la T-RFLP (Terminal-Restriction Length Polymorphism) (Liu et al 1997), la ARDRA (Amplified Ribosomal DNA Restriction Analysis) (Liu et al 1997) ou encore la A-RISA (Automated-Ribosomal Intergenic Spacer Analysis) (Fisher and Triplett 1999). Ces méthodes ont également été appliquées aux communautés méthanogènes pour donner des empreintes ADNr 16S et *mcrA* de différents environnements (Casamayor et al 2001, Casamayor et al 2002, Kemnitz et al 2004, Ramakrishnan et al 2001, Wright and Pimm 2003, Yu et al 2007). D'une manière générale, ces techniques sont rapides et permettent une analyse comparative simultanée de plusieurs échantillons. Elles ont été mises au point pour observer des changements ou des différences entre les communautés microbiennes, mais elles ne permettent pas une identification taxonomique directe des communautés.

La PCR quantitative (qPCR) fournit quant à elle une méthode sensible et permettant de quantifier des communautés microbiennes d'intérêt dans des environnements complexes (Zhang and Fang 2006). Ainsi la qPCR ciblant l'ADNr 16S et le gène *mcrA* sont également très utilisées pour quantifier les communautés méthanogènes de l'environnement. Différents couples d'amorces ont été déterminés dont les amorces généralistes Met630F/Met803R notamment utilisées pour la quantification des méthanogènes du rumen des vaches laitières (Hook et al 2008). De même, d'autres amorces ont été déterminées et appliquées à la technique de détection Taqman, permettant la quantification des différents ordres et de certaines familles (Yu et al 2005). Ces amorces ont ultérieurement servi pour la quantification des méthanogènes acétoclastes dans des boues de traitement des eaux usées et ont permis de mettre en évidence que ces mêmes populations étaient affectées par la concentration en acétate dans les eaux usées (Yu et al 2006). Des couples ciblant le gène *mcrA* sont également disponibles et permettent la quantification des méthanogènes et des ANME dans des boues de digestat et dans les sédiments (Inagaki et al 2004, Nunoura et al 2006) D'autres couples sont disponibles pour la quantification des méthanogènes en ciblant le gène *mcrA* dans le rumen (Denman et al 2007) ou encore dans la plaque dentaire humaine (Vianna et al 2008).

2.2.2.c Limites de l'amplification PCR pour l'analyse des communautés

Les méthodes basées sur la PCR couplées ou non à des techniques d'analyse des amplicons comme le clonage/séquençage ou encore les techniques d'empreintes génétiques, apportent certes des informations pertinentes sur la structure et/ou la fonction des communautés microbiennes, mais elles restent toutefois encore incomplètes et biaisées. En effet, toutes ces méthodes sont soumises à différentes limites techniques inhérentes à l'extraction des acides nucléiques et à l'amplification PCR. Ces dernières apparaissent à l'heure actuelle clairement identifiées pour les études en écologie microbienne et peuvent être classées en deux catégories. Elles peuvent résulter d'artéfacts de séquences liés à des erreurs de PCR, ou fausser la distribution des amplicons liée à une efficacité d'amplification PCR non homogène (Acinas et al 2005, Polz and Cavanaugh 1998, von Wintzingerode et al 1997), pouvant conduire à une surreprésentation de certains fragments et altérer la vision des communautés microbiennes présentes dans un environnement.

Les artéfacts de séquences peuvent être dus soit (1) à la formation de chimères se produisant au niveau d'un cycle de PCR lors d'une extension incomplète des amorces. Dans ce cas, le produit d'extension tronqué peut servir d'amorce pour le cycle suivant ; (2) à la

formation d'hétéroduplexes (séquences amplicons hétérologues) se formant préférentiellement durant les derniers cycles de PCR, au niveau de la phase de plateau de la cinétique d'amplification (Thompson et al 2002) ; (3) aux erreurs de *Taq* polymérase où une faible fidélité de cette dernière génère des erreurs d'incorporation durant la synthèse du brin d'ADN aboutissant à des substitutions de base (von Wintzingerode et al 1997). Ces séquences artéfactuelles apparaissent problématiques en écologie microbienne, car elles aboutissent à une surestimation de la diversité présente dans un environnement et à une identification erronée de nouveaux variants génétiques. De plus, la présence de bases mal incorporées au cours de l'élongation du brin d'ADN, induite par une faible fidélité de la *Taq* polymérase, peut être problématique notamment lorsque ces bases sont situées au niveau de sites moléculaires d'intérêt comme par exemple ceux nécessaires à la détermination de sondes ou d'amorces.

L'amplification PCR non homogène peut être due soit (1) à l'inhibition de l'amplification par la présence de molécules co-extraites avec les acides nucléiques comme par exemple les acides humiques, (2) à une différence d'amplification principalement due à la composition en bases des gènes cibles en lien avec le pourcentage GC des séquences ou encore (3) à la spécificité des amorces utilisées pouvant conduire à une surreprésentation de certains fragments (Polz and Cavanaugh 1998, Suzuki and Giovannoni 1996, von Wintzingerode et al 1997).

2.3 Méthodes d'analyse globale des communautés et émergence des outils haut-débit

L'analyse des séquences codantes pour des biomarqueurs phylogénétiques et fonctionnels est couramment utilisée en écologie microbienne pour l'exploration des environnements complexes. Cependant, même en utilisant des molécules très conservées pour permettre une bonne base d'affiliation taxonomique, certains biomarqueurs, comme le gène ADNr 16S, ne fournissent pas une résolution suffisante pour permettre dans tous les cas une discrimination au niveau de l'espèce ou de la souche (Konstantinidis et al 2006). Les techniques moléculaires basées sur l'exploitation de l'ensemble des séquences des génomes présents au sein d'un environnement offrent donc une vision plus exhaustive de la diversité génétique (Handelsman et al 1998).

2.3.1 La métagénomique

La métagénomique, également connu sous le terme de génomique environnementale ou génomique des communautés, se définit comme l'étude globale de l'ensemble des génomes des communautés microbiennes multi-espèces extraits directement à partir d'un échantillon environnemental et ne nécessitant pas au préalable une connaissance ou une mise en culture des communautés microbiennes (Handelsman et al 1998, Riesenfeld et al 2004). D'une manière générale, les techniques utilisant la métagénomique sont basées sur le principe suivant : l'ensemble des données génomiques des communautés microbiennes de l'environnement peut être criblé et/ou séquencé de la même manière que le génome entier issu par exemple d'une culture bactérienne pure. Des études métagénomiques ont été conduites au niveau de différents environnements tels que les sols, les lacs, les océans ou encore les drainages miniers acides pour permettre d'avoir accès à la diversité phylogénétique et fonctionnelle d'organismes non cultivés (Delmont et al 2010, Handelsman 2004, Tyson et al 2004). Ainsi, la métagénomique est primordiale pour la compréhension, au sein d'un environnement, des rôles des microorganismes non cultivés et de leurs interactions. Les banques environnementales construites à partir de métagénomes se sont avérées être très utiles pour la découverte de nouvelles enzymes microbiennes et d'antibiotiques d'intérêt, avec des applications potentielles au niveau des biotechnologies, de la médecine et de l'industrie. Les banques métagénomiques peuvent être criblées de manière à identifier les séquences impliquées dans des phénotypes d'intérêt. Ainsi, une grande variété d'activités biochimiques ont été mises en évidence comme de nouveaux antibiotiques (turbomycine, terragine), des enzymes microbiennes (cellulases, lipases, amylases) ou encore des protéines identifiés à partir des banques métagénomiques issues d'échantillons de sols (Rondon et al 2000). Toutefois cette stratégie nécessite une expression optimale des gènes dans un système hétérologue comme des bactéries ou des levures, ce qui n'est pas le cas pour l'ensemble des gènes d'intérêts. Il est donc important de disposer de vecteurs appelés vecteurs navettes capables de se propager à la fois dans différents hôtes bactériens et levures. En effet, au niveau d'une banque métagénomique, la fréquence associée à l'identification de gènes actifs exprimant un phénotype d'intérêt est relativement basse lorsqu'un seul type de système hétérologue est utilisé. A titre d'exemple, une étude au niveau d'une banque métagénomique, créée à partir d'un échantillon de sol et utilisant un système hétérologue *E. coli* a pu montrer seulement un clone sur 730 000 possédant une activité lipolytique d'intérêt (Henne et al 2000).

Tableau 5. Comparaison des différentes plateformes de séquençage de première et deuxième génération

	Séquenceur (Société)	Méthode d'amplification	Méthode de séquençage	Longueur des lectures	Débit (Mb par run)	Temps de séquençage	Coût (par Mb)	Disponibilité commerciale
1ère génération	**3730xl** (Applied Biosystems by Life Technologies)	PCR avec utilisation de didesoxyribonucléotides (ddNTPs)	Séquençage par synthèse (Sanger)	600-1000	0,06	2h	1500$	oui
2ème génération	**454 GS Jr. Titanium** (Roche/454)	PCR en émulsion (emPCR)	Séquençage par synthèse (Pyroséquençage)	400	50	10h	22$	oui
	454 FLX Titanium (Roche/454)	PCR en émulsion (emPCR)	Séquençage par synthèse (Pyroséquençage)	400	500	10h	17$	oui
	454 FLX+ (Roche/454)	PCR en émulsion (emPCR)	Séquençage par synthèse (Pyroséquençage)	700	900	18-20h	7$	oui
	Illumina MiSeq (Illumina/Solexa)	PCR en ponts (bridge PCR)	Séquençage par synthèse (terminaison réversible: CRT)	150	1020	26h	0,74$	oui
	Illumina HiSeq 1000 (Illumina/Solexa)	PCR en ponts (bridge PCR)	Séquençage par synthèse (terminaison réversible: CRT)	100	100 000	8 jours	0,10$	oui
	Illumina HiSeq 2000 (Illumina/Solexa)	PCR en ponts (bridge PCR)	Séquençage par synthèse (terminaison réversible: CRT)	100	200 000	8 jours	0,10$	oui
	SOLiD – 4 (Applied Biosystems by Life Technologies)	PCR en émulsion (emPCR)	Séquençage par ligation	50	71 400	12 jours	<0,11$	oui
	SOLID – 5500xl (Applied Biosystems by Life Technologies)	PCR en émulsion (emPCR)	Séquençage par ligation	75	155 100	8 jours	<0,07$	oui
	Helicos (Hélicos)	Aucune	Séquençage par synthèse (terminaison réversible: CRT)	35	28 000	ND	ND	oui

CRT: Cyclic Reversible Termination

Le criblage du métagénome peut aussi être réalisé *via* l'utilisation de sondes moléculaires ciblant des gènes d'intérêt et chaque fragment d'ADN caractérisé sera alors séquencé. Finalement, l'exploitation des banques métagénomiques peut également se faire par séquençage massif de l'ensemble des fragments générés. Ce type d'approche permet notamment de mettre en lumière d'importantes caractéristiques et organisations au niveau génomique, ainsi que des caractères acquis par des transferts horizontaux de gènes (Handelsman 2004).

Malgré la nouvelle vision du monde microbien apportée par l'essor de la métagénomique, qui permet d'avoir accès à l'ensemble des microorganismes présents dans un environnement, la création des banques environnementales demeure toutefois techniquement difficile à mettre en œuvre (Rajendhran and Gunasekaran 2008). Actuellement, l'émergence de nouvelles technologies facilitent l'étude des échantillons par l'approche de séquençage systématique, en proposant de séquencer directement les acides nucléiques extraits et donc de s'affranchir des étapes de clonage (Shendure and Ji 2008). Par exemple, le séquençage direct a pu mettre en évidence la diversité des communautés méthanogènes impliquées dans la production de biogaz au sein d'un bioréacteur au niveau du milieu de fermentation (Schlüter et al 2008, Wirth et al 2012) ou de biofilms (Rademacher et al 2012)

2.3.2 La révolution des techniques de séquençage

Grâce au développement de nouvelles technologies de séquençage qui depuis ces dix dernières années ont connu une révolution sans précédent, il est désormais possible d'avoir une vision globale et plus intégrative de l'ensemble des évènements se déroulant dans un environnement. Cette révolution concerne à la fois les technologies, l'appareillage, l'informatique ou encore les outils de traitement et de stockage des données.

2.3.2.a De la première à la deuxième génération de séquençage

Au début des années 1990, le séquençage était représenté par une première génération basée sur la technique de Sanger (Hunkapiller et al 1991, Sanger et al 1977, Swerdlow et al 1990). Même si ce type de séquençage est de moins en moins utilisé du fait de son coût élevé et de son faible débit (**Tableau 5**), il a permis la réalisation de projets d'envergure comme le séquençage du métagénome de la mer des Sargasses (Venter 2004) ou encore celui de la surface des océans (intitulé « the Sorcerer II Global Ocean Sampling Expedition ») (Rusch et al 2007). Mais de puis ces dix dernières années le séquençage par la méthode de Sanger a laissé place aux techniques de séquençage dites de deuxième génération ou « Next Generation

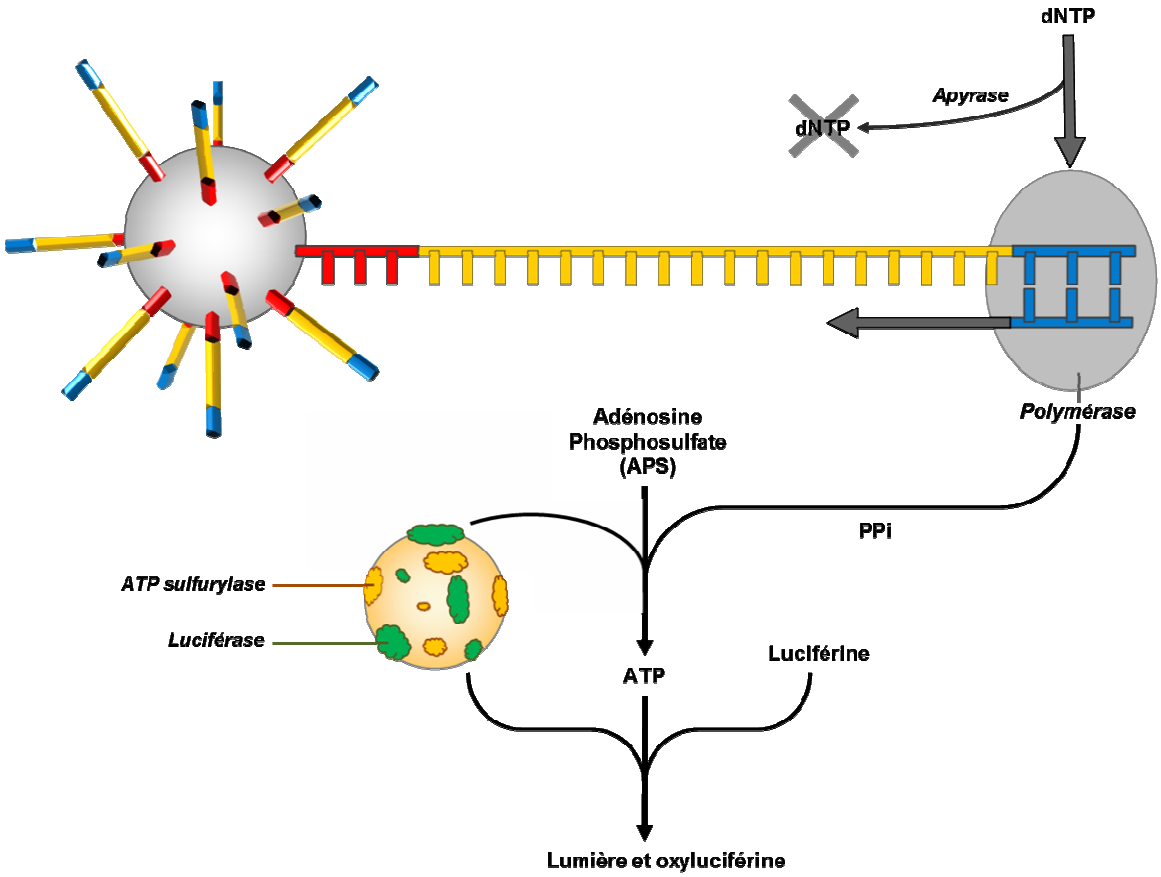

Figure 21. Principe de la méthode du pyroséquençage. PPi = phosphate inorganique. (Modifié et redessiné d'après Metzker, 2010)

Sequencing » (NGS), qui permettent de s'affranchir des étapes de clonage des fragments d'ADN, de réduire fortement les coûts et le temps d'acquisition des données, et donc d'augmenter considérablement les quantités de données de séquences produites (Ansorge 2009, Metzker 2010, Shendure and Ji 2008). Ces nouvelles approches dites à haut-débit sont indispensables pour permettre une exploration fine et présenter une vision non biaisée de la composition phylogénétique et de la diversité fonctionnelle des communautés microbiennes au sein des environnements complexes. Actuellement deux principales technologies de séquençage de deuxième génération connaissent un essor considérable en terme de développement technologique et d'application pour les études en écologie microbienne : le pyroséquençage 454 (454 Life Sciences / Roche Applied Science) et le séquençage Illumina (Solexa).

i. Le pyrosequençage 454

La technologie 454 est basée sur le principe du pyroséquençage développé depuis le milieu des années 1980 (Hyman 1988, Nyrén and Lundin 1985) puis amélioré au milieu des années 1990 avec la possibilité de multiplexage (Ronaghi et al 1996). Le pyroséquençage est un séquençage par synthèse. Les quatre désoxyribonucléotides sont ajoutés un par un de manière itérative et un capteur à transfert de charge (dispositif CDD ou Charge Coupled Device) permet de détecter la lumière produite transformant ainsi le signal lumineux en impulsion électrique. Au niveau moléculaire, l'addition au cours de la polymérisation d'un désoxyribonucléotide par l'ADN polymérase aboutit au relargage d'un pyrophosphate inorganique PPi. Ce même pyrophosphate inorganique couplé à l'adénosine phosphosulfate est pris en charge par l'ATP sulfurylase pour produire de l'ATP. Enfin, l'ATP néoformé couplé à la D-luciférine aboutit à la production d'oxyluciférine et de lumière par la luciférase. Une apyrase quant à elle est chargée de dégrader les désoxyribonucléotides non incorporés et l'ATP résiduel entre deux cycles d'incorporation de bases au niveau du brin néosynthétisé (Ahmadian et al 2006) (**Figure 21**).

La technologie de séquençage de 454 Life Sciences / Roche Applied Science développée en 2005 (Margulies et al 2005) est issue de la combinaison du pyroséquençage avec l'utilisation de plaques picotitrées (PicoTiterPlate, PTP), de la PCR en émulsion (emPCR) ainsi que des technologies informatiques pour l'acquisition et le traitement des images. Actuellement, des plateformes comme le GS FLX Titanium permettent de réduire fortement les coûts et le temps d'acquisition des données (**Tableau 5**) (Glenn 2011).

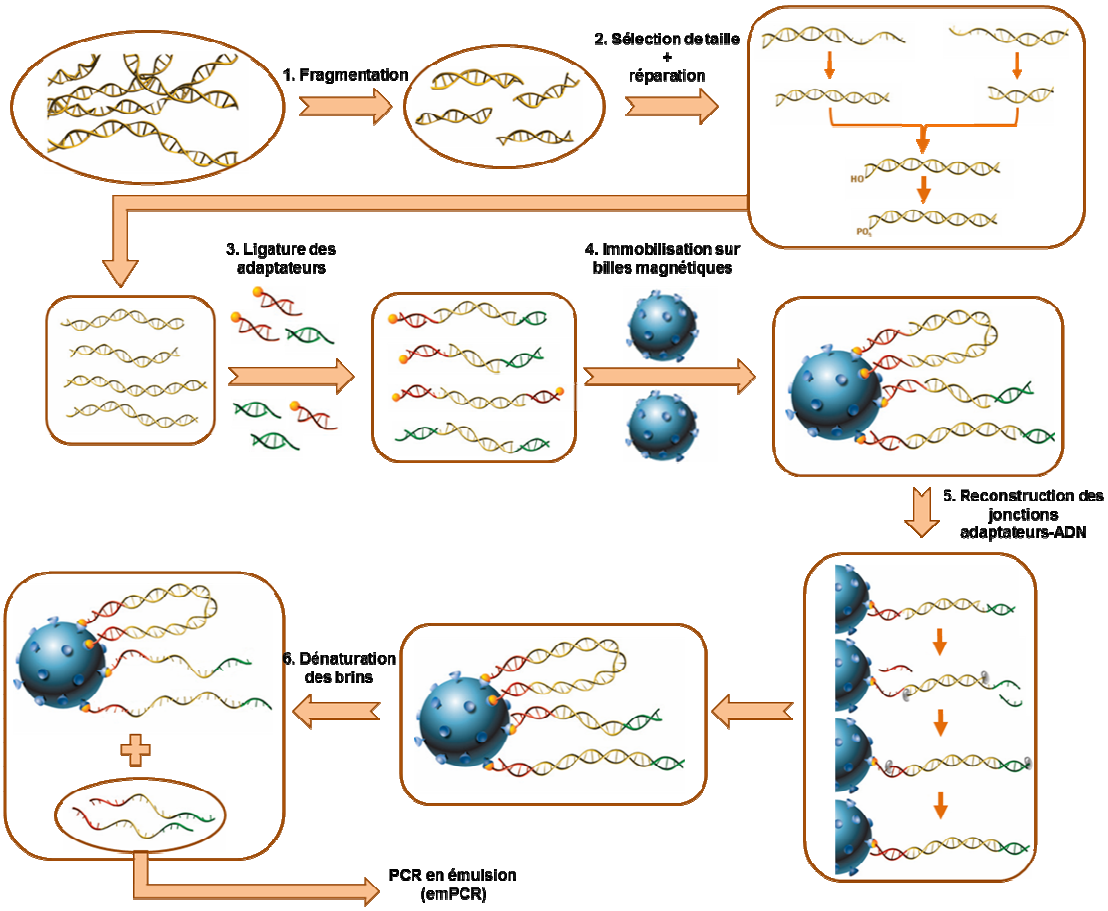

Figure 22. Préparation d'une banque d'ADN génomique simple brin pour le séquençage d'un échantillon de manière *de novo*. (1) fragmentation de l'ADN par nébulisation correspondant à une fragmentation mécanique aléatoire générant des fragments entre 400 et 1000 pb ; **(2)** La deuxième étape est la sélection des fragments dont la taille est compatible avec la technologie de séquençage GS FLX Titanium, comprise entre 500 et 800 pb puis la réparation des extrémités des fragments d'ADN de façon à générer des extrémités franches par action d'une polynucléotide kinase T4 (rephosphorylation des extrémités 5'- OH) et d'une ADN polymérase T4 (possédant une activité 5' 3' polymérase et une activité 3' 5' exonucléase) ; **(3)** Ligature des adaptateurs aux deux extrémités des fragments d'ADN réparés. Ces adaptateurs (nommés adaptateurs A et B) constituent les régions d'ancrage pour les étapes d'amplification et de séquençage. Ils contiennent une séquence unique non-palindromique de quatre bases appelée « clé » utilisée pour la calibration du séquenceur et la reconnaissance des fragments à séquencer. L'adaptateur B contient un marquage à la biotine au niveau de son extrémité 5'-P permettant l'immobilisation de la banque d'ADN sur des billes magnétiques nécessaire à la récupération finale d'une banque d'ADN simple brin. Ces adaptateurs ont la possibilité de contenir au niveau de leur séquence après la clé, une étiquette de onze bases appelée MID (Multiplex Identifier) permettant le multiplexage d'échantillons. Tous les adaptateurs ont été dessinés de façon à permettre une ligature unidirectionnelle au niveau des fragments d'ADN réparés. En effet, ces derniers possèdent une extrémité 5' cohésive et une extrémité 3' franche permettant uniquement à cette dernière de se lier aux fragments d'ADN de l'échantillon. Enfin, les adaptateurs sont non phosphorylés au niveau de leurs extrémités 5', ce qui permet de limiter la formation d'éventuels dimères. Enfin, les produits de ligature devront être réparés avec une ADN polymérase à activité de déplacement de brin. **(4)** Immobilisation de la banque sur des billes magnétiques recouvertes de streptavidine par l'intermédiaire de l'adaptateur B portant une biotine sur son extrémité 5'. Suite à l'étape de ligature des adaptateurs, les fragments d'ADN peuvent soit ne pas porter d'adaptateur, soit un seul à une extrémité soit présentés les associations suivantes : AB, AA, BB. Seuls les fragments d'ADN portant au moins un adaptateur B biotinylé seront retenus au niveau de billes. **(5)** Réparation des jonctions extrémité 5'adaptateurs – extrémité 3' des fragments d'ADN. Les adaptateurs n'étant pas phosphorylés au niveau de leurs extrémités 5', il y a une interruption au niveau de la chaîne nucléotidique entre les extrémités 5'des adaptateurs et les extrémités 3' des fragments d'ADN. Cette réparation se fait grâce à l'action d'une ADN polymérase à activité de déplacement de brin qui va reconnaître ces interruptions, déplacer les brins d'adaptateurs commençant par une extrémité 3'OH et resynthétiser l'ADN double brin ; **(6)** Récupération de la banque d'ADN simple par dénaturation chimique. Les banques comportant deux adaptateurs B de chaque côté ne pourront pas permettre le relargage d'un brin d'ADN non biotinylé, les banques comportant un seul adaptateur B pourront être récupérées mais pas amplifiées au niveau de l'étape de PCR en émulsion (emPCR). Il n'est récupéré uniquement qu'une banque d'ADN simple brin flanqué de part et d'autre part d'un adaptateur A en extrémité 5' et d'un adaptateur B en extrémité 3'.

Différentes applications de séquençage sont possibles grâce à la technologie 454, avec notamment la possibilité de séquencer à la fois à très haut-débit (un million de lectures), de façon globale et sans à priori, des métagénomes de manière dite *de novo* **(Figure 22)** (Petrosino et al 2009). Le principal avantage d'utilisation du pyroséquençage 454 demeure la longueur des lectures produites (un million de lectures de 400 à 700 bases suivant la plateforme) dans un temps de séquençage à haut-débit relativement court (entre 10 et 20h) (Glenn 2011). Cette perspective est très intéressante pour des études métagénomiques où l'identification de fragments de grande taille apporte une information phylogénétique et taxonomique plus pertinente (Wommack et al 2008). De plus, cette approche facilite l'assemblage des lectures et permet donc d'isoler des gènes ou des opérons entiers. Cependant, certaines limites technologiques sont à prendre en considération comme par exemple le coût relativement élevé de cette méthode (environ 13€ la mégabase) et des erreurs de séquençage en relation notamment avec la présence d'homopolymères. En effet, le pyroséquençage n'ayant pas de terminaison pendant la synthèse empêchant l'incorporation multiple de nucléotides au cours d'un cycle de séquençage, ce dernier se base sur l'intensité de la lumière émise pour déterminer le nombre de bases incorporées. Ainsi, le problème fréquemment rencontré est la perte de la relation de linéarité entre l'intensité de lumière émise et le nombre de nucléotides incorporés, aboutissant soit à des insertions de bases, soit à des délétions (Margulies et al 2005, Rothberg and Leamon 2008). Ces erreurs de séquençage peuvent conduire à une surreprésentation de la diversité, ce qui a pu être observé lors d'une étude des communautés archées dont les méthanogènes dans des sédiments (Porat et al 2010).

ii. Le séquençage Illumina

Cette technologie de séquençage est basée sur le principe de terminaison réversible (Cyclic Reversible Termination, CRT) par utilisation de désoxyribonucléotides triphosphates (dNTPs) modifiés et fluorescents (Guo et al 2008, Metzker 2005), et implique une méthode de synthèse en trois étapes : l'incorporation, la mesure de la fluorescence et le clivage (Metzker 2005, Metzker 2010). Premièrement, une ADN polymérase va initier la synthèse du brin complémentaire au niveau de l'amorce de séquençage en incorporant un dNTP modifié portant un fluorophore et un groupement protecteur au niveau de l'extrémité 3'-O du ribose. Deuxièmement, suite à l'incorporation, les dNTP modifiés résiduels seront éliminés par lavage et la fluorescence émise est enregistrée en temps réel permettant de déterminer la nature de la base incorporée au niveau de la séquence. Enfin, l'étape de clivage, suivie d'une nouvelle étape de lavage, élimine le groupement protecteur en 3'-O inhibant la réaction de

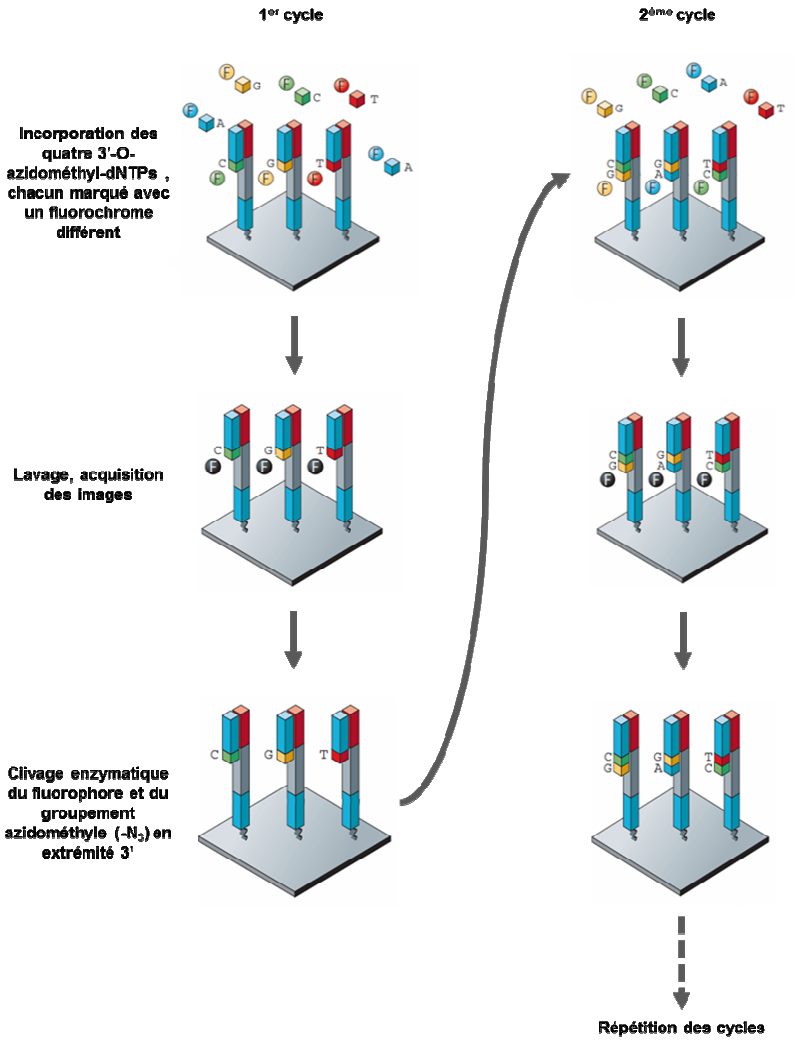

Figure 23. Principe du séquençage Illumina par la technique CRT « Cyclic Reversible Termination ». (Modifié d'après Metzker, 2010).

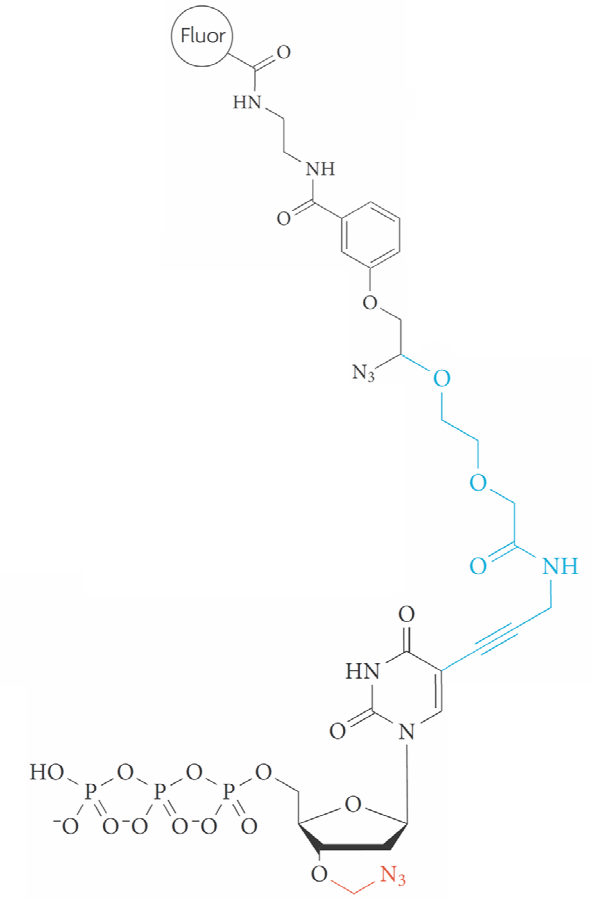

Figure 24. Formule chimique d'un 3′-O-azidométhyl-dNTPs utilisé pour le séquençage de type Illumina. La flèche indique le site de clivage séparant le fluorophore du nucléotide, en rouge est surligné le groupement 3' protecteur et en bleu le groupement résiduel attaché à la base et s'accumulant au cycle suivant.

polymérisation, ainsi que le fluorophore pour permettre une nouvelle étape d'incorporation (**Figure 23**). Ce principe est dérivé de la méthode de Sanger, où contrairement à cette dernière qui utilise des didésoxyribonucléotides triphosphates (ddNTPs) bloquant la polymérisation, la méthode CRT offre la possibilité de bloquer la polymérisation de manière réversible. La clé de cette méthode réside en l'utilisation de bases bloquantes, comme les 3′-O-azidométhyl-dNTPs, couplées à des fluorophores et portant un groupement azidométhyle ($-N_3$) sur l'extrémité 3'-O du ribose, pouvant être clivés chimiquement pour restaurer une extrémité 3'OH libre et ainsi rétablir la polymérisation (**Figure 24**).

La technologie de séquençage Illumina (Solexa) est issue de la combinaison de la méthode CRT utilisant quatre fluorophores différents (Turcatti et al 2007), de l'amplification par PCR en ponts sur phase solide (Adessi et al 2000), des nanotechnologies (Fedurco 2006) et des technologies informatiques pour l'acquisition et le traitement des images. Comme pour la technologie 454, la technologie Illumina offre un coût de séquençage réduit (Glenn 2011) (**Tableau 5**), avec également la possibilité de séquencer des métagénomes de manière *de novo* à un débit encore plus important que pour la technologie 454 (un milliard de lectures). La technologie Illumina a apporté une révolution sans précédent au niveau du débit de séquençage (un milliard de lectures d'une taille de 100 bases) et du coût (environ 0,1€ la mégabase). Ainsi, le très haut-débit proposé par cette plateforme, associé à son coût relativement réduit, est un atout énorme pour les études métagénomiques, qui recherchent une très grande profondeur de séquençage pour étudier la diversité phylogénétique et fonctionnelle des communautés bactériennes (Li et al 2011, Qin et al 2010, Yu and Zhang 2012). Cependant, le séquençage Illumina possède certaines limites dont notamment la faible longueur des lectures (au maximum 100 bases voire 2×100 bases séquencées à partir des deux extrémités) et le temps de séquençage important (environ huit jours). De même des erreurs de séquençage sont observées lors d'une mauvaise incorporation ou d'une mauvaise interprétation de la fluorescence émise (« base calling »), aboutissant à des substitutions de bases. Ces substitutions sont fréquemment rencontrées au niveau des sites moléculaires précédés d'un G avec substitution préférentielle d'une adénine par une cytosine (Dohm et al 2008, Qu et al 2009).

2.3.2.b Vers une troisième génération de séquençage

Malgré l'énorme révolution des techniques de séquençage aboutissant à l'émergence des techniques dites de deuxième génération, le problème majeur à l'heure actuelle réside

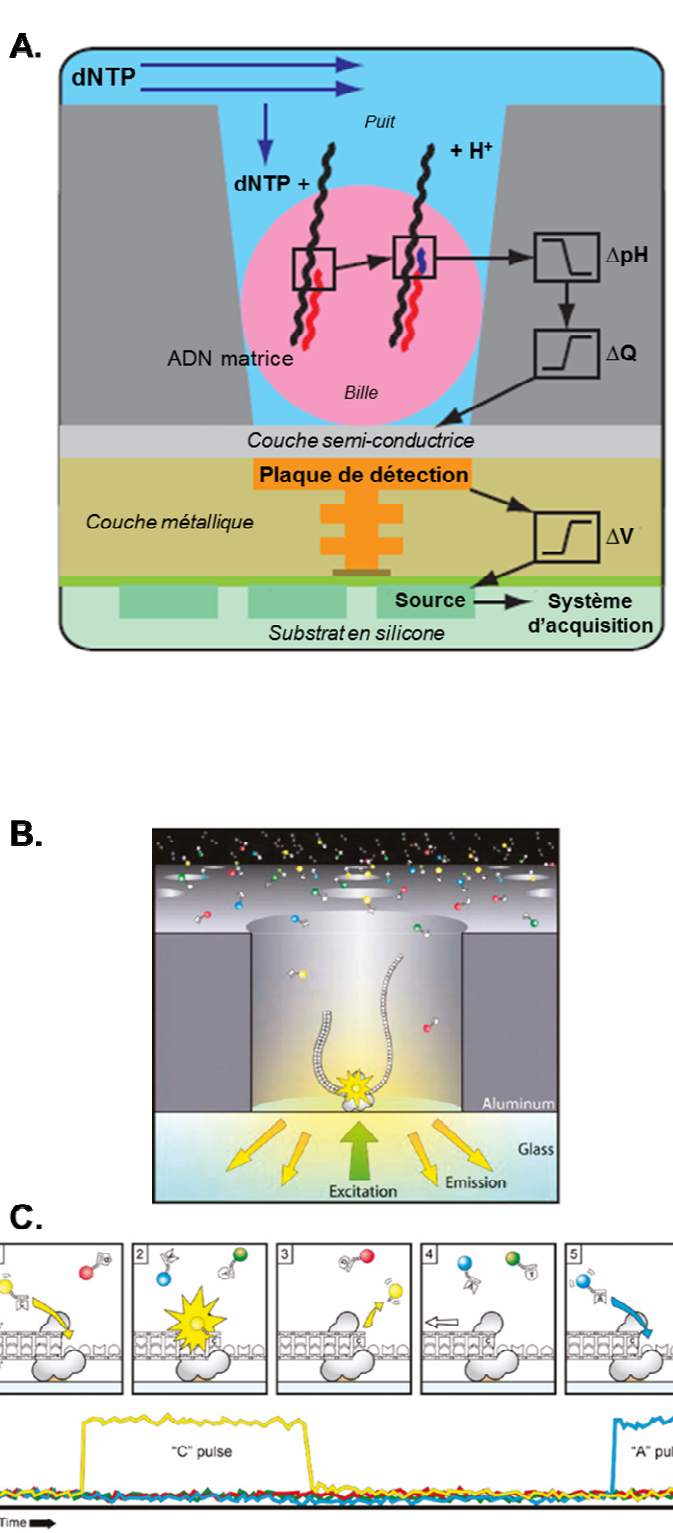

Figure 25. Représentation schématique du mode de fonctionnement des nouvelles technologies de séquençage dites de troisième génération. (A) Système Ion Torrent où lors de l'incorporation d'un dNTP des ions H^+ sont libérés modifiant le pH (ΔpH) à l'intérieur du puits. Ce changement de pH est converti *via* une couche semi-conductrice et une plaque de détection en une différence de potentiel (ΔV), qui sera par la suite retranscrite en un signal numérique; **(B)** Système PacBio RS au niveau d'un nanopuits « Zero Mode Waveguide » contenant une ADN polymérase fixé au fond du puits. La configuration du ZMW permet une détection uniquement de la fluorescence sur le fond du puits ; **(C)** Représentation au niveau du système PacBio RS de l'incorporation de deux nucléotides portant deux fluorophores différents. (Redessiné et adapté de Rothberg et al., 2011 et Niedringhaus et al., 2011)

dans la taille relativement courte des fragments séquencés, compliquant fortement les étapes d'assemblage et la reconstruction de génomes complets (Morales and Holben 2011). En outre, malgré la quantité de données massives générées par ces nouvelles technologies de séquençage, ces dernières ne suffisent pas à accéder aux génomes de l'ensemble des populations microbiennes présentes au sein des environnements complexes. En effet, Quince et al. (2008) ont estimé que pour certains environnements complexes, l'effort de séquençage actuel est insuffisant et doit être multiplié par 10 000 afin de couvrir 90% de la diversité microbienne qu'ils hébergent. Même si les capacités de séquençage ont fortement évolué, cette couverture de la diversité n'est pas envisageable autant sur un plan technologique que financier (Quince et al 2008).

Actuellement une troisième génération de séquençage est apparue, exploitant au maximum les progrès des nanotechnologies permettant de simplifier au maximum les technologies de séquençage (Blow 2008, Glenn 2011, Munroe and Harris 2010, Pareek et al 2011). Ces nouvelles méthodes utilisent une enzyme (une ADN polymérase ou une exonucléase) immobilisée de manière individuelle sur un support solide et permettant de séquencer une seule molécule d'ADN à la fois. Deux nouvelles technologies de séquençage dites de troisième génération sont actuellement disponibles et commercialisées : le système PacBio RS (Pacific Biosciences) basé sur le principe de séquençage SMRT « Single Molecule Real Time Technology » (Eid et al 2009, Korlach et al 2010, McCarthy 2010) et le système Ion Torrent (Life technologies) basé sur le principe « Ion semiconductor sequencing » (Rothberg et al 2011) (**Figure 25**). Ces deux méthodes utilisent un réseau de nanopuits comportant une ADN polymérase permettant la réaction de séquençage. Le système PacBio RS utilise des puits nanophotoniques appelés « zero-mode waveguide » (Levene et al 2003) d'un volume de l'ordre du zeptolitre (10^{-21} litre) permettant de canaliser et d'éviter la propagation de la lumière visible de grande longueur d'onde hors des puits, et permettant une détection efficace des signaux fluorescents émis lors de l'incorporation des nucléotides par l'ADN polymérase. Le système Ion Torrent n'utilise pas de dNTP fluorescents, mais un semi-conducteur détectant une différence de potentiel créée par la libération d'ions H^+ suite à l'incorporation d'un dNTP par l'ADN polymérase. Cette technologie fonctionne, à la différence du PacBio RS, par un ajout séquentiel des dNTP qui sont apportés les uns après les autres. Cette troisième génération de séquençage améliore encore de manière significative les capacités de séquençage (1 Gigabase en 2h pour Ion Torrent et 10 Mégabases en 2h pour PacBio RS) avec l'obtention de lectures de grande taille pour certaines technologies (>100

Tableau 6. Comparaison des différentes plateformes de séquençage de troisième génération

	Séquenceur (Société)	Méthode de séquençage	Longueur des lectures	Débit (Mb par run)	Temps de séquençage	Coût (par Mb)	Disponibilité
3ème génération	**Ion Torrent – '316'chip** (Life technology)	Séquençage par synthèse (détection H+)	>100	>100	2h	<7,5$	oui
	Ion Torrent – '318'chip (Life technology)	Séquençage par synthèse (détection H+)	>100	>1000	2h	~0,93$	oui
	PacBio RS (Pacific Biosciences)	Séquençage de molécules individuelles en temps réel (SMRT)	860-1100	5-10	0,5 - 2h	11-180$	oui
	GridION (Oxford Nanopore)	Séquençage de molécules individuelles (exonucléase et ADN polymérase Phi 29)	<100kb	ND	<1 jour	ND	2013
	MinION (Oxford Nanopore)	Séquençage de molécules individuelles (exonucléase et ADN polymérase Phi 29)	<100kb	>1000	<1 jour	ND	2013
	Optipore (Noblegen Biosciences)	Séquençage de molécules individuelles	ND	ND	ND	ND	2014

ND: non déterminé

SMRT: Single Molecule Real Time Technology

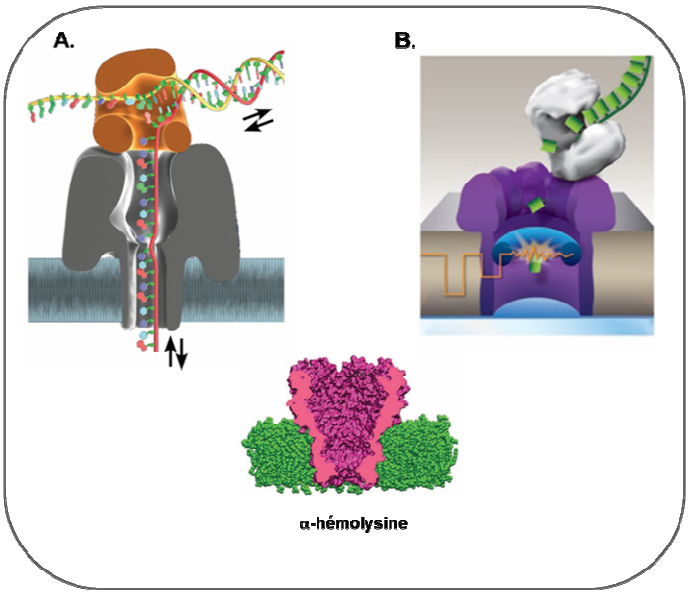

Figure 26. Applications du séquençage de troisième génération de type Nanopore. (A) L'ADN est inséré à travers le nanopore avec une vitesse contrôlée par l'ADN polymérase (marron). Un nanopore d'α-hémolysine (gris) est intégré au sein d'une bicouche lipidique séparant deux chambres contenant un tampon à base de KCl. Le brin d'ADN matrice (rouge) est introduit à travers le pore en appliquant un champ électrique. Son mouvement vers l'intérieur ou l'extérieur du nanopore (flèches) est contrôlé par le champ électrique appliqué et à l'activité de l'ADN polymérase. L'information de séquence est obtenue par les changements du courant ionique à travers le nanopore. **(B)** Une exonucléase (gris) fixé sur un nanopore d'α-hémolysine (bleu) dégrade le brin d'ADN en faisant tomber les bases (vert) une par une à travers le nanopore. L'information de séquence est déterminée par une modification de la différence de potentiel à travers le nanopore. (Modifié d'après Munroe et Harris, 2010 et Schneider et Dekker, 2012).

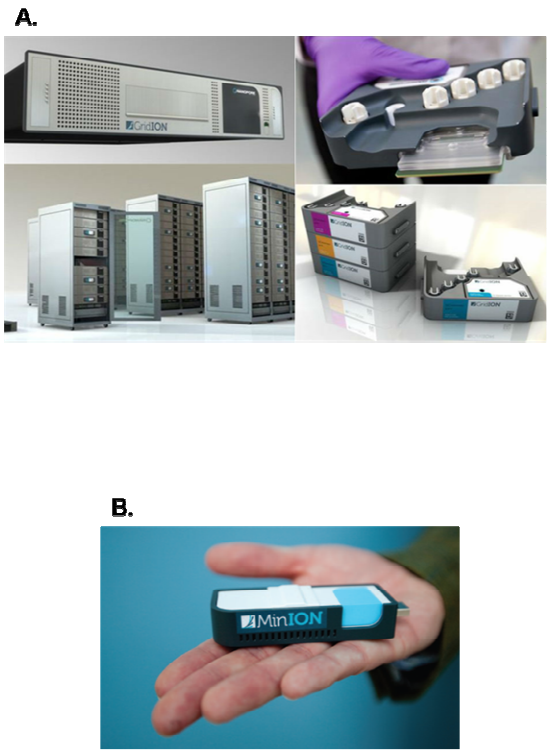

Figure 27. Systèmes de séquençage de troisième génération Nanopore prochainement commercialisés. (A) le système GridION et **(B)** le système MinION se matérialisant sous la forme d'une clé USB.

bases pour Ion Torrent et >1000 bases pour PacBio RS) (**Tableau 6**) (Glenn 2011). Ces nouvelles techniques ont été récemment appliquées pour le séquençage de génomes complets notamment celui du sérotype O104 : H4 d'*Escherichia coli* responsable d'une épidémie par consommation de graines germées en mai 2011 en Allemagne. Le génome complet de la souche a pu être séquencé *via* l'utilisation du système Ion Torrent, puis assemblé en seulement deux jours (Ahmed et al 2011). De même l'utilisation du PacBio RS a permis le séquençage complet des génomes de cinq souches de *Vibrio cholerae*, dont celle responsable de l'épidémie de choléra en Haïti en octobre 2010, et ceci en seulement 3 heures pour permettre des études ultérieures de génomique comparative (Chin et al 2011). Toutefois, même si ces technologies révolutionnent le domaine du séquençage, d'autres systèmes sont en cours de développement et doivent permettre à long terme de séquencer des molécules de grande taille pouvant atteindre une centaine de kilobases. Il est possible de citer la technologie Nanopore (Oxford Nanopore Technologies) proposant deux types d'application de séquençage : un système en cours de développement dans lequel une exonucléase est fixée sur un nanopore d'α-hémolysine (Clarke et al 2009, Timp et al 2010) et un système prochainement commercialisé courant 2012 utilisant un nanopore à base d'une porine A de *Mycobacterium smegmatis* (MspA) et une ADN polymérase phi29 (**Figure 26**) (Cherf et al 2012, Manrao et al 2012, Schneider and Dekker 2012). Ce système s'utilisera au sein de deux plateformes : l'une appelée GridION et l'autre MinION correspondant à un mini système de séquençage se présentant sous la forme d'une clé USB pouvant être branché à un ordinateur portable (Eisenstein 2012) (**Figure 27**). Une autre application des nanopores est exploitée par la société Noblegen Biosciences qui ambitionne de commercialiser en 2014 le système « optipore » (pour « optical detection » et « nanopore »), un séquenceur de paillasse de troisième génération combinant les nanotechnologies et un système de lecture optique (McNally et al 2010, Singer et al 2012).

Ces nouvelles méthodes ont été conçues pour faciliter de manière considérable la reconstruction de génomes complets, mais leurs applications restent encore limitées pour les échantillons métagénomiques du fait d'un taux d'erreur de séquençage relativement important (jusqu'à 16%) pour le système PacBio RS (Glenn 2011) et de la complexité des échantillons rendant problématique les étapes d'assemblage et d'annotation des séquences. Même si ces technologies émergentes sont très prometteuses pour des applications en écologie microbienne et plus précisément pour l'étude des environnements complexes, leur application à l'échelle du laboratoire reste problématique d'une part par le coût de l'équipement en

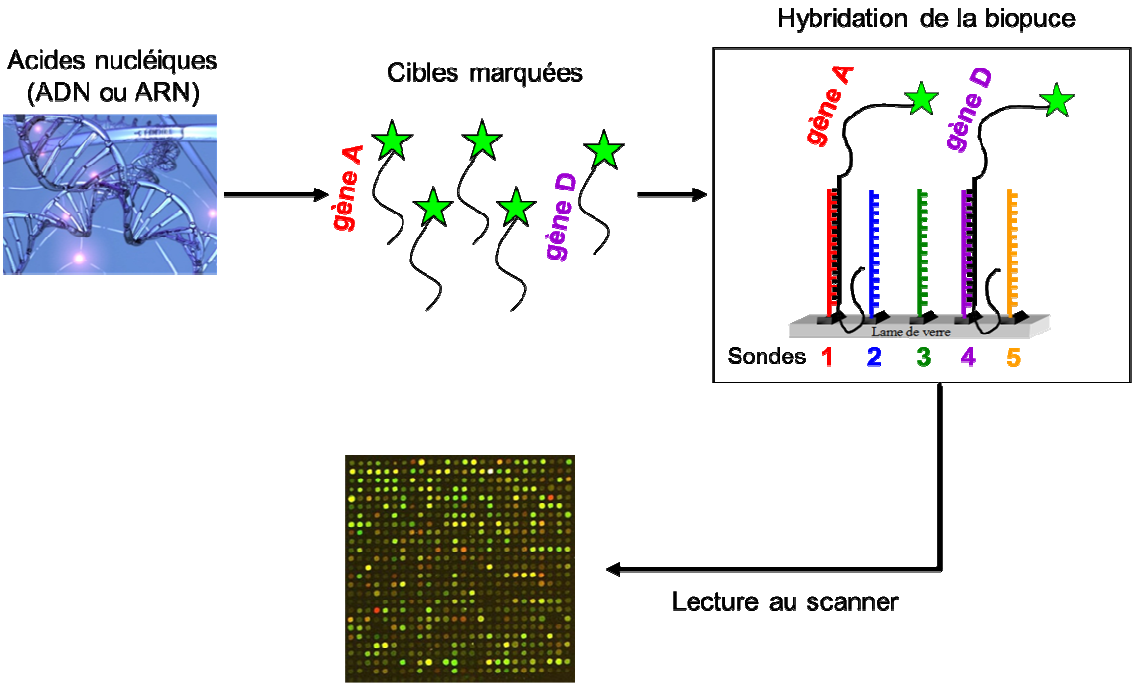

Figure 28. Principe des biopuces à ADN. Les cibles marquées à l'aide d'un fluorochrome s'hybrident spécifiquement avec les sondes qui leur sont complémentaires. L'analyse de l'image obtenue permet de déterminer quels sont les gènes présents dans l'échantillon.

machines performantes (qui du reste sont en perpétuelle évolution), mais surtout par la masse de quantité de données produites restant très délicate à analyser. Cette perspective implique obligatoirement la collaboration de plusieurs équipes de recherche, de disposer de moyens de calcul, de traitement et de stockage conséquents. Des approches moléculaires alternatives sont à privilégier comme les biopuces à ADN, quin permettent à l'heure actuelle d'aborder les problématiques d'écologie microbienne pour des environnements complexes et sont qualifiées d'outils haut-débit nécessitant des temps de traitement des données beaucoup plus courts.

2.4 Les biopuces à ADN

Apparues au milieu des années 90 suite au séquençage des premiers génomes et issues de la rencontre des domaines tels que la microélectronique, les microsystèmes et la biologie, les biopuces ont été au départ mises au point pour l'étude simultanée de l'expression de tous les gènes d'un organisme (Schena et al 1995). Les biopuces à ADN ont connu un essor considérable ces vingt-cinq dernières années pour répondre aux problématiques de l'écologie microbienne.

2.4.1 Principe

Les biopuces à ADN sont dérivées des techniques d'hybridation des acides nucléiques du Southern blot (Southern 1975) et du Dot blot (Kafatos et al 1979), mais contrairement à ces dernières, l'hybridation est dite inverse puisque ce sont les sondes et non pas les cibles qui sont fixées sur un support solide (Ehrenreich 2006). Les sondes peuvent correspondre à de l'ADNg, des produits PCR, de l'ADNc ou encore à des oligodésoxyribonucléotides fixées sur une lame de verre, et elles vont agir comme des « hameçons » moléculaires en reconnaissant leurs cibles par complémentarité des bases. Les cibles sont des produits PCR, de l'ADNg, de l'ADNc des ARN. Plusieurs échantillons peuvent être hybridés simultanément en utilisant des marquages à l'aide de fluorophores différents (généralement les cyanines Cy3 et Cy5) (**Figure 28**). Une seule biopuce, sur laquelle une compartimentation physique est possible, peut donc permettre l'analyse de différents échantillons biologiques en une seule expérience ainsi que l'identification d'un grand nombre de séquences, puisque les formats actuels de biopuces permettent la fixation de plusieurs millions de sondes différentes. Suite à l'étape d'hybridation entre les sondes et les cibles fluorescentes, les duplex formés sont détectés à l'aide d'un scanner. Un faisceau laser va balayer toute la surface de la lame et exciter les fluorophores pour entraîner une émission de lumière. Les intensités lumineuses sont

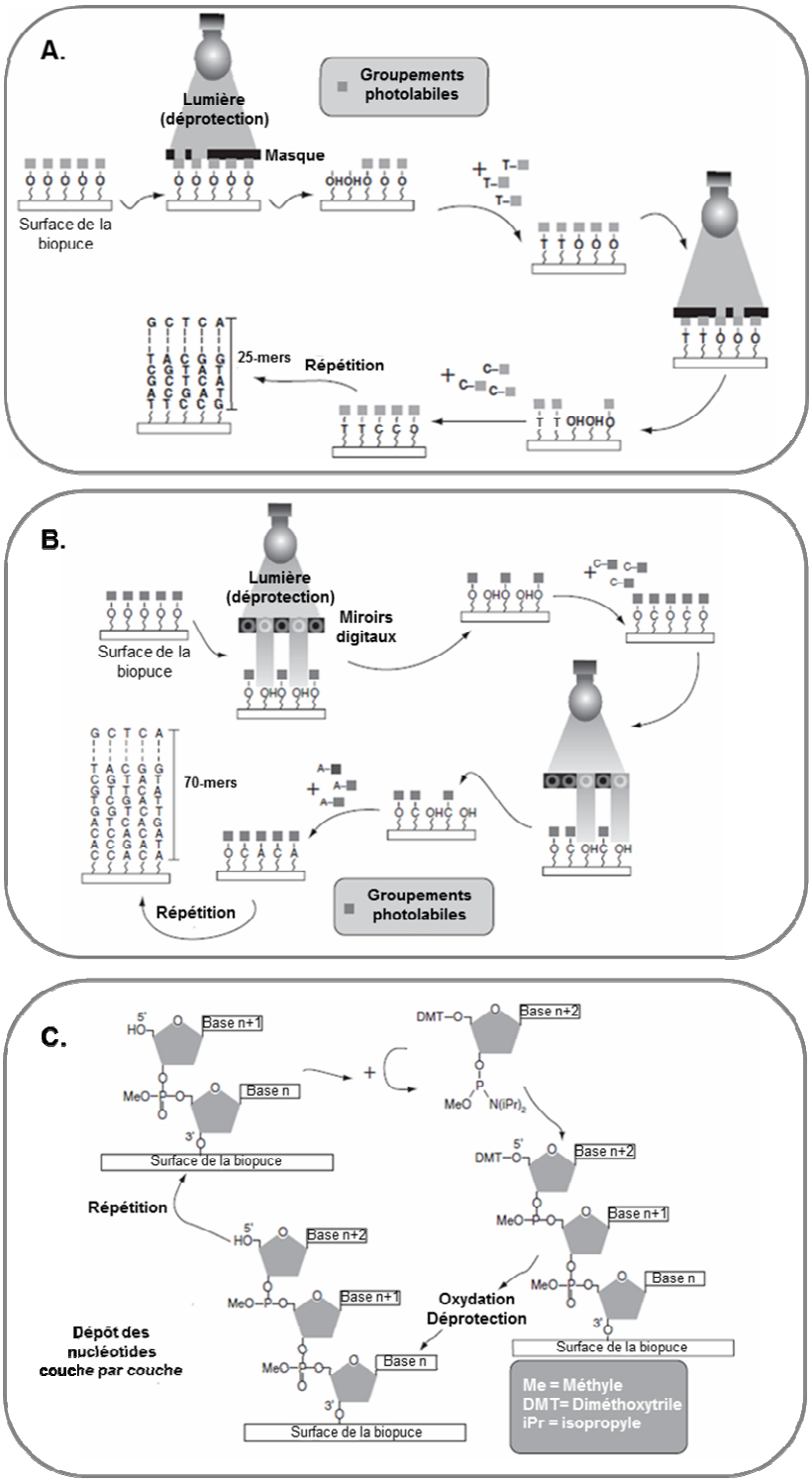

Figure 29. Méthode de synthèse des biopuces à ADN dites *in situ*. (A) Synthèse dirigée par la lumière utilisant des masques photolithographiques ; **(B)** synthèse dirigée par la lumière utilisant des miroirs digitaux ; **(C)** synthèse séquentielle par incorporation directe par la méthode au phosphoramidite. (Modifié et redessiné d'après Nakaya et al., 2007)

collectées puis transformées en signal électrique permettant d'évaluer quantitativement et qualitativement les échantillons hybridés.

Actuellement, les sondes oligodésoxyribonucléotidiques sont privilégiées du fait de leur faible coût et de leur facilité de synthèse (Relógio et al 2002). Il existe deux principales technologies de fabrication des biopuces à ADN : les biopuces dites *ex situ* nécessitant une préparation des sondes au préalable avant leur greffage sur le support solide, et les biopuces dites *in situ* pour lesquelles la synthèse des sondes est directement réalisée sur la lame de verre. Cette dernière technologie de synthèse de biopuce à ADN est à privilégier pour les études haut débit en écologie microbienne du fait de son coût moindre et de la possibilité de greffer à très haute densité (Dufva 2005, Dufva 2009, Kawasaki 2006).

2.4.2 Les biopuces *in situ*

Les biopuces dites *in situ* permettent une flexibilité de synthèse des sondes oligodésoxynucléotidiques. Cette synthèse peut se faire par une méthode de photolithographie dérivée des techniques de fabrication de l'industrie des semi-conducteurs (Dalma-Weiszhausz et al 2006). Les lames de verres sont traitées par une solution de silane (N-(3-triethoxysilylpropyl)-4-hydroxy-butyramide) où les groupements OH libres des amides terminaux sont bloqués par un groupement photolabile tel que le méthylnitropoperonyloxycarbonyl (MeNPOC). Les groupements OH libres des amides terminaux à la surface du support solide sont ensuite activés par déprotection des groupements photolabiles. Cette déprotection est réalisée à l'aide de différents masques photolithographiques constitués de zones soit à base de chrome pour bloquer la lumière, soit à base de quartz pour la laisser passer au niveau des zones à activer. La synthèse est réalisée dans le sens 3'→5' par la méthode au phosphoramidite (White 1988) en incorporant au niveau des zones déprotégées des synthons ou désoxynucléotides modifiés portant également en 5' un groupement MeNPOC protecteur photolabile. Le cycle de traitement est ainsi répété jusqu'à ce que les oligodésoxynucléotides soient complétement synthétisés (**Figure 29A**). Cette méthode utilisant les masques photolithographiques permet de générer des sondes pouvant atteindre 25-mers, qualifiées de sondes courtes. D'autres méthodes de synthèse *in situ* basées sur le principe de la photolithographie permettent avec beaucoup plus de flexibilité, la conception des biopuces à ADN en réduisant les coûts et en augmentant la taille des sondes. Ces méthodes utilisent un jeu de miroirs digitaux pour remplacer les masques photolithographiques (Nuwaysir 2002). Cette méthode permet d'obtenir une très haut-densité

(des millions de sondes greffées) avec une taille maximale beaucoup plus importante (jusqu'à 80-mers) (**Figure 29B**). D'autres technologies existent également pour augmenter la flexibilité et réduire les coûts de synthèse comme la projection de bases, analogue à une projection de type « jet d'encre » (Wolber et al 2006). Cette technologie se base aussi sur une synthèse dans le sens 3'→5' par la méthode au phosphoramidite, en incorporant de manière séquentielle les quatre synthons (**Figure 29C**). Cette technique permet de contourner l'utilisation de masques et d'obtenir des sondes longues jusqu'à 70-mers, mais avec une densité de dépôt moins importante.

L'évolution des techniques de synthèse des sondes permet de disposer actuellement d'une technologie de biopuces à ADN à très haut débit utilisable en écologie microbienne pour étudier la structure et caractériser les capacités métaboliques des communautés microbiennes des environnements complexes.

2.4.3 Les biopuces à ADN en écologie microbienne

La première biopuce à ADN appliquée à l'écologie microbienne date de 1997. Elle était composée de neuf sondes ciblant le gène codant pour l'ARNr 16S et permettait l'identification de bactéries nitrifiantes (Guschin et al 1997). Depuis les biopuces à ADN ont été utilisées dans de nombreuses études en écologie microbienne (Wagner et al 2007, Zhou 2003). Différentes catégories de biopuces ont ainsi été utilisées dont les « Whole Genome Array » (WGA) permettant de cibler dans son intégralité les gènes d'un ou plusieurs microorganismes et pouvant être utilisées notamment pour caractériser des souches ou des *consortia* isolés d'environnements complexes (Wu et al 2004). Cependant, l'utilisation de ce type de biopuces pour l'étude *in situ* d'échantillons environnementaux est limitée en raison de l'importante complexité des communautés microbiennes composées en grande majorité de souches non caractérisées et pour lesquelles il n'existe aucune information de séquence. C'est pourquoi, l'utilisation de biopuces dites phylogénétiques (« Phylogenetic Oligonucleotide Array », POA) ou fonctionnelles (« Functional Gene Array, FGA) ciblant respectivement des biomarqueurs phylogénétiques et fonctionnels apparaît plus adaptée pour l'écologie microbienne (Gentry et al 2006, Wagner et al 2007). La difficulté majeure concernant l'utilisation de ce type de biopuces est la détermination et la sélection des sondes. Ainsi, la précision concernant la caractérisation des communautés microbiennes des environnements complexes repose sur l'efficacité des sondes où celles-ci doivent être (1) spécifiques, en reconnaissant uniquement leurs cibles sans entraîner d'hybridations aspécifiques, (2)

sensibles, en permettant la détection de cibles peu abondantes dans un échantillon, (3) uniformes, en ayant le même comportement et les mêmes capacités d'hybridation dans une condition donnée.

2.4.3.a Biopuces phylogénétiques

Les différentes méthodes moléculaires développées et appliquées à l'écologie microbienne pour l'identification et la classification des communautés microbiennes appartenant au domaine des procaryotes (bactéries et archées) ciblent principalement la séquence codante pour l'ARNr 16S. La présence de régions plus ou moins conservées entre les séquences d'ADNr 16S permettent d'accéder à une grande proportion des communautés. En effet, les régions hautement conservées de l'ADNr 16S permettent de déterminer des sondes ciblant des niveaux taxonomiques supérieurs comme la famille, l'ordre ou la classe, alors que les régions hypervariables peuvent discriminer les microorganismes à des niveaux plus fins comme le genre ou l'espèce (Huyghe et al 2008). De plus, l'accessibilité à une multitude de séquences présentes au sein des bases de données dédiées comme SILVA (Pruesse et al 2007), Greengenes (DeSantis et al 2006) et RDP (Ribosomal Database Project) (Cole et al 2009) permet d'affiner la détermination des sondes mais également de tester la spécificité des sondes en récupérant l'ensemble des séquences ADNr 16S disponibles ou caractéristiques de l'environnement étudié.

Dès lors de nombreuses biopuces phylogénétiques dédiées à l'écologie microbienne ont été mises au point avec des sondes présentant une taille comprise entre 18 et 25-mers. Elles ont été utilisées pour l'étude (1) de groupes bactériens spécifiques à différents niveaux taxonomiques comme la division des *Acidobacteria* (Liles et al 2010), des *Proteobacteria* (Sanguin et al 2006a, Sanguin et al 2006b) et des *Actinobacteria* (Kopecky et al 2011), l'ordre des *Rhodocyclales* (Loy et al 2005) et des *Actinomycetales* (Kyselková et al 2008), ou encore au genre *Burkholderia* (Schönmann et al 2009) ; (2) de communautés microbiennes d'environnements spécifiques comme des sols pollués par du hexachlorocyclohexane (Neufeld et al 2006) ou du tetrachloroethylène (Nemir et al 2010), la rhizosphère du blé (Sanguin et al 2009), le microbiote intestinal (Rajilić-Stojanović et al 2009), le microbiome oral humain (Preza et al 2008) ou encore de façon plus originale le métagénome de la cigarette (Sapkota et al 2009) ; (3) de groupes fonctionnels particuliers comme les sulfato-réducteurs (Loy et al 2002b) ou les bactéries nitrifiantes (Kelly et al 2005). Concernant l'étude des communautés méthanogènes, une biopuce ANAEROCHIP a été mise au point.

Elle est composée de 103 sondes ciblant les méthanogènes à différents niveaux taxonomiques (famille, genre et espèce) (Franke-Whittle et al 2009b). Elle a été appliquée dans d'autres études visant à évaluer la diversité des méthanogènes comme par exemple sur des résidus industriels d'olive (Goberna et al 2010) ou encore des boues provenant de différentes usines de production de biogaz (Walter et al 2012). En parallèle, une biopuce phylogénétique plus généraliste comme la PhyloChip (Brodie et al 2006), contenant environ 500 000 sondes de 25-mers et ciblant près de 9000 OTUs, permet de couvrir la quasi totalité de la diversité des communautés procaryotes répertoriée dans les bases de données dont les méthanogènes. Cet outil a été utilisé pour étudier les communautés microbiennes issues d'environnements complexes : l'air des avions (Korves et al 2012), des nappes phréatiques contaminées par du trichloroéthylène (Lee et al 2012), des sols de prairies (Cruz-Martínez et al 2009, DeAngelis et al 2008, He et al 2011a), les roches volcaniques (Kelly et al 2010, Kelly et al 2011), des sédiments de rivières contaminés par des métaux lourds (Rastogi et al 2011), des anciennes mines contaminées (Rastogi et al 2009, Rastogi et al 2010), ou encore des sols de l'Antarctique (Yergeau et al 2008). Cependant, en raison d'une connaissance partielle des microorganismes de l'environnement qui apparaissent encore être en grande majorité inconnus (Amann et al 1995, Pace 1997), les informations obtenues par ces approches ne permettent d'identifier que les espèces pour lesquelles des séquences sont disponibles. Par conséquent des stratégies de détermination de sondes dites « exploratoires » ont vu le jour afin d'appréhender toute la diversité sans *a priori* sur les séquences d'ADNr 16S ciblées (Dugat-Bony et al 2012b), avec notamment l'utilisation de logiciels de sélection de sondes exploratoires comme PhylArray (Militon et al 2007).

Les biopuces phylogénétiques ont démontré leur pertinence en écologie microbienne et leur complémentarité aux nouvelles techniques de séquençage. En effet, des études ont montré une forte corrélation entre les résultats des biopuces et des NGS. Par exemple, l'utilisation de la HITChip *versus* le pyroséquençage des régions V4 et V6 de l'ADNr 16S au sein d'échantillons de selles de patients âgés, a montré une bonne corrélation des résultats au niveau du phylum (r = 0,94), de la classe (r = 0,93) ou encore de l'ordre (r = 0.94) (Ahmed et al 2009). Le même résultat a pu être retrouvé au niveau du microbiome humain, avec l'utilisation de la Human Oral Microbe Identification Microarray (HOMIM) (Preza et al 2008) *versus* le pyroséquençage des régions V3-V5 de l'ADNr 16S, où une forte corrélation des résultats au niveau phylum et du genre a pu être obtenue (Ahn et al 2011). Néanmoins, même si les biopuces phylogénétiques apportent des informations précises sur la composition

des communautés microbiennes au sein d'environnements complexes, elles ne permettent pas l'identification des capacités métaboliques. Ceci est d'autant plus problématique quand différents membres d'un même groupe de microorganismes présentent des capacités métaboliques différentes et qu'ils ne peuvent être différenciés sur la seule base de leurs signatures moléculaires d'ADNr 16S. Ainsi, l'étude de capacités métaboliques particulières comme la production de méthane, ou la biodégradation de polluants aromatiques, nécessite l'utilisation préférentielle de biopuces fonctionnelles (FGA) ciblant directement les gènes impliqués dans les processus métaboliques d'intérêt (He et al 2008).

2.4.3.b Biopuces fonctionnelles

L'utilisation des biopuces fonctionnelles nécessite au préalable un choix des gènes cibles basé sur des critères précis. Ces gènes doivent (1) coder pour une enzyme clé dans la voie métabolique ciblée ; (2) présenter des régions suffisamment conservées permettant la détermination de sondes mais aussi suffisamment variables pour discriminer les gènes codants pour des enzymes différentes ; (3) être représentés par un maximum de séquences dans les bases de données permettant de couvrir un maximum de variants. Généralement, les sondes longues (50 à 70 mers) sont privilégiées pour la conception de telles biopuces fonctionnelles, puisqu'elles offrent une meilleure sensibilité tout en gardant une spécificité suffisante en relation avec la variabilité généralement rencontrée entre les gènes fonctionnels (Gentry et al 2006). Il a été montré que l'utilisation de sondes de 50-mers permet de différencier des cibles possédant en général moins de 88% d'identité avec leur sonde tout en conservant une sensibilité suffisante pour détecter au minimum des cibles à hauteur de 5 à 10 ng d'ADNg seul ou à hauteur de 50 à 100 ng d'ADNg en mélange (Rhee et al 2004). De tels seuils de détection correspondent à la mise en évidence de cibles ADN provenant seulement de 10 cellules ou de 5% des populations présentes dans une communauté bactérienne (Gentry et al 2006). En outre, il a été montré que pour des quantités entre 1 à 100ng d'ADNg (pur ou en mélange) une relation linéaire pouvait être établie entre intensité du signal et quantité de cible hybridée, et permettre une analyse semi-quantitative au niveau de cette gamme (Wu et al 2001).

La première biopuce fonctionnelle était composée d'environ 100 sondes construites à base de produits PCR et ciblant différents gènes du cycle de l'azote (Wu et al 2001). Depuis, l'utilisation de sondes oligonucléotidiques a permis, au cours de ces dix dernières années, la conception à haute densité et à moindre coût de biopuces fonctionnelles dédiées ciblant un ou

plusieurs métabolismes particuliers, et d'autres plus généralistes s'intéressant aux nombreux gènes impliqués dans la plupart des réactions biochimiques et des cycles biogéochimiques. Les biopuces fonctionnelles spécifiques ont été élaborées pour répondre à des questions biologiques précises comme celles en relation avec la résistance à un antibiotique (Call et al 2003), à la dégradation d'hydrocarbures aromatiques polycycliques et la résistance aux métaux (Rhee et al 2004), à la dégradation de solvants chlorés (Dugat-Bony et al 2012a), à la dégradation des polychlorobyphényls (PCBs) (Denef et al 2003), du benzène (Iwai et al 2007, Iwai et al 2008), à des facteurs de virulence bactérienne (Jaing et al 2008, Miller et al 2008), au cycle du méthane (Bodrossy et al 2003, Stralis-Pavese et al 2011), de l'azote (Duc et al 2009, Steward et al 2004, Taroncher-Oldenburg et al 2003, Tiquia et al 2006, Ward and Bouskill 2011, Ward et al 2007) et du soufre (Rinta-Kanto et al 2011). Plus récemment une biopuce fonctionnelle dédiée à l'étude des procédés de bioremédiation, la « BiodegPhyloChip » a été mise au point pour détecter 1057 gènes impliqués dans la dégradation de 133 polluants et mettre ainsi en avant les capacités métaboliques des communautés microbiennes de différents sites contaminés (Pathak et al 2011). Cependant, même si cet outil permet de couvrir une large gamme d'informations sur la dégradation de nombreux polluants, les sondes déterminées ne ciblent pas tous les gènes impliqués dans le processus de biodégradation ni même l'ensemble des variants de chaque gène.

Des biopuces fonctionnelles plus généralistes ont été mises au point et sont actuellement disponibles comme la GeoChip (He et al 2011b) qui a subi différentes évolutions, permettant de passer d'une première biopuce ciblant 2402 gènes (GeoChip 1.0) (He et al 2007) à 57 000 (GeoChip 3.0) (He et al 2010b). L'évolution de cet outil se poursuit avec une prochaine version 4.0 constituée d'environ 84 000 sondes permettant la détection de 152 000 gènes et couvrant les cycles du carbone (comprenant la méthanogénèse), de l'azote, du soufre, du phosphate, la réduction et la résistance aux métaux, la biodégradation de contaminants organiques mais aussi les voies de réponses aux stress ainsi que les phages et les éléments impliqués dans la virulence bactérienne. Les différentes versions de la GeoChip sont actuellement les biopuces fonctionnelles les plus utilisées en écologie microbienne pour caractériser la diversité fonctionnelle des environnements complexes. Leurs applications ont été diverses avec par exemple l'étude de sites pollués par des fuites de pétrole (Beazley et al 2012), par des lixiviats provenant de décharge (Lu et al 2012), par de l'uranium (Liang et al 2012, Van Nostrand et al 2009, Van Nostrand et al 2011, Xu et al 2010), par divers métaux (Epelde et al 2010, Xie et al 2010), par des huiles (Liang et al 2009, Liang et al 2010), par de

l'arsenic (Xiong et al 2010), de la dégradation des hydrocarbures aromatiques polycycliques (Ding et al 2012), des PCBs (Leigh et al 2007), de l'hexachlorobenzène (Tas et al 2009) ou encore de sols de l'Antarctique en lien avec le réchauffement (Yergeau et al 2011). Cependant, même si ces outils ciblent une large gamme de gènes impliqués dans tous les processus métaboliques globaux, leurs applications restent limitées à l'image des biopuces phylogénétiques, par leur incapacité à appréhender la diversité fonctionnelle encore non décrite. Certaines stratégies sont actuellement capables, sur la base des séquences disponibles et répertoriées dans les bases de données, de sélectionner des sondes exploratoires aptes à identifier certes les séquences des gènes ciblés présentes dans les bases, mais également celles de nouveaux variants n'ayant pas encore été caractérisées. Il est possible de citer les logiciels Metabolic Design (Terrat et al 2010) et HiSpOD (Dugat-Bony et al 2011) élaborés pour la détermination de sondes pour biopuces fonctionnelles et intégrant le caractère dit exploratoire.

2.5 Conclusion

Les méthodes moléculaires ont révolutionné les capacités d'études des microorganismes au sein des environnements complexes en offrant une vision de plus en plus intégrative et exhaustive des communautés microbiennes. Les méthodes classiques basées sur la culture ou sur les empreintes moléculaires ont laissé la place au développement des techniques à haut-débit permettant de s'affranchir des faiblesses des méthodes pionnières en écologie microbienne. Les biopuces à ADN apparaissent être des outils moléculaires ayant un potentiel pour pouvoir lier la structure et la fonction des microorganismes et les nouvelles techniques de séquençage apportent de nombreuses informations de séquences à des niveaux jusqu'alors inégalés. Ces deux méthodes peuvent être combinées pour permettre la mise en place de nouvelles approches comme la capture de gènes pour une exploration fine et ciblée des communautés microbiennes au sein des environnements complexes.

3. La capture de gènes appliquée à la métagénomique

L'exploration des environnements dans leur globalité nécessite un effort de séquençage très important, dépassant même les capacités actuelles des NGS (Delmont et al 2012, Quince et al 2008, Vogel et al 2009). A l'heure actuelle, l'utilisation de ces nouvelles technologies reste limitée pour explorer finement les environnements complexes et coûteuse pour de nombreuses structures de recherche (Bentley 2006, Roh et al 2010). Une alternative intéressante est de pouvoir réduire la complexité des échantillons en ciblant, en isolant spécifiquement et en enrichissant de manière efficace des séquences nucléiques d'intérêt pour permettre l'application des NGS sur ces mêmes séquences ciblées.

3.1 Les nouvelles techniques d'enrichissement

La plus ancienne et la plus familière est sans doute la PCR qui utilise un couple d'amorce pour amplifier de manière spécifique une région génomique. Comme pour n'importe quelle molécule d'ADN, les produits PCR peuvent être utilisés comme matrice pour le séquençage utilisant les NGS. Il est possible à l'heure actuelle grâce notamment à la PCR dite multiplexe, de générer différents types d'amplicons en une seule réaction. Malgré un progrès considérable ces dernières années, la PCR multiplexe, utilisée en combinaison avec les NGS pour l'enrichissement de régions génomiques d'intérêt (Tewhey et al 2009, Varley and Mitra 2008) reste difficile à optimiser. En effet, elle est soumise à de nombreux biais (cf § 2.2.2.c) avec en outre, une perte d'efficacité au cours de la réaction PCR en relation avec la taille et le nombre de régions à amplifier (Good 2011). Un des défis majeurs actuels en écologie microbienne est de pouvoir explorer les environnements complexes en isolant des gènes ou des opérons en entier tout en s'affranchissant de la PCR. De nouvelles méthodes moléculaires d'enrichissement proposent de lever ce verrou technique pour enrichir spécifiquement de grandes régions génomiques d'intérêt.

3.1.1 Les techniques d'hybridation soustractives

Les techniques d'hybridation soustractives ont été utilisées en génomique bactérienne et eucaryote pour l'enrichissement de fragments d'ADN suite à l'action d'enzymes de restriction. Cette méthode est un bon complément au séquençage des génomes entiers afin de mettre en avant les différences de séquences qu'il peut exister entre deux génomes. Une variante de cette technique, appelée hybridation soustractive suppressive, porte quant à elle sur les différences d'expression des gènes (Diatchenko et al 1996, Huang et al 2007). Ainsi d'une manière générale, les approches de soustraction peuvent être utilisées comme méthode

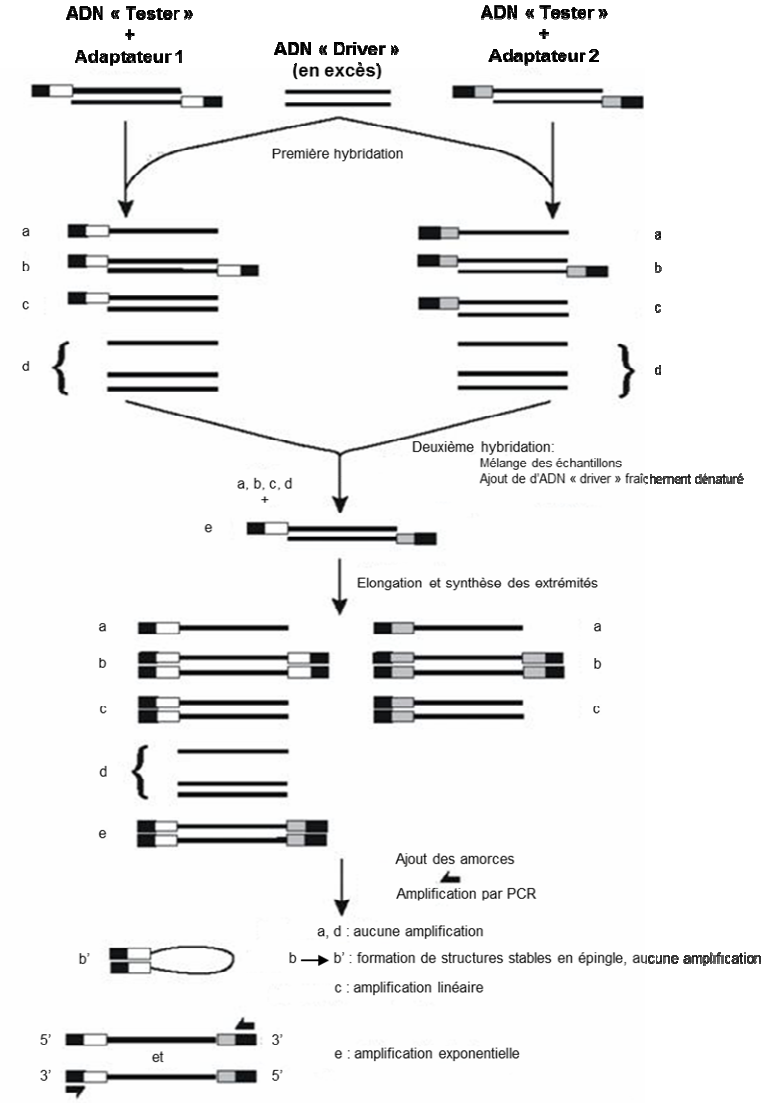

Figure 30. Principe de l'hybridation soustractive suppressive. La première étape commence par l'extraction des acides nucléiques des deux souches bactériennes à comparer (ADN « tester » et « driver »). Ces différents acides nucléiques sont ensuite digérés par une enzyme de restriction coupant ces derniers de manière fréquente. L'ADN « tester » est ensuite subdivisé en deux groupes, chacun étant lié à un adaptateur différent nommés adaptateurs 1 et 2. Une première hybridation implique la mise en solution d'une quantité en excès d'ADN « driver » au niveau des deux groupes d'ADN « tester ». Les deux groupes d'acides nucléiques sont ensuite dénaturés séparément puis soumis à une renaturation progressive aboutissant à la formation de différentes conformations moléculaires dans chacun des deux groupes, nommées A, B, C et D. Les molécules de type A correspondent à des fragments spécifiques de l'ADN « tester », alors que les fragments non spécifiques forment des molécules de type C avec l'ADN « driver ». Les copies multiples forment quant à elles des molécules de type B. Enfin, les molécules de type D sont des séquences de l'ADN « driver » en excès, simple ou double brins et ne possédant pas d'adaptateurs. Au niveau de la deuxième hybridation, les deux groupes d'ADN « tester » issus de la première hybridation sont mélangés sans subir de dénaturation avec de l'ADN « driver » fraîchement dénaturé. Cette étape permet uniquement aux molécules simple brin de l'ADN « tester » (molécule de type A) de s'hybrider pour former des molécules de type B et C mais surtout une nouvelle conformation hybride dite de type E, représentant des fragments différentiellement exprimés entre l'ADN « tester » et « driver ». Les différentes conformations moléculaires (A, B, C, D et E) sont soumises à une réaction de PCR permettant l'amplification spécifique des séquences de l'ADN « tester » où les conformations de type A et D ne sont pas amplifiées car elles ne possèdent pas de sites moléculaires (absence de séquences complémentaires à l'adaptateur) permettant l'hybridation des amorces. Les molécules B quant à elles possèdent des séquences complémentaires en leurs extrémités, et suite à la PCR ces dernières vont former des structures stables en épingle à cheveux. Ces structures ne seront pas amplifiées préférentiellement du fait d'une hybridation plus favorable et plus stable des séquences longues des adaptateurs en comparaison avec les amorces qui elles sont beaucoup plus courtes, aboutissant ainsi à un effet dit suppressif de la PCR. Les molécules de type C, possédant quant à elles un seul site moléculaire permettant l'hybridation des amorces, ne pourront être amplifiées que linéairement. Ainsi seules les molécules de type E, normalisées et enrichies en séquences spécifiques de l'ADN « tester » et possédant deux sites d'hybridation pour les amorces, peuvent être amplifiées de manière exponentielle.

d'enrichissement pour identifier des séquences présentes au sein d'un génome bactérien et absentes d'un autre.

3.1.1.a Principe

Le principe de la soustraction reste similaire quelle que soit la méthode utilisée. L'échantillon d'ADN génomique contenant les séquences d'intérêt est appelé « tester » et l'échantillon dit de référence est appelé « driver ». L'objectif est de récupérer des fragments spécifiques du « tester ». Au niveau de l'hybridation soustractive suppressive, il s'agit d'hybrider les ADN « tester » et « driver » et d'amplifier par PCR les séquences spécifiques du « tester » correspondant aux duplexes « tester-tester ». Après PCR, le milieu réactionnel se trouve donc majoritairement enrichi en duplexes « tester-tester ». La méthode peut être schématisée et décomposée en différentes étapes (**Figure 30**) (Huang et al 2007). Même si ces techniques d'hybridation soustractives ont été initialement mises au point pour mettre en évidence un niveau d'expression différentielle des gènes ou l'identification des séquences génomiques différentes entre différents organismes (De Long et al 2007, Huang et al 2007). Ces techniques ont été le point de départ pour le développement d'autres méthodes similaires avec des applications dites méta« omiques ».

3.1.1.b Applications méta« omiques »

Ces dernières années, les techniques d'hybridation soustractives ont été appliquées pour la caractérisation des communautés microbiennes lors d'études métagénomiques et métatranscriptomiques.

i. En métagénomique

Les méthodes d'hybridation soustractives suppressives ont été premièrement utilisées dans le but d'isoler des fragments présents au niveau d'un métagénome microbien et absent d'un autre, dans une optique dite de métagénomique comparative. Les applications ont notamment ciblé le métagénome du rumen (Galbraith et al 2004) où il a pu être comparé les différences phylogénétiques et fonctionnelles des communautés microbiennes entre deux échantillons. Cette étude a également révélé une forte proportion (environ 50%) de séquences isolées suite à la soustraction, n'ayant pas de similarités dans les bases de données. Une autre étude sur le microbiote intestinal du porc (Chew and Holmes 2009) a permis, suite à la mise en place d'une méthode d'hybridation soustractive couplée à de la DGGE, de mettre en évidence des différences dans la composition de la flore intestinale chez deux porcs de 24 et 31 jours, traduisant une variation temporelle des communautés microbiennes.

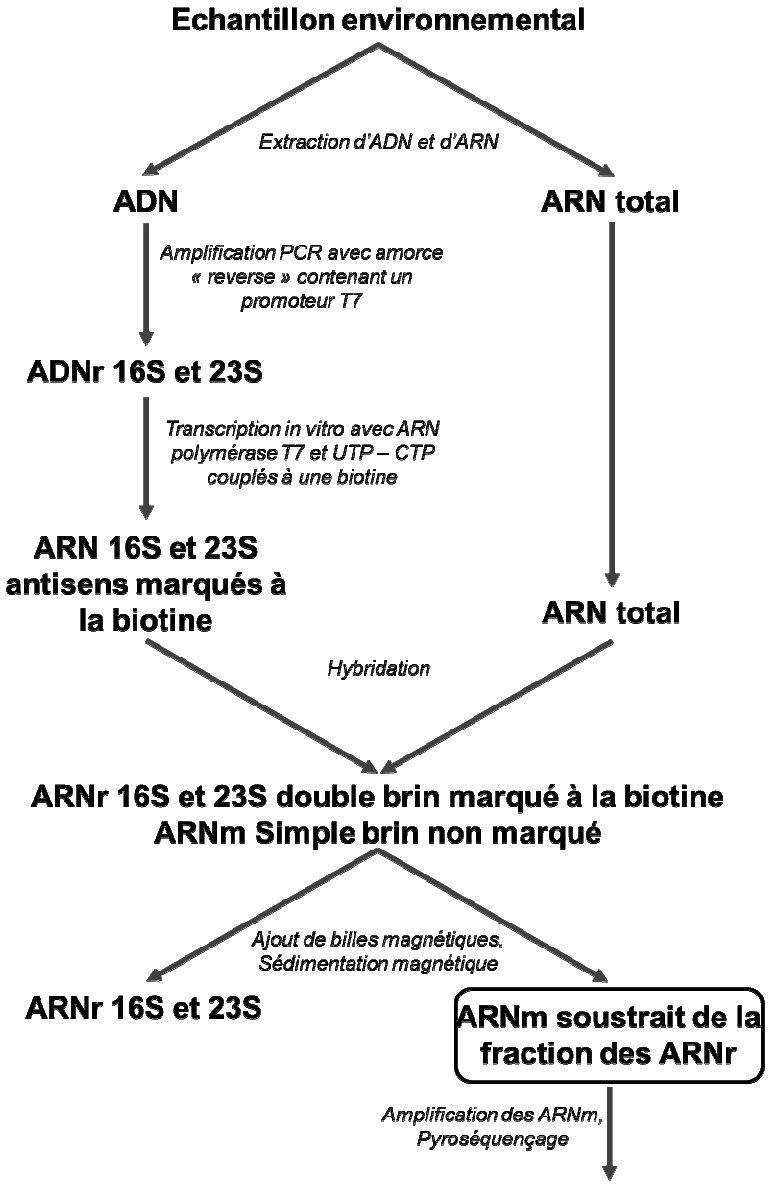

Figure 31. Représentation schématique de la méthode de soustraction des ARNr en métatranscriptomique utilisée pour des échantillons environnementaux. (Modifié d'après Stewart et al., 2010)

ii. En métatranscriptomique

Les applications en métatranscriptomique des méthodes de soustraction ont été combinées à l'utilisation de billes magnétiques. Elles ont été utilisées pour soustraire la fraction majoritaire représentée par les ARN ribosomiques retrouvés dans les échantillons d'ARN totaux, et permettre ainsi d'accéder plus facilement par séquençage à la fraction minoritaire (entre 3-5%) des ARN messagers. Ces méthodes utilisent des sondes ARN antisens ciblant les gènes codant pour l'ARNr 16S et 23S. Ces sondes sont préparées par PCR en utilisant des amorces universelles avec une amorce (reverse) contenant un promoteur T7, qui permettra d'initier une réaction de transcription *in vitro* incorporant un UTP portant une molécule de biotine. Les sondes ARN biotinylées sont hybridées avec les ARN totaux et les duplexes ainsi formés sont piégés par des billes magnétiques recouvertes de streptavidine (**Figure 31**). Ces méthodes de soustraction ont été validées sur des métatranscriptomes artificiels (He et al 2010a) ou environnementaux, issus notamment des communautés marines de bactérioplancton (Stewart et al 2010). Il est intéressant de noter que ces méthodes de soustraction, basées sur ce même principe, sont actuellement commercialisées par la société Ambion sous la forme du kit, « MICROBExpress Bacterial mRNA Enrichment kit (Ambion) ».

Même si ces techniques de soustraction démontrent un potentiel d'application très intéressant pour enrichir des échantillons environnementaux, elles comportent certains biais. Ces derniers sont principalement dus à l'utilisation prédominante de la PCR lors des méthodes de soustraction décrite ci-dessus. De plus, malgré l'élimination des séquences non désirées permettant un enrichissement de la fraction d'intérêt, la complexité de l'échantillon environnemental reste importante. Cette forte complexité limite encore l'utilisation directe des nouvelles techniques de séquençage sur les échantillons environnementaux. D'autres méthodes d'enrichissement utilisant des sondes oligonucléotidiques spécifiques de régions d'intérêt (gènes) peuvent être utilisées.

3.1.2 Les techniques de capture de gènes

Ce concept d'enrichissement par hybridation a longtemps été utilisé en génétique humaine, où une étude pionnière a permis l'enrichissement d'ARN (sous forme d'ADNc) transcrits à partir de grandes régions d'ADN génomique (Lovett et al 1991). Les auteurs ont utilisé un principe de sélection directe entre une banque d'ADNc et un YAC (Yeast Artificial

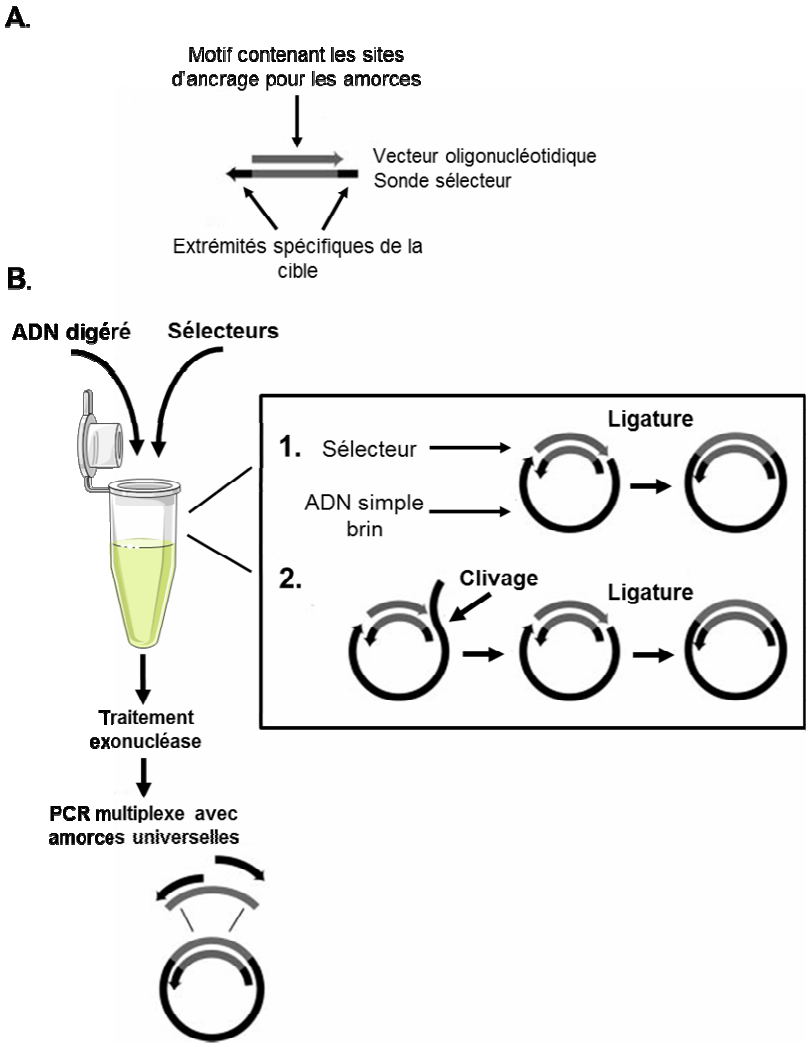

Figure 32. Principe de la capture par circularisation sélective. (A) Le sélecteur se compose de deux oligonucléotides : une sonde et un vecteur. La sonde possède deux extrémités spécifiques et complémentaires au niveau de la cible (noir) et un motif universel (gris). Le vecteur est complémentaire à ce motif et comporte des séquences d'ancrage pour des amorces PCR. **(B)** La méthode de circularisation débute avec une digestion de l'ADN qui sera par la suite mélangé avec les sélecteurs. Le mélange est dénaturé puis renaturé progressivement permettant l'hybridation des sélecteurs au niveau de leur cible. L'hybridation peut avoir lieu de deux manières différentes : (1) un sélecteur s'hybride au niveau des deux extrémités de la cible et les extrémités sont connectées entre elles par une réaction de ligature ou (2) un sélecteur s'hybride au niveau d'une extrémité 3' de la cible et l'autre au niveau d'une séquence plus en interne de la cible formant une structure en branche qui peut être clivée par l'activité endonucléasique d'une ADN Taq polymérase particulière. Les deux extrémités sont donc prêtes pour être liées à l'ADN simple brin de manière à former une molécule circulaire. Dans les deux cas de figure, le mélange est par la suite traité par une exonucléase puis amplifié par PCR en utilisant des amorces universelles. (modifié et redessiné d'après Dahl et al., 2007)

Chromosome) contenant le gène codant pour l'érythropoïétine (EPO). Ils ont pu, en un seul cycle de sélection, aboutir à un facteur d'enrichissement de 1000 en transcrits codant pour l'EPO. Cette stratégie originelle de sélection directe a été récemment appliquée au reséquençage de grandes régions génomiques à l'aide des approches NGS (Gnirke et al 2009, Hodges et al 2007). Ces méthodes reposent sur une stratégie de détermination de sondes oligonucléotidiques dites en « tilling ». Les différentes sondes, correspondant à une région d'ADN, sont déterminées de manière à ce qu'elles se chevauchent sur 1 à 10 bases pour permettre d'obtenir différents degrés de résolution. En effet, il est possible de détecter avec une précision d'une base, des régions d'un génome présentant des caractéristiques particulières. Cette stratégie de détermination permet donc la détection par exemple de Single Nucleotide Polymorphisms (SNP) ou de remaniements en s'affranchissant de la PCR. Les sondes utilisées pour ces nouvelles stratégies de capture peuvent être (1) en solution ou (2) greffées au niveau sur une surface solide (par exemple une biopuce à ADN) (Garber 2008, Good 2011, Mamanova et al 2010, Mertes et al 2011, Summerer 2009, Turner et al 2009b)

3.1.2.a La capture de gènes en solution

En dehors des techniques classiques d'enrichissement par PCR, plusieurs méthodes ont été mises au point et décrites comme utilisant une phase en solution pour l'hybridation des sondes avec leurs cibles. Deux stratégies principales ont été développées récemment. La première utilise différents types de sondes circulaires, des enzymes de restriction et des étapes d'amplification permettant d'obtenir les séquences cibles (Dahl 2005, Dahl et al 2007, Porreca et al 2007, Turner et al 2009a). La deuxième stratégie repose sur l'emploi de sondes ARN biotinylées produites de manière enzymatique, et immobilisées sur des billes magnétiques avec leurs cibles suite à l'hybridation, afin de permettre les étapes de lavage et d'élution des fragments capturés (Gnirke et al 2009).

i. Capture par circularisation

Cette stratégie de capture implique la circularisation des séquences cibles *via* l'utilisation d'une ADN ligase (Dahl 2005, Porreca et al 2007). Cette étape de circularisation a deux buts principaux : (1) lier chaque séquence cible par le biais d'un « sélecteur » qui contient les séquences permettant de sélectionner la cible et une séquence universelle utilisée pour initier l'amplification PCR avec un même et unique couple d'amorces ; (2) protéger les fragments capturés de l'action des exonucléases utilisées pour dégrader les molécules d'ADN non ciblées (**Figure 32**). Le principal avantage de cette approche est d'autoriser le

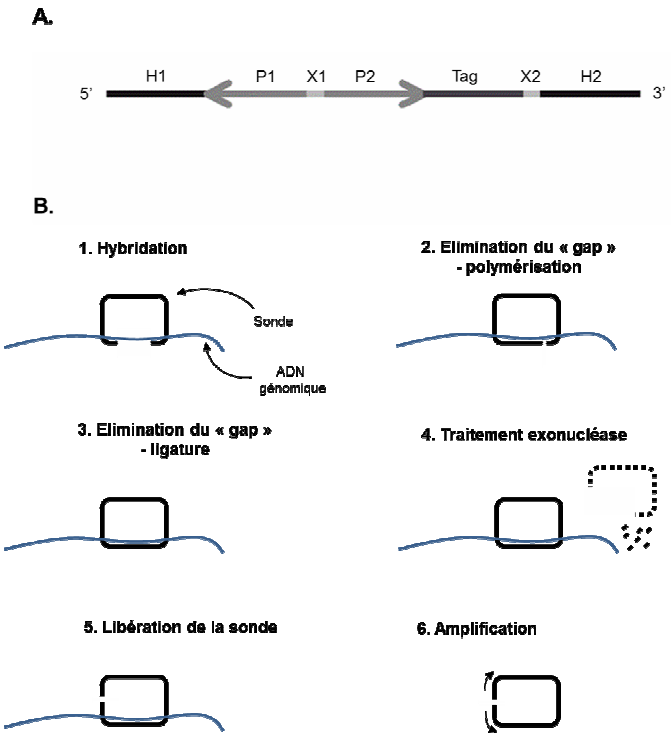

Figure 33. Principe de la capture utilisant des sondes cadenas ou MIP « Molecular Inversion probes ». (A) Nomenclature des sondes MIP. La sonde utilisée comporte sept parties ; deux régions d'homologie à la séquence cible et unique à chaque sonde (H1 et H2) au niveau des extrémités, deux régions d'ancrage pour des amorces PCR commune à toutes les sondes, une étiquette ou code-barres (tag) et deux sites de clivage X1 et X2 utilisés pour le décrochage de la sonde depuis l'ADN génomique. (B) Méthode. (1) l'ADN génomique est mélangé avec les sondes, une ligase thermostable et une polymérase, puis dénaturé et enfin renaturé permettant aux sondes de s'hybrider avec leurs cibles. (2) les dNTP sont ajoutés puis la polymérase comble le « gap » et (3) la ligase permet la jonction au niveau des extrémités de façon à former une molécule circulaire autour du brin d'ADN génomique. (4) une exonucléase est ajoutée pour digérer les sondes linéaires non circularisées. Le milieu réactionnel est par la suite chauffé de manière à inactiver les exonucléases. (5) les sondes sont libérées de l'ADN génomique par clivage enzymatique puis (6) amplifiées par PCR. (Modifié et redessiné d'après Hardenbol et al., 2003)

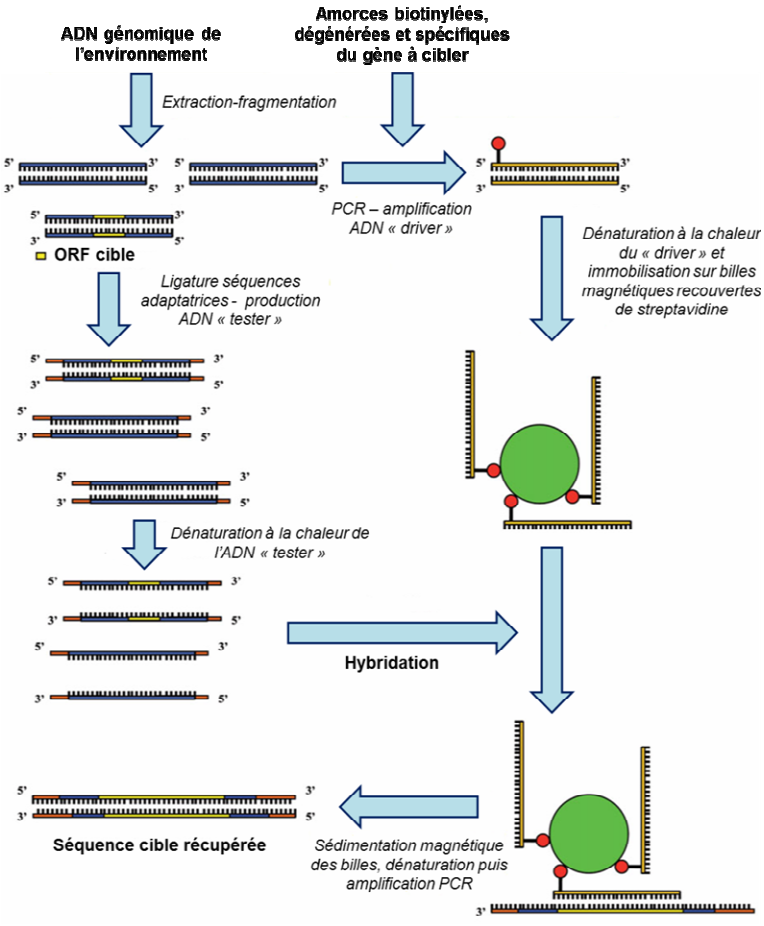

Figure 34. Représentation schématique du principe de capture au sein d'un métagénome par hybridation soustractive utilisant des billes magnétiques. (Modifié d'après Meyer et al., 2007)

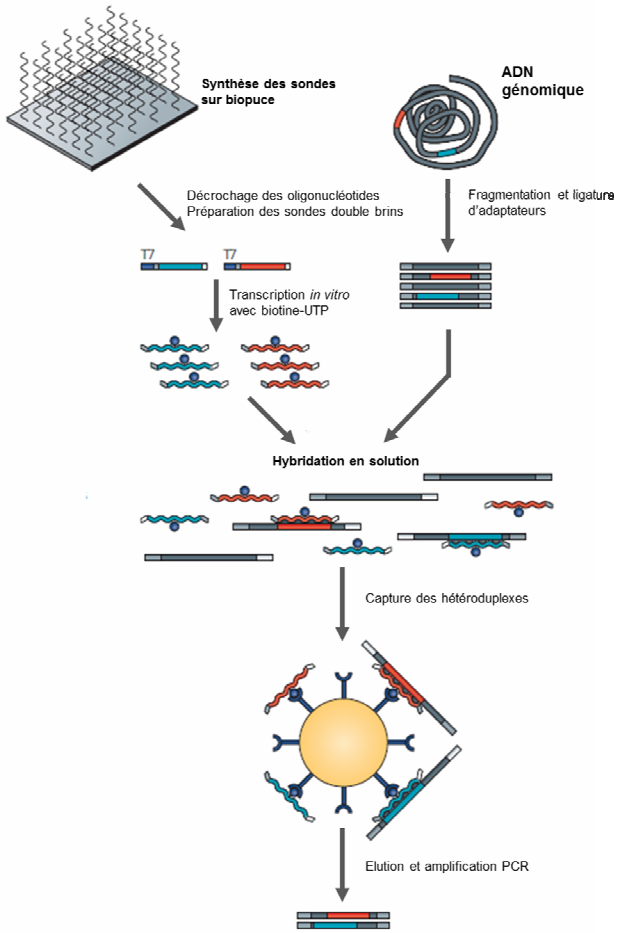

Figure 35. Principe de la capture de gènes par hybridation et sélection en phase liquide. (Modifié d'après Metzker, 2010)

multiplexage par rapport aux techniques classiques d'enrichissement comme la PCR. Une première méthode, dite de circularisation sélective, appliquée initialement en cancérologie humaine, a permis le reséquençage de 177 exons issus de dix gènes en utilisant 425 (Dahl et al 2007). Une deuxième méthode, basée sur la circularisation, utilise des sondes cadenas, ou « Molecular Inversion probes » (MIP), qui se présentent sous la forme d'oligonucléotides contenant deux séquences complémentaires à la séquence cible et séparées par une séquence « linker » (Hardenbol et al 2003, Nilsson et al 1994). Cette approche moléculaire est décrite **Figure 33**.

ii. Capture par hybridation et sélection en phase liquide

Cette méthode, couramment appelée capture par hybridation en solution, a été appliquée pour la première fois pour enrichir des séquences d'intérêt (ADN humain de Neandertal), issues d'ADN métagénomique préhistorique (Noonan et al 2006). L'objectif de cette étude était de sélectionner des séquences d'ADN de l'Homme de Neandertal au sein d'un échantillon contenant également de l'ADN microbien. Pour cela, les auteurs ont appliqué une hybridation et une sélection sur de l'ADN métagénomique préhistorique en utilisant des sondes de capture dessinées à partir d'ADN humain « comtemporain ». Plus précisément, ces sondes de capture biotinylées ont été produites par PCR et elles correspondent aux régions génomiques souhaitant être ciblées. Les sondes ont été hybridées en solution avec des amplicons provenant d'une amplification PCR préalable réalisée sur une banque métagénomique construite à partir de l'ADN préhistorique. Les duplexes ainsi formés ont ensuite été capturés par des billes magnétiques recouvertes de streptavidine. Une seule étude présente l'application d'une telle approche de capture au sein d'un échantillon métagénomique environnemental permettant l'enrichissement des gènes codant pour des « multi-copper oxidases » (**Figure 34)** (Meyer et al 2007). Même si cette étude dévoile peu d'informations quant aux résultats biologiques obtenus, elle dévoile une méthodologie séduisante pour une application à grande échelle en écologie microbienne.

Une autre méthode de capture en solution a été développée ces dernières années. Elle consiste en l'utilisation de sondes ARN longues (Gnirke et al 2009) (**Figure 35**). Cette méthode de capture en solution possède différents points forts avec (1) la présence de sondes ARN sens en excès ; (2) l'utilisation d'une quantité réduite d'ADNg cible pour l'hybridation (entre 500 ng et 3 µg) ; (3) une réaction en solution potentiellement automatisable ; (4) l'utilisation de sondes de capture longues ; (5) la possibilité de synthétiser par transcription *in*

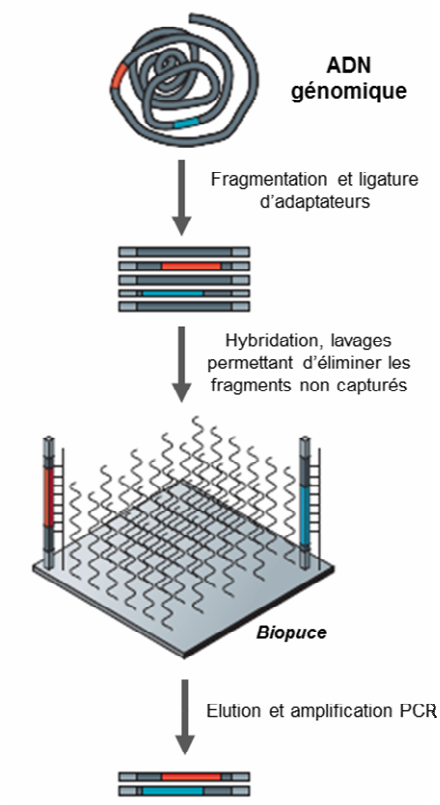

Figure 36. Principe de la capture de gènes sur support solide. (Modifié d'après Metzker, 2010).

Tableau 7. Comparaison des méthodes de capture sur support solide et en solution

	Réactions enzymatiques	Quantité de matériel	spécificité	Coût	Automatisation
Capture sur support solide	Minimales *(PCR)*	Elevée *(~15 µg)*	Bonne	Elevée *(coût des biopuces)*	Faible
Capture en solution	Elevées *(PCR, TIV)*	Moyenne *(1-3 µg)*	Très bonne *(possibilité de piéger plusieurs régions sur une même molécule)*	Moyen	Elevée

vitro de très grandes quantités de sondes (utilisables pour plusieurs capture) ; (7) la possibilité de cibler des régions génomiques courtes, longues ou discontinues et enfin (8) une grande efficacité avec 85-90% des séquences obtenues suite à la capture correspondant au régions initialement ciblées (Gnirke et al 2009, Turner et al 2009b). Il est intéressant de noter que cette approche de capture en solution a été récemment commercialisée par la société Agilent Technologies.

3.1.2.b La capture de gènes sur support solide

La capture sur support solide a été appliquée lors de nombreuses études en génétique humaine, animale et végétale (Albert et al 2011, Almomani et al 2010, Amstutz et al 2010, Antipova et al 2009, Bau et al 2008, Burbano et al 2010, Chang et al 2011, Chou et al 2009, Cosart et al 2011, D'Ascenzo et al 2009, Daiger et al 2010, Hodges et al 2009, Hong et al 2012, Jiang et al 2011, Lee et al 2009, Naumova et al 2010, Ng et al 2009, Summerer et al 2010, Ustek et al 2012, Wei et al 2011, Winfield et al 2012). Seuls deux types de support solide dédié à cette méthode de capture sont actuellement utilisés et permettent tous deux une synthèse *in situ* des sondes et une grande flexibilité au niveau des séquences ciblés. Le premier support est de type lame de verre et est classiquement utilisée pour la technologie des biopuces à ADN (Albert et al 2007, Hodges et al 2007, Okou et al 2007, Summerer et al 2009). Actuellement, les sociétés Agilent et Nimblegen (Roche) proposent des protocoles dédiés à l'utilisation des biopuces de capture, dont la mise en œuvre est beaucoup plus simple que la capture en solution (**Figure 36**). Le second type de support correspond aux biopuces en verre dites microfluidiques et compartimentées. Celles-ci contiennent différents réseaux et permettent d'effectuer différentes captures sur une même lame dans un faible volume réactionnel (Bau et al 2008, Summerer 2009, Summerer et al 2009, Summerer et al 2010).

Les nouvelles méthodes de capture décrites ci-dessus proposent des applications très intéressantes pour l'enrichissement de grandes régions génomiques d'intérêt. Même si elles présentent des performances similaires, chacune d'entre elles présente ses avantages et inconvénients (Teer et al 2010). De plus, certaines considérations techniques et financières sont à prendre en compte pour l'application de ces méthodes de capture à haut-débit (**Tableau 7**).

3.2 Conclusion

Les nouvelles approches de capture de gènes peuvent avoir de nombreuses applications très intéressantes notamment en couplant leur utilisation au séquençage haut-

débit. En effet, l'enrichissement *via* l'utilisation de sondes spécifiques, sensibles et exploratoires permettraient de limiter l'effort de séquençage pour la caractérisation des gènes d'intérêt des communautés microbiennes présentes au sein d'environnements complexes. Ces nouvelles stratégies de capture de gènes permettraient d'isoler de grands fragments d'ADN avec la possibilité d'étendre l'analyse au-delà des régions ciblées et donc d'explorer leurs régions flanquantes. En outre, ces nouvelles techniques pourraient explorer plus finement la diversité des communautés microbiennes des environnements complexes, en s'affranchissant des biais des méthodes classiques basées sur la PCR.

Conclusion générale

Depuis le début de l'ère industrielle, les activités anthropiques ont perturbé le cycle du carbone aboutissant à l'émission d'une quantité non négligeable de gaz à effet de serre dans l'atmosphère, ou à la contamination de différents écosystèmes par des hydrocarbures. Ces activités humaines engendrent sur les environnements des répercussions fortes regroupées sous le terme de changement global. Les microorganismes sont des acteurs clés au centre de ces changements globaux car ils peuvent amplifier les effets néfastes liés à ces changements ou au contraire participer à la réduction de ces effets. Cependant, les informations disponibles sur ces microorganismes restent à l'heure actuelle très incomplète du fait de l'extraordinaire diversité des communautés microbiennes. Les méthodes culturales ou encore des méthodes moléculaires générant des données parcellaires basées sur l'amplification de biomarqueurs, ont laissé la place à l'essor de nouvelles techniques d'étude globale des écosystèmes. Ainsi, le développement de la métagénomique, l'utilisation du séquençage massif, des méthodes moléculaires à haut-débit comme les biopuces à ADN ou les nouvelles techniques de capture de gènes, permettent d'explorer la diversité microbienne au sein des environnements complexes.

Ces nouvelles méthodes, en permettant d'accéder à une quantité non négligeable de données moléculaires, assurent une meilleure compréhension de la diversité phylogénétique et fonctionnelle des communautés microbiennes d'écosystèmes, et donc du fonctionnement global de ces écosystèmes. L'efficacité de ces méthodes repose essentiellement sur la qualité des sondes sélectionnées en termes de sensibilité, de spécificité et d'uniformité (Loy and Bodrossy 2006, Wagner et al 2007). Actuellement, une limite de ces techniques réside dans la détermination des sondes, en effet ces dernières sont généralement déduites des séquences d'organismes présentes dans les bases de données. La plus grande majorité de la diversité microbienne restante encore non caractérisée, celle-ci ne peut être appréhendée par de telles sondes. L'enjeu majeur actuel est de définir des sondes dites exploratoires ciblant des variants génétiques encore non référencés dans les bases de données (Dugat-Bony et al 2012b). Il est donc nécessaire de faire évoluer les stratégies de détermination des sondes en intégrant ce concept exploratoire. De plus, il est primordial de prendre en compte l'apport exponentiel des séquences dans les bases de données, qui de par son importance nécessite aussi de repenser les algorithmes de détermination des sondes pour réduire les temps de calcul. Les différentes approches de détermination de sondes feront donc l'objet d'un chapitre complet de ce

manuscrit (PARTIE 2 : Outils logiciels pour la sélection de sondes oligonucléotidiques). Par la suite, un nouveau logiciel de détermination de sondes sera présenté (PARTIE 3 : Développement d'un logiciel de sélection de sondes oligonucléotidiques). Finalement, *via* l'utilisation de sondes spécifiques de gènes d'intérêt, mais sans a priori sur les variants de séquences, une nouvelle méthode de capture dédiée à l'écologie microbienne sera présentée (PARTIE 4 : Développement d'une méthode innovante de capture de gènes en solution couplée a du séquençage haut-débit pour l'exploration métagénomique ciblée des environnements complexes).

PARTIE 2 : Outils logiciels pour la sélection de sondes oligonucléotidiques

1. Contexte

A l'heure de la révolution des techniques moléculaires, les biopuces à ADN se présentent comme des outils à haut-débit de choix en écologie microbienne (Cook and Sayler 2003, Sessitsch et al 2006, Wagner et al 2007, Zhou and Thompson 2002). Différents types de biopuces sont disponibles et ces dernières apportent des informations sur la structure des génomes, leurs parentés, les expressions géniques, les capacités métaboliques ainsi que sur la structure et la dynamique des communautés microbiennes. Les biopuces permettent donc de caractériser, grâce à des signatures moléculaires spécifiques, la présence de milliers de microorganismes mais également les potentialités métaboliques présentes dans un échantillon environnemental, et ceci au cours d'une même expérience. Cette caractérisation à haut-débit repose principalement sur l'utilisation de sondes oligonucléotidiques du fait du faible coût de synthèse, de la flexibilité de fabrication des biopuces *in situ* et des masses de données disponibles issues du séquençage. Même si les sondes sélectionnées doivent toutes présenter des caractéristiques semblables en termes de sensibilité, spécificité et uniformité, cette détermination reste le point essentiel de l'approche biopuces.

2. Objectifs

L'évolution constante des puissances de calcul et des infrastructures informatiques, de même que le nombre de séquences disponibles dans les bases de données, offrent de nouvelles perspectives pour les stratégies de détermination de sondes oligonucléotidiques dédiées aux problématiques d'écologie microbienne. De plus, l'augmentation des connaissances, bien qu'encore incomplètes sur les hybridations sur phase solide, permet d'optimiser la qualité de ces sondes. De manière à proposer une vue d'ensemble des stratégies bioinformatiques disponibles pour la détermination de sondes oligonucléotidiques, la rédaction d'un chapitre d'un ouvrage prochainement disponible et intitulé « Microarrays: Current Technology, Innovations and Applications », a été réalisé sous la direction du Docteur Zhili He de l'Université d'Oklahoma (USA).

Chapitre livre : Software tools for selection of oligonucleotide probes for microarrays

Nicolas Parisot[1,2*], Jérémie Denonfoux[1,2*], Eric Dugat-Bony[1,2], Eric Peyretaillade[1,2] and Pierre Peyret[1,2]

[1]Clermont Université, Université d'Auvergne, Centre de Recherche en Nutrition Humaine Auvergne, EA 4678, Conception, Ingénierie et Développement de l'Aliment et du Médicament, BP 10448, F63000 Clermont-Ferrand, France

[2] Clermont Université, Université d'Auvergne, UFR Pharmacie, Clermont-Ferrand, France

[*] These authors contributed equally to this work.

Nicolas Parisot: nicolas.parisot@udamail.fr

Jérémie Denonfoux: jdenonfoux@yahoo.fr

Eric Dugat-Bony: eric.dugat@univ-bpclermont.fr

Eric Peyretaillade: eric.peyretaillade@udamail.fr

Pierre Peyret: pierre.peyret@udamail.fr

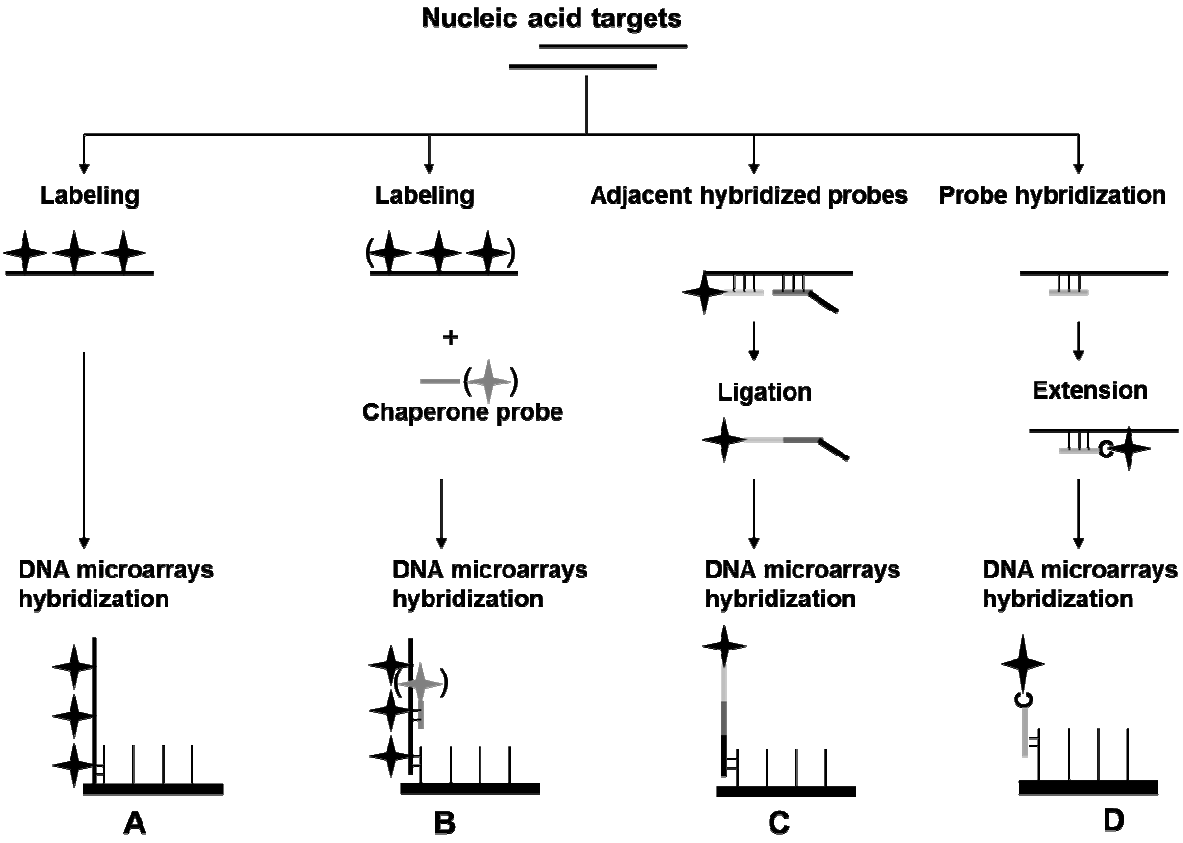

Figure 1. Schematic representation of different oligonucleotide-based approaches used in DNA microarray technology. (A) The most widely used approach allows the direct recognition of labelled targets by specific probes fixed on a solid surface. (B) The two-probe proximal chaperone detection system uses helper probes (chaperones) to resolve the secondary or tertiary structure of the targets, thereby improving accessibility for efficient matching between the specific probe located on the solid surface of the DNA microarrays and the target. Labelling could be directed towards the targets (black) or the chaperones (grey). (C) Enzymatic ligation uses a high-selectivity ligase, which requires the perfect complementarity of the double-stranded DNA structure to successfully catalyse the covalent joining of two adjacently hybridised probes. Detection is only possible if the probes are linked together. In the universal microarray approach, the common probe has a tag sequence (cZip code in black) that directs the hybridisation on the capture probe (Zip) on the solid surface of the DNA microarrays. (D) The single nucleotide extension labelling allows the reverse complement probe to be labelled in a linear amplification reaction given the availability of the corresponding target sequence. The labelled reverse complement probe is then captured by specific probes located on the solid surface of the DNA microarrays.

Abstract

Oligonucleotide microarrays have been widely used for gene detection and quantification of gene expression. Recently, they have been adapted for profiling microbial communities in a flexible and easy-to-use manner. In fact, it is possible to analyse both the microbial diversity and the metabolic capacity of complex communities in one experiment. However, the quality of the result is largely dependent on the quality of designed probes. Probe design, which is not a trivial task, should thus take into account multiple parameters such as the oligonucleotide sequence and its binding capacity in order to ensure high specificity, sensitivity, and uniformity as well as potentially quantitative capability for each probe. Furthermore, the exploration of the not-yet-described fraction of complex communities requires consideration of the explorative power of oligonucleotide probes. To design such probes, multiple tools have been developed based on different algorithms. These algorithms and the different probe criteria that they used are described in the present chapter. However, the best algorithm to guarantee a high-quality design must be chosen with the knowledge of biological questions and biological samples.

Introduction

With exponential growth in the availability of complete genome sequences and metagenomic data sets and the low cost of DNA synthesis, oligonucleotide arrays have become the most widely used type of microarrays. Furthermore, with the advancement of microarray technology (e.g., *in situ* synthesis technologies), high-density oligonucleotide microarrays can hold millions of probes on a single microscopic glass slide with multiplexing capacities. These molecular tools can be easily synthesised on demand, in small batches, and at low cost. This flexibility combined with rapid data acquisition, management and interpretation allow oligonucleotide microarrays to continue to advance next-generation sequencing in various applications. Several strategies using oligonucleotides probes have been developed to improve the specificity and sensitivity of gene detection **(Figure 1).** The most widely used strategy **(Figure 1A)** is based on the determination of specific sub-sequences in the targeted genes that serve as probes. The subsequent steps involve the hybridisation of the labelled targets,

followed by the image processing. Other supplementary steps are added to improve the specificity and sensitivity of detection.

The capture of targets is strongly influenced by their secondary and tertiary structures; therefore, probes should be directed toward accessible regions. However, measuring or predicting the effect of secondary structure is still difficult. The shearing of target molecules into small fragments is one widely utilised technique. Alternatively, to overcome secondary structure constraints, a two-probe proximal chaperone detection system **(Figure 1B)** that consists of a species-specific capture probe and a chaperone probe (sometimes also used as a labelled detector) that reduces secondary structure formation was developed (140)(Small et al 2001). The term “chaperone probe” has been used rather than the term “stacking probe” (169), which was originally used in the context of polymorphism detection (95). However, chaperone detector probes located in the immediate proximity of the capture probe provide detectable, non-specific, non-target binding, presumably because of base-stacking effects (21). In some cases, the addition of specific DNA helper oligonucleotides improved detection (73). However, the use of helper oligonucleotides is not always practical because of the difficulty of designing helper probes with the same specificity as the capture probe but without non-target detection (114).

Enzymatic ligation **(Figure 1C)** is another microarray-based method that has also been used for the detection of environmental microorganisms (17-19, 48, 64). The reaction is performed separately from array hybridisation, which enables the use of address (also known as tag or zip) oligonucleotides to equalise probe hybridisation conditions. The enzymatic ligation step is the primary source of specificity. The principle of detecting specific DNA templates by enzymatic ligation was developed to overcome some of the limitations of oligomeric hybridisation probes in distinguishing single base mutations associated with genetic diseases. Enzymatic ligation relies on the high selectivity of the ligase, which requires the perfect complementarity of a double-stranded DNA structure to successfully catalyse the covalent joining of two adjacently hybridised probes. The probes constitute a target-specific probe pair that becomes detectable only if the probes are linked together. The so-called discriminating probe is designed such that the 3' end matches the target at a unique position containing a nucleotide that distinguishes the target from other species. A common probe is designed to hybridise adjacent to the discriminating probe, which enables ligation if an

appropriate target is present in the reaction mixture. In the universal microarray approach (48), the common probe has a 3'-tag sequence (cZip code) that directs it to the correct address on the array, whereas the discriminating probe is fluorescently labelled.

The advantages of the universal array lie in the uniform hybridisation conditions of all zip sequences and in flexibility, as the same array platform can be used with multiple ligation probe sets. ORMA (Oligonucleotide Retrieving for Molecular Applications) is a set of scripts for searching discriminating positions and selecting oligonucleotide probes for such an approach (Ligase Detection Reaction; LDR) or for Minisequencing/Primer Extension (139)(Severgnini et al 2009). A variant strategy that utilises a cleavable padlock probe has recently been developed (156) that eliminates the probe amplification of the initial padlock probe assay (146), resulting in a background-free assay. Padlock probes are long oligonucleotides that contain asymmetric target complementary regions at both their 5’ and 3’ ends to confer specific target detection. Upon hybridisation to the target, the two ends are brought into contact, which allows probe circularisation by ligation. In the first assay after exonuclease treatment, the circularised probes are amplified and hybridised on DNA microarrays. The central part of the probe harbours sequences for PCR amplification and DNA microarray capture. In a recent improvement to the method, in addition to the sequence complementary to the probe on DNA microarrays, padlock probes now harbour a cleavage site in their central part near the labelling position. After cleavage, only the originally ligated padlock probes can be visualised on the DNA microarray.

Finally, a microbial diagnostic microarray approach using single nucleotide extension labelling (SSELO: sequence-specific end labelling of oligonucleotides) has been developed (77). Reverse complements of the capture oligonucleotides (RC oligonucleotides) are end-labelled in a linear amplification reaction based on the availability of the corresponding target sequence **(Figure 1D)**. The entire mixture is hybridised to the microarray to identify the sequences that have been labelled. The specificity of the assay was shown to be determined primarily by the stringency of the annealing step during labelling rather than that of the subsequent hybridisation.

Regardless of the strategy used to develop DNA microarrays, probe selection remains the key element in obtaining an efficient detection tool. Several criteria related to probe

characteristics influence the efficiency of detection and should be assessed with caution before fabricating DNA microarrays.

General criteria for probe design

Specificity is defined according to the ability of the probe to not cross-hybridise with non-target sequences (*i.e.*, probes should discriminate well between the intended target and all other sequences present in the target pool). Sensitivity is defined as the strength with which a probe binds to its target. This parameter influences the level of the detection signal, and consequently, the relevance of obtained information (*i.e.*, probes should detect differences in target concentrations under given hybridisation conditions). Uniformity corresponds to the similarity of hybridisation behaviour for a given probe set, *i.e.*, similar thermodynamic characteristics under the same experimental conditions (*e.g.*, temperature, salt and formamide concentration), which could also influence sensitivity to some extent. In fact, the structural properties of several probes, including probe length, GC content, melting temperature (*Tm*) and the Gibbs free energy reflecting binding capacities (ΔG), are optimised in this case.

***2.1** Specificity*

The ability to minimize or eliminate cross-hybridisation is an important parameter and represents a current bottleneck in the design of microarray probes (162). In fact, the specificity of the hybridisation of a probe with its target is one of the most important parameters that determines the quality of the microarray result (71, 76). Specificity is defined as the ability of a probe to bind to a target sequence without hybridisation to non-targets. Currently, most probe design software uses the BLAST algorithm (2) to search for potential cross-hybridisation against custom databases constructed in concordance with microarray experiments and applications. Probe specificity assessment with BLAST uses a homology threshold that determines whether the oligonucleotide is specific. Kane's recommendations for long oligonucleotides are based on the discarding of probes that share a total identity greater than 75–80% or contiguous stretches of identity greater than 15 nucleotides with a non-target sequence (71)(Kane et al 2000). Alternatively, some probe design software uses a suffix array approach to overcome BLAST's limitations (96). Rather than performing several local alignments, the suffix array method utilises an efficient and space-saving data structure

that quickly identifies and records, in alphabetical order, all possible substrings or suffixes and their locations in the input sequences. The theory of suffix arrays states that the longest common prefix (LCP) shared by any two non-adjacent suffixes must be equal to or shorter than the LCP of any two neighbouring suffixes between them in the suffix array (23, 96). The main limitation of this approach is the memory storage of the suffix structure. For example, the human genome, which has 3 billion characters, requires 12 GB for storage of the entire suffix array (131). Thermodynamic calculations are also used to evaluate the strength of cross-hybridisations by determining the binding-free energy between the probe and the non-target sequence to give an indication of the duplex's stability. As the probe is bound to a solid surface rather than being free in solution, the calculation appears as an approximation (116). Finally, other probe design software uses custom methodologies to evaluate probe specificity based, for instance, on global alignments or hierarchical clustering approaches (80). Even with the use of BLAST or suffix array tools combined with thermodynamic prediction, several other criteria must be considered during the design process to improve probe specificity.

Low-complexity regions such as those containing long homopolymers may also contribute to probe specificity and consequently must be avoided during the probe design process (81, 161). To overcome this problem, many probe design algorithms, such as CommOligo (87), ROSO (123) or HiSpOD (39), apply a filter or mask these particular nucleotide repeats, whereas YODA (107) can discard specific regions defined beforehand as prohibited for the probe design. These particular regions can also be highlighted by more complex calculations using a lossless compression algorithm such as the LZW compression algorithm (172)(Ziv and Lempel 1977), a suffix array structure or custom calculations for complexity scoring. Otherwise, low-complexity regions can be masked using the DUST programme (52) included in the software that uses the BLAST algorithm for the assessment of potential cross-hybridisation.

Just as it can enhance probe sensitivity, the probe's position on the sequence could also influence the oligonucleotide specificity (155)(Tomiuk and Hofmann 2001). For example, the 3' untranslated region (3'UTR) in eukaryotic mRNA is considered the less-conserved region because of the usage of alternative polyadenylation signals (155). Consequently, the choice of 3'UTRs for probe design reduces the probability of cross-

hybridisation with closely related paralogs. However, the potential alternative polyadenylation signals found in 3'UTR combined with a propensity for repetitive elements has to be taken in consideration. Thus, some programmes compute a localisation score based on the distance to the centre or to the 3' or 5' end of the sequence (*e.g.*, OligoWiz (163)) or let the user localise the designed probes in a 3' or 5' range (*e.g.*, OligoPicker (161)). Others can display (*e.g.*, YODA (107)) all of the non-overlapping probes or only those located in the 3' end, 5' end or in the centre of the sequence.

2.2 Sensitivity

The term sensitivity is closely related to the affinity of a probe to its target, which is mediated by hybridisation and is characterised by the free energy difference ΔG that measures the binding affinity for the two strands to form a duplex. ΔG can be estimated from the probe sequence using nearest neighbour models that provide a reasonable approximation of ΔG for strands hybridising in solution (133). Furthermore, in microarray experiments where quantitative detection is required, microarray probes should also exhibit a sensitive and predictable response to concentrations of specific targets (99). Although in-solution parameters, e.g., base composition, temperature and salt concentration, are typically used for such calculations, the estimated *Tm* of the nucleotide duplex is a good proxy for the sensitivity of the probe to some extent. Nevertheless, even though the thermodynamic properties of nucleic acid duplex formation and dissociation in solution are well known (133), the thermodynamic properties during hybridisation at the solid-liquid interface in a microarray context remain unclear (116). Thus, several parameters have to be considered to increase probe sensitivity and allow microarray probes to exhibit a sensitive and predictable response to a target concentration.

Although the secondary structure must be considered as the main sensitivity criterion for microarray design, the probe length, number of probes per target and probe position can also be considered. The choice of which criteria to use will typically depend on the probe design strategy and microarray synthesis technology.

2.2.1 Secondary structure

To achieve maximum probe sensitivity, the design must exclude oligonucleotides that are able to form homo-dimers or stable intra-molecular secondary structures such as hairpins or stem–loops that may impact hybridisation efficiency by preventing stable target hybridisation (80). Thus, the objective is to prevent the formation of any such structures at the hybridisation temperature by assessing the secondary structure. Some probe design software uses alignment-based strategies for a self-annealing assessment combined with scoring calculations (75), thermodynamic calculations based on the Mfold tool (173) or suffix array data (96) to assess secondary structure stability in combination with a specificity test to evaluate potential cross-hybridisation.

2.2.2 Probe length

Probe sensitivity generally increases with probe length, as the binding energy for longer probe-target duplexes is typically higher and hybridisation kinetics are irreversible (40, 63, 83, 122). Long oligonucleotide probes (50-60 mers) have a comparable sensitivity to PCR-based probes with a length of 300-400 nucleotides. Fifty-mer probes demonstrate good specificity as long as the similarity with non-targeted sequences is less than 75% or there is no stretch of 15 perfectly matching nucleotides (71). The use of 60-mer oligonucleotide probes for hybridisation could allow the detection of targets with eight-fold higher sensitivity than the use of 25-mer probes (23), whereas an identity of less than 77% between a 60-mer probe and its target results in a lack of signal (63)(Hughes et al 2001). Generally, the threshold for differentiation between targets is 75-90% (40, 71, 148, 154) identity for such long probes, which indicates low specificity (87). In contrast, short oligonucleotide (18-30-mer) probes are more specific, as they allow the discrimination of single nucleotide polymorphisms under optimal hybridisation conditions but with reduced sensitivity (122). The GoArrays strategy (124)(Rimour et al 2005) combines both advantages by designing long probes (high sensitivity) composed of two short sub-sequences (high specificity).

2.2.3 Number of probes per target

As noted above, longer oligonucleotides provide higher sensitivity than shorter probes (40, 63, 122). However, the use of one probe per gene with long oligonucleotide microarrays appears limiting even though oligonucleotide hybridisation is highly sequence-dependent (23, 153). In fact, the binding of an oligonucleotide probe to different regions of the target yields different signal intensities (63, 138) and thus complicates the prediction of whether an oligonucleotide probe will bind efficiently to its target and yield a good hybridisation signal based on sequence information alone (23). Thus, multiple probes per gene have been used in oligonucleotide array designs to obtain reliable quantitative information for gene expression (63, 138) as well as gene detection in complex environmental samples (38, 39). Five probes per gene has been suggested as a suitable number for 30-mer probes (122), but this number could increase with probe length. A perfect case would be to select a minimal probe set that ensures good hybridisation signals and test it experimentally to successfully detect targets even at low levels, but such a large-scale screening process remains extremely time-consuming and costly.

2.2.4 Probe position

The positioning of the probe along the target may also impact the hybridisation signal, especially in gene expression experiments (162). The signal may decrease near the 5’ end when using poly-T-primed cDNA synthesis, which requires a multiple-probe design along the length of the transcripts. The decreased signal is a consequence of the stability of the RNA (4) and enzymatic reactions during sample preparation such as reverse transcription that have a tendency to terminate early (162). Thus, the probes used for gene expression are preferentially positioned near the 3’ end of eukaryotic transcripts. In contrast, for cDNA synthesis using random priming in prokaryote gene expression experiments, decreases in signal can be observed for probes positioned at the very end of the 3’ end of the gene (162).

2.3 Uniformity

Microarray technology relies on the simultaneous hybridisation of many probes under the same conditions (e.g., salt concentrations, temperature); therefore, uniform thermodynamic

behaviour for the selected probes is crucial (40, 91, 160). The easiest way to reach this objective is to select probes with homogeneous probe lengths, but several other parameters must be considered, such as the melting temperature (*Tm*) and GC content.

2.3.1 Tm uniformity

To achieve maximum homogeneity in the probe set, a primary objective is to select probes that share similar melting temperatures (*Tm*), which ensures quantitative comparison of gene expression and detection as well as similar microarray hybridisation for all genes targeted in the study. Chemical compounds such as tetra-alkyl ammonium salts (66), which have been applied in dot-blot experiments with degenerate oligonucleotide hybridisation probes (165), are known to eliminate the dependence of *Tm* on base composition. However, these salts have not been widely applied in microarray experiments. Thus, an alternative is to select oligonucleotide probes with melting temperatures that fall within a narrow range. Several methods are available to calculate the *Tm* of a probe; the most frequently used is the application of the nearest-neighbour (NN) model using parameters from SantaLucia (133) or from Rychlik *et al*. (129)(Rychlik et al 1990). The *Tm* can be calculated directly by the probe design software, by an external programme or by using a custom method. Most probe design software, for example CommOligo (87) and HiSpOD (39), allow the user to select a *Tm* range in which the selected probes will be designed. Some, such as OligoArray (128), OligoPicker (161), PICKY (24) and YODA (107), perform optimisation calculations by adapting some parameters, such as the probe length, to select probes in the expected *Tm* range. For some programmes, such as ArrayOligoSelector (13) and ProbeSelect (84), *Tm* is not considered and selection is based solely on the similar *Tm* values of probes with uniform lengths and GC content. Finally, all of the available formula calculate the *Tm* for oligonucleotides that are free in solution, and not oligonucleotide probes bound to a glass surface. However, the probe's behaviour in solution could be different from that when attached to a slide, and thus, it is more suitable that probes fall into a *Tm* range rather than having a precise *Tm* value.

2.3.2 GC content

Table 1. Comparison of probe design software features for phylogenetic oligonucleotide arrays (POAs).

Software	Reference	Application	Availability	URL
ARB (v 5.3)	(94)	POA	Downloadable, standalone GUI (L, M)	http://www.arb-home.de/
CaSSiS (v 0.5.0)	(5)	POA	Downloadable, command-line (L)	http://cassis.in.tum.de
PhylArray	(101)	POA	Web Interface	http://g2im.u-clermont1.fr/serimour/phylarray
ORMA	(139)	POA, FGA	Matlab Script	Upon request
KASpOD	(112)	POA, FGA, WGA-ORF	Web interface or command-line (L)	http://g2im.u-clermont1.fr/kaspod/

POA: phylogenetic oligonucleotide array. FGA: functional gene array. WGA-ORF: open reading-frame oriented whole-genome array. GUI: graphical user interface. L: Linux. M: MacOS.

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
ARB (v 5.3)	Fixed by the user (10-100)	No localisation specified	All probes reaching selection criteria	No	No	Yes	Yes	No	No
CaSSiS (v 0.5.0)	Fixed by the user or range chosen by the user	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	No	Yes	Yes	No	No
PhylArray	Fixed by the user (20-70)	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	No	Yes	Yes	No	Yes
ORMA	Fixed by the user	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	Yes	No	Yes	No	Yes
KASpOD	Fixed by the user (18-31)	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	No	No	No	No	Yes

Software	Organism	Cross-hybridisation assessment	Database for specificity test	Input files
ARB (v 5.3)	No limitation	Local alignment and thermodynamic calculations	ARB-Silva database	ARB database. Nucleotide sequences.
CaSSiS (v 0.5.0)	No limitation	ARB Positional Tree server and distance calculations	Input sequence dataset	FASTA file with all target and non-target sequences and a list or a tree file containing targeted sequence identifiers. Nucleotide sequences.
PhylArray	Prokaryotes	BLAST and Kane's specifications	Custom non-redundant SSU rRNA database (95MB)	No input files are required.
ORMA	No limitation	No	No	Multiple sequence alignment file (Clustal-like, Multi Sequence Files, or aligned FASTA format). Nucleotide sequences.
KASpOD	No limitation	Global alignment and distance calculations	External FASTA file	Two FASTA files (targeted sequences and non-target sequences). Nucleotide sequences.

The oligonucleotide GC content is another parameter to consider and is closely related to the melting temperature (*Tm*). Some probe design software, such as OligoWiz (163), OligoPicker (161) or ProbeSelect (84), does not consider the GC content as a potential criterion for the oligonucleotide probe selection process. In contrast, other software such as CommOligo (87), OligoArray (128) or YODA (107) allow the user to select a GC content range and filter candidate probes that do not fulfil this range from the final probe list. Generally, the programmes use a preferential range between 40-65% (80), and some are able to perform a *Tm* optimisation by using the GC content range defined by the user to select the best probe candidates. This strategy appears useful for oligonucleotide probe design that involves sequences with very high or very low GC content.

Probe design algorithms for microbial DNA microarrays

3.1 Phylogenetic Oligonucleotide Arrays (POAs)

To rapidly characterise the members of microbial communities present in complex environments, numerous phylogenetic oligonucleotide arrays (POAs) have been developed using the SSU rRNA biomarker (14, 15, 33, 54, 92, 110, 164). Fully automated software and manual approaches have both been developed to design POAs **(Table 1)**.

3.1.1 Alignment-based strategies

Initially, probe design software for POAs was primarily based on aligned sequence sets such as PRIMROSE (3), PROBE (117), ARB-Probe Design (94), PhylArray (101) and ORMA (139). Probe design software programmes based on aligned input data or on performing a multiple sequence alignment as the first step of the algorithm is well suited for the design of probes with an optimal coverage of the target group. Multiple sequence alignments are generally converted into consensus sequences that account for the sequence variability at each position. Then, probe design programmes search for conserved regions to select oligonucleotides.

The Probe Design tool included in the ARB programme package (94) has been widely used to develop low-density, custom-made POAs for reduced groups of organisms (43, 93, 106). The ARB Probe Design is able to design oligonucleotides with a length of 10 to 100 nucleotides using a three-step algorithm. First, the user selects the target group through the ARB interface. Second, the programme searches for potential target sites (avoiding repetitive regions) and subsequently returns a ranked list of candidate oligonucleotides according to several compositional and thermodynamic criteria. Finally, the proposed oligonucleotide probes are evaluated against the entire database using the Probe Match tool. Local alignments are determined between the probe and the most similar sequences in the database, and up to 5 mismatches are allowed.

Among the alignment-based oligonucleotide design software programmes for POA, ORMA (Oligonucleotide Retrieving for Molecular Applications) appears to be suitable for the determination of discriminating positions within a set of highly similar sequences (139). This software is well adapted for the ligation or extension strategies described in the introduction to this chapter. ORMA relies on a Single Base Seeker (SBS) algorithm to locate positions that are able to discriminate one sequence from a set of closely related sequences. First, the user selects the sequences that are to be considered as targeted from among the dataset; the remaining sequences are subsequently used as the group from which the discriminating positions must be different. Then, for each non-degenerate position, the SBS algorithm calculates the sum of sequences carrying the same base as the considered sequence. If the only sequence harbouring this base is the targeted sequence, the position is identified as discriminant. This last step is reiterated, replacing each degenerate position (except for undetermined and subsequently non-discriminant positions referred to as N's) with its two or three alternative bases. Candidate oligonucleotides are then defined at these discriminating positions by retrieving flanking sequences. A series of constraints and quality filters is used to assign a quality score to each putative probe (*i.e.*, length, melting temperature, number of degenerate bases, low-complexity regions). Moreover, intra-group (*i.e.*, coverage) and inter-group (*i.e.*, specificity) scores are calculated, and the probes that maximise the intra-group score and have the lowest inter-group score are selected.

The design strategies described above are not solely dedicated to high-density microarrays (*i.e.*, those with tens of thousands of oligonucleotide probes). High-density POAs

are, however, the most promising approach to comprehensive screening all known bacterial and archaeal taxa with a single microarray (40). Many strategies for designing large probe sets are not fully automated and thus not provided in the form of autonomous software. For instance, the PhyloChip (33), which is the most widely used high-density POA, was constructed using a semi-automated procedure explained in the supplementary material of Hazen *et al.* (54). All known 16S rRNA sequences containing at least 1,300 nucleotides were extracted from the NAST multiple sequence alignment (34) of the Greengenes (35) database. Then, sequences were filtered to remove putative chimeras using the Bellerophon software (62) and also to remove low-complexity sequences (*i.e.*, sequences with more than three homopolymers with a length greater or equal to 8) and sequences with ambiguous nucleotides (*i.e.*, sequences with ambiguous base calls greater than or equal to 0.3%). Retained 16S rRNA sequences were then clustered at 0.5% sequence divergence in 59,959 operational taxonomic units (OTUs). The 59,959 OTUs represented 1,464 families, 1,219 orders, 1,123 classes, 147 phyla and 2 domains. For each OTU, each of the sequences within the OTU was separated into overlapping 25-mers segments, and these potential targets were used to select the probe set. Candidate 25-mer oligonucleotides were selected from the sub-alignment according to thermodynamic constraints (*i.e.*, GC content, secondary structure, melting temperature, and self-dimerisation). Potential targets were ranked according to their universality among the OTU; those having data for all members of the OTU were preferred over those found in only a fraction of the OTU members. Candidate probes that matched exactly with well-ranked putative targets were selected for microarray fabrication.

Computational alignment of a large multiple sequences is a time-consuming task. Thus, to accelerate the computations, the probe design software tools have to be retooled to permit the computation of many probes based on large sequence datasets.

3.1.2 Alignment-free strategies

CaSSiS (Comprehensive and Sensitive Signature Search) was developed to address the limited ability of previous probe design software to handle large collections of sequences (5). CaSSiS is able to perform fast and comprehensive probe design based on a three-step algorithm. First, CaSSiS extracts and assesses each possible probe. The results are stored in a

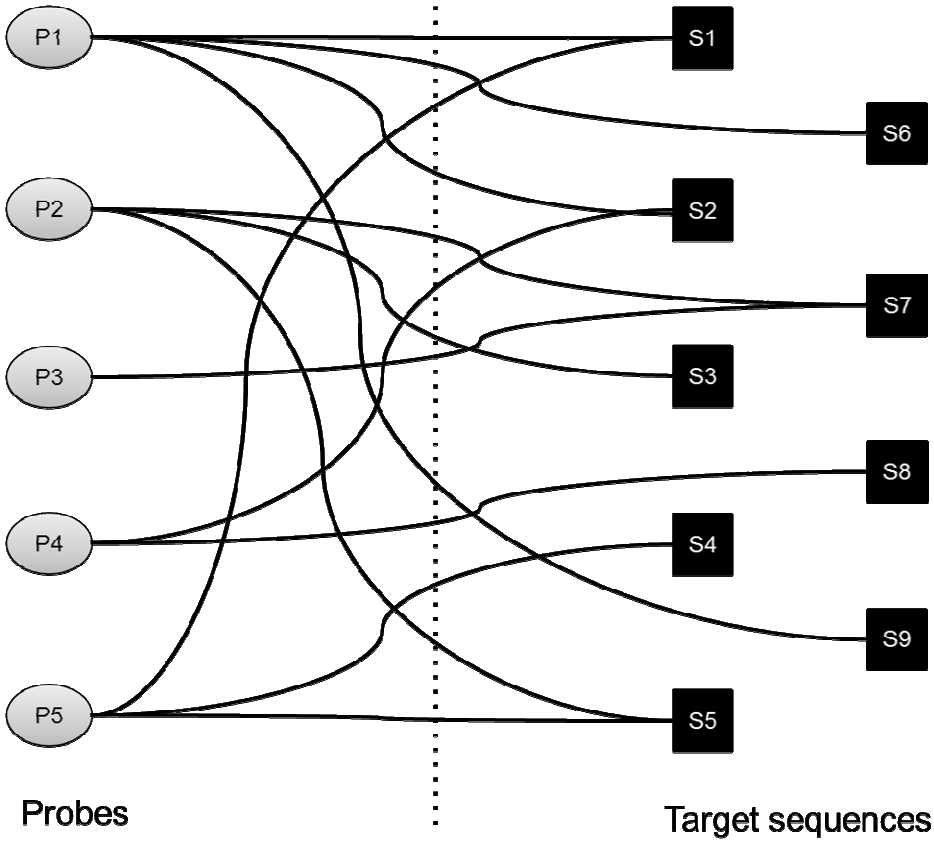

Figure 2. A Bipartite graph. This data structure provides a representation of the probes' coverage. Nodes represent a probe (P) or a sequence (S), and edges indicate which sequences are matched with which probe.

bipartite graph where the probes' coverage within the overall dataset is represented as edges **(Figure 2)**. Evaluating all of the probes could be a time-consuming task, but CaSSiS uses the ARB Positional Tree Server (PT-Server) (94) to rapidly identify exact and inexact matches. Based on predefined parameters such as length or the number of mismatches allowed, the PT-Server, using a truncated suffix tree, returns all matches of the query probe. Because CaSSiS supports a relaxed search within the database, the user can specify the number of mismatches allowed within the targeted sequences and the mismatch threshold for non-target hits. The second stage of the CaSSiS algorithm consists of ranking candidate oligonucleotides according to their specificity scores. The last step extracts the probes that harbour the highest coverage and have up to *n* non-target matches (outgroup hits), where *n* is user-defined.

Even if the probe design software for POAs was able to handle millions of sequences, its capabilities would always be restricted to surveying known microorganisms with sequences that have been deposited in a public database. However, in spite of the high number of recorded sequences, our current vision of microbial diversity is still incomplete, partially because of the tremendous diversity of microbial species, ecological niches and technological limits. Detection of 90% of the richness in some complex environments could require tens of thousands of times the current sequencing effort (120). A major challenge, therefore, is to develop new strategies for designing explorative probes to target sequences that have not yet been described.

3.1.3 Explorative probe design strategies for POA

There are two ways to detect unknown microorganisms: using probes defined from known high phylogenetic levels and using explorative probes that correspond to new sequence variants of existing phylogenetic signatures that are not yet deposited in public databases but potentially present in the environment.

The “multiple probe concept” consists of several probes to target an organism at different phylogenetic levels (*e.g.*, genus, family, order). This strategy dramatically reduces the risk of misidentification and substantially increases the resolution of the analysis by

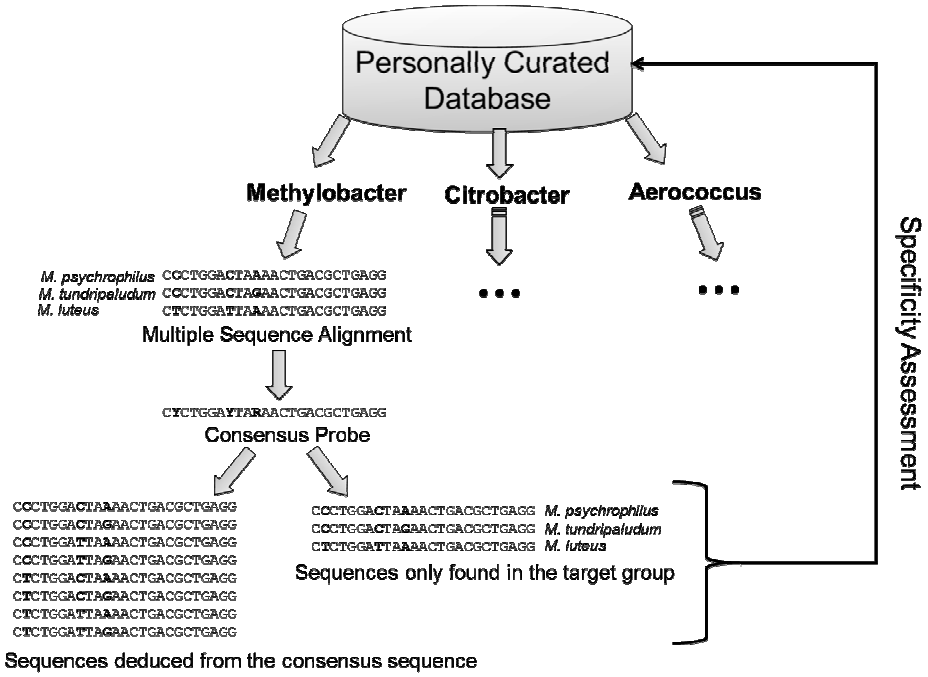

Figure 3. PhylArray programme workflow. The PhylArray programme is composed of four steps: (i) sequence extraction for each taxon, (ii) multiple sequence alignment, (iii) degenerate consensus sequence production and probe selection and (iv) specificity tests against the 16S rRNA database.

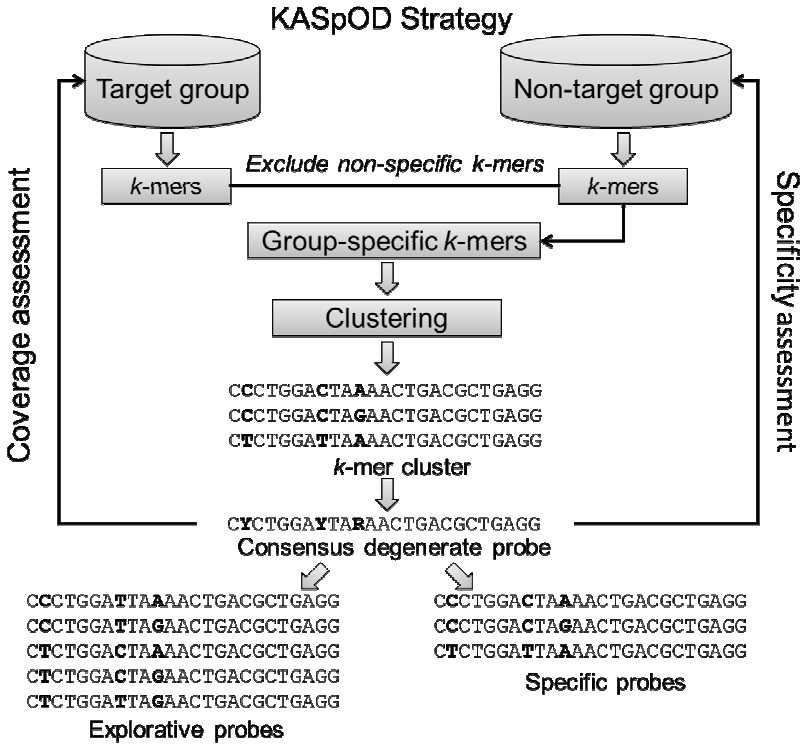

Figure 4. KASpOD programme workflow. The KASpOD programme is composed of three computational steps: (i) search for group-specific *k*-mers, (ii) consensus *k*-mer building, and (iii) coverage and specificity assessment.

discriminating bacteria down to the species level (65, 89, 91, 136, 137). The use of this strategy to construct POAs is well suited to ensuring the detection of unknown microorganisms by probes defined at higher taxonomic levels. Nevertheless, such probes are strictly complementary to known sequences and do not harbour the explorative power to detect microorganisms with uncharacterised phylogenetic signatures (40).

The first software programme dedicated to POAs that offered the possibility of designing explorative probes was the PhylArray programme (101). PhylArray was developed to survey whole microbial communities, including known and unknown microorganisms, in complex environments. The first step of the PhylArray algorithm **(Figure 3)** is the extraction of all available sequences corresponding to a targeted taxon from a custom 16S rRNA curated database. Retrieved sequences are then aligned using the ClustalW programme (151). A degenerate consensus sequence is then deduced from this multiple sequence alignment, taking into account the sequence variability at each position. Degenerate candidate probes are then selected along the consensus sequence, and all non-degenerate combinations are checked for cross-hybridisations against the 16S rRNA database. Among the combinations derived from each degenerate probe, some correspond to sequences that have not yet been deposited in public databases, namely explorative probes. Such probes should, therefore, allow the detection of undescribed microorganisms belonging to the targeted taxon. Probes defined using this software, which were recently used to evaluate the bacterial diversity in soils (31), yield a higher sensitivity and specificity than probes designed using the PRIMROSE and ARB strategies (101). PhylArray was designed to account for all of the sequence variability within the targeted sequences, but because it relies on multiple sequence alignment, it is limited in its ability to manage large input datasets. Consequently, new probe design strategies are needed to define explorative probes based on large databases.

KASpOD (112) software was developed to overcome this limitation. KASpOD (K-mer Based Algorithm for Highly Specific and Explorative Oligonucleotide Design) consists of three computational stages **(Figure 4)**. The user first provides two datasets that correspond to the target group and the non-target group. The first stage is the extraction of every *k*-mer from the target and the non-target groups using the Jellyfish programme (97). For large target groups containing more than 100 sequences, a noise reduction step is performed to remove untrustworthy *k*-mers that occur only once. Every *k*-mer found in both the target and the non-

target groups is removed from the list of oligonucleotide candidates. The selected *k*-mers are then clustered together using CD-HIT (85) at an 88% identity threshold (*i.e.*, allowing three mismatches for 25-mer probes). Only fully overlapping *k*-mers are clustered to gather *k*-mers from the same genomic location. For each cluster, a degenerate consensus is constructed that accounts for the sequence variability within the cluster. Among the combinations derived from each degenerate oligonucleotide, some correspond to sequences not previously included in the target group and therefore represent explorative probes.

Finally, the last stage of the KASpOD algorithm consists of assessing the coverage and specificity of each degenerate consensus *k*-mer. The coverage is evaluated against the target group using the PatMan programme (119), which allows the user to perform an exhaustive search with mismatches and indels to identify all occurrences of a high number of short sequences within a large database. The user defines the upper limit of tolerated mismatches. Specificity is assessed in the same way using the non-target group. KASpOD is provided as both a web service (http://g2im.u-clermont1.fr/kaspod/) and a stand-alone package. The software was used to design 25-mer probes for 1,295 prokaryotic genera based on the recently published Greengenes taxonomy (98). The defined probe set allows each of the 252,183 high-quality and non-redundant 16S rRNA sequences to be covered by at least three different probes. Finally, 22,613 group-specific signatures were designed and are freely available on the KASpOD web site (http://g2im.u-clermont1.fr/kaspod/about.php). The alignment-free strategy allows computations to be completed in approximately two weeks. Furthermore, this approach enables the definition of probes for large groups such as the *Corynebacterium* genus (20,093 sequences) where an alignment-based algorithm would have failed.

3.2 Functional Gene Arrays (FGAs)

Microbes mediate almost every conceivable biological process, and some researchers have estimated that individual environmental samples such as soil may contain between 10^3 and 10^7 different bacterial genomes (Curtis et al 2002, Gans et al 2005)(28, 29, 45), each harbouring thousands of genes. In this context, high-density oligonucleotide FGAs provide the best high-throughput tools to access this tremendous diversity (59). Currently, the most

Table 2. Comparison of probe design software features for functional gene arrays (FGAs).

Software	Reference	Application	Availability	URL
DEODAS (v 0.1.2)	(44)	FGA	Downloadable, GUI (L)	http://deodas.sourceforge.net/
Metabolic Design	(149)	FGA	Downloadable, GUI (W)	ftp://195.221.123.90/
ProDesign	(42)	FGA	Web Interface	http://www.uhnresearch.ca/labs/tillier/ProDesign/ProDesign.html
ArrayOligoSelector (v 3.8.4)	(13)	FGA, WGA-ORF	Downloadable, command-line (L)	http://arrayoligosel.sourceforge.net
CommOligo (v 2.0)	(87)	FGA, WGA-ORF	Downloadable, standalone GUI (W)	http://ieg.ou.edu/software.htm
HiSpOD	(39)	FGA, WGA-ORF	Web Interface	http://g2im.u-clermont1.fr/hispod
MProbe (v 2.0)	(86)	FGA, WGA-ORF	Downloadable, GUI (W)	http://www.biosun.org.cn/mprobe/
OligoArray (v 2.1)	(128)	FGA, WGA-ORF	Downloadable, command-line (L)	http://berry.engin.umich.edu/oligoarray2_1/
OligoPicker (v 2.3.2)	(161)	FGA, WGA-ORF	Downloadable, command-line (L)	http://pga.mgh.harvard.edu/oligopicker/
OligoWiz (v 2.2.0)	(163)	FGA, WGA-ORF	Downloadable client programme, GUI (L, W, M)	http://www.cbs.dtu.dk/services/OligoWiz2
PRIMEGENS (v 2.0)	(167)	FGA, WGA-ORF	Web interface or command-line standalone (L, W)	http://primegens.org/
UPS 2.0	(22)	FGA, WGA-ORF	Web interface	http://array.iis.sinica.edu.tw/ups/

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
DEODAS (v 0.1.2)	Range chosen by the user	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	No	No	No	No	Yes
Metabolic Design	Fixed by the user	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	No	No	No	No	Yes
ProDesign	Range chosen the user (20-70)	Read input sequences from 5'-end to 3'-end	Maximum number of probes chosen by the user	Yes	Yes	Yes	Yes	Yes	No
ArrayOligoSelector (v3.8.4)	Fixed by the user	Probes ranking according to the 3' end distance	Chosen by the user	Yes	Yes	Yes	No	No	No
CommOligo (v 2.0)	Fixed by the user (8-128)	Design starting from the 3'- or 5'-end	Chosen by the user	Yes	Yes	Yes	Yes	No	No
HiSpOD	Fixed by the user (18-120)	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	Yes	Yes	Yes	No	Yes
MProbe (v 2.0)	Range chosen by the user (20-100)	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	Yes	No	Yes	Yes	No	No
OligoArray (v 2.1)	Fixed by the user (15-75)	Distance to the 3'-end specified by the user (max 1500)	Chosen by the user	Yes	Yes	Yes	Yes	No	No
OligoPicker (v 2.3.2)	Fixed by the user (20-100)	Design chosen for the 5'- or the 3'-end	Chosen by the user (up to 5)	Yes	Yes	No	Yes	No	No
OligoWiz (v 2.2.0)	Fixed by the user	Localisation score based on centre, 5' or 3' distance	All probes reaching selection criteria	Yes	No	No	Yes	No	No
PRIMEGENS (v 2.0)	Fixed by the user	No localisation specified	Chosen by the user	Yes	No	Yes	Yes	No	No
UPS 2.0	Fixed by the user (20-120)	Read input sequences from 5'-end to 3'-end	Chosen by the user (up to 10)	Yes	Yes	Yes	Yes	Yes	Yes

Software	Organism	Cross-hybridisation assessment	Database for specificity test	Input files
DEODAS (v 0.1.2)	No limitation	EMBOSS	GenBank	A FASTA file containing targeted sequences. Protein sequences.
Metabolic Design	No limitation	BLAST and Kane's specifications	EnvExBase (10GB) Complete CDS database	No input files are required.
ProDesign	No limitation	Spaced seed hashing and Kane's specifications	Input sequence dataset	A FASTA file containing targeted sequences and optionally a cluster file. Nucleotide sequences.
ArrayOligoSelector (v 3.8.4)	No limitation	BLAST and thermodynamic calculations	External FASTA file (typically single organism genome)	Two FASTA files (targeted sequences and the complete genome). Nucleotide sequences.
CommOligo (v 2.0)	No limitation	Global alignment and thermodynamic calculations	Input sequence dataset	A FASTA file containing targeted sequences. Nucleotide sequences.
HiSpOD	No limitation	BLAST and Kane' specifications	EnvExBase (10GB) Complete CDS database	A FASTA file containing targeted sequences. Consensus or non-degenerate nucleotide sequences.
MProbe (v 2.0)	No limitation	BLAST and Kane's specifications	Input sequence dataset	A GenBank, EMBL or FASTA file containing targeted sequences. Nucleotide sequences.
OligoArray (v 2.1)	No limitation	BLAST and thermodynamic calculations	External FASTA file (typically single organism genome)	A FASTA file containing targeted sequences. Nucleotide sequences.
OligoPicker (v 2.3.2)	No limitation	BLAST	Input sequence dataset or external FASTA file (typically single organism genome)	A FASTA file containing targeted sequences. Nucleotide sequences.
OligoWiz (v 2.2.0)	All organisms found on the server	BLAST, Kane's specifications and thermodynamic calculations	Single organism genome (among a list of organisms found on the server)	A FASTA file containing targeted sequences or a tab-delimited file containing both sequences and annotations. Nucleotide sequences.
PRIMEGENS (v 2.0)	No limitation	BLAST, Kane's specifications and multiple sequence alignment	External FASTA file or a complete genome sequence among a list of organisms found on the server	A FASTA file containing targeted sequences. Nucleotide sequences.
UPS 2.0	No limitation	BLAST and thermodynamic calculations	Input sequence dataset, complete genome sequence among a list of organisms found on the server, NCBI Nucleotide database or external FASTA file	A FASTA file containing targeted sequences and optionally a fasta file with non-target sequences. Nucleotide sequences.

comprehensive FGA is the GeoChip (55, 57), which has evolved over several generations to be able to monitor most microbial functional processes, such as carbon, nitrogen, sulphur and phosphorus cycling, energy metabolism, antibiotic resistance, metal resistance, and organic contaminant degradation (56, 58, 60). Although most strategies are limited to the determination of probes that target specific gene sequences within a single genome dataset, few strategies offer the opportunity to design probes that permit broad coverage of multiple sequence variants for a given gene family **(Table 2)** (40, 80).

3.2.1 Probe design for FGAs using nucleic sequences

GeoChips are composed of 50-mer probes designed using a modified version of CommOligo (87). The experimental assessment of optimal probe design criteria (61) permitted CommOligo to be implemented to combine three different parameters for sensitivity and specificity evaluation: sequence identity, free energy and continuous stretch. For each sequence, the first stage consists of masking oligonucleotides according to different filters including distance to the 3' untranslated region (UTR), GC content, complexity, degeneracy and specificity (*i.e.*, significant matching of oligonucleotides with non-targets). Continuous matches of a user-specified length with non-targets are assessed using an algorithm similar to that of OligoPicker (161) by storing all possible 10-mers within the sequences in a hash table data structure. Thus, the hash key is a 10-mer sequence, and the hash value corresponds to the relative sequence indices and positions where this particular 10-mer is found. Strictly identical 10-mers shared between probes and non-targets are not retained. This data structure is also used to assess the self-annealing of each unmasked oligonucleotide by searching for continuous matches of a user-defined length within the tested oligonucleotide itself. Probes that show self-annealing are filtered out. The remaining probes are tested for specificity against non-targets using both a global alignment algorithm (104) and a binding free energy calculation rather than the classically used Basic Local Alignment Search Tool (BLAST) (2). Sequence identity is therefore inferred from the percentage of matches in a global gapped alignment, and oligonucleotides with high identity (*i.e.*, higher than a cut-off value defined by the user) to non-targets are filtered out. Oligonucleotides with medium identity but low free energy are also removed. Then, the programme computes the best interval of melting temperatures that covers most targets and probes and removes all candidate oligonucleotides

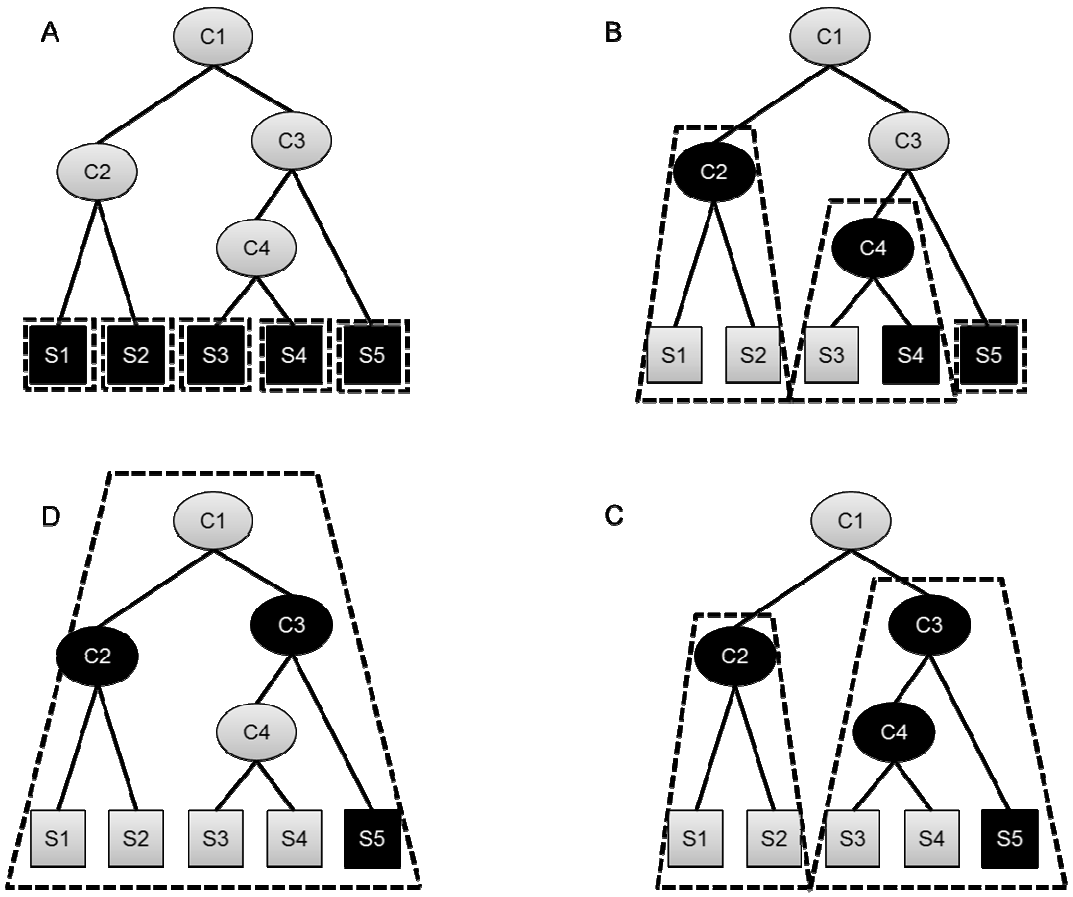

Figure 5. HPD probe candidate selection process. Sequences are hierarchically clustered (A), and a bottom-up approach (B, C) is performed to search for putative group-specific probes that can target the whole dataset (D). Black circles or squares within dotted regions indicate that probe candidates exist for that cluster or sequence. Grey circles or squares within the dotted regions indicate that no probe candidate exists for that cluster or sequence. The circles outside of the dotted region indicate the clusters that have not yet been explored.

outside of this range. Finally, a sequence may have more probes than needed, in which case CommOligo is able to select oligonucleotides using a multi-criterion optimisation algorithm where cross-hybridisation, positions and identity between probes are taken into account. Gene-specific probes can be selected using CommOligo with the following parameters: < 90% sequence identity, < 20-base continuous stretch, and > -35 kcal/mol free energy with non-targets (58, 88). Additionally, a group-specific probe design can be performed by adding these supplemental criteria: >96% sequence identity, > 35-base continuous stretch, and < -60 kcal/mol free energy within the targeted group (58, 61). CommOligo performs complex and time-consuming calculations. However, the version available for download is not well suited to conducting high-throughput analyses, and with the increasing availability of sequences corresponding to protein-coding genes (complete genome sequencing and environmental studies from specific functional markers), new software has been developed in the last decade that takes this wide diversity into account.

Hierarchical Probe Design (HPD) software (25) was the first programme dedicated to FGAs that was based on the concept of cluster-specific probes. The first step of the algorithm consists of the multiple sequence alignment of input sequences using ClustalW (151). A hierarchical clustering is then performed using either a neighbour-joining (132) or a UPGMA (141) method. All candidate probes are subsequently generated, and cluster-specific probes are selected using a bottom-up approach **(Figure 5)**. The specificity of candidate oligonucleotides is checked against clusters that are one level higher. If a probe of one sibling cluster harbours sufficient specificity to discriminate among these clusters, it remains in the sibling cluster. If not, the candidate is transferred to the upper cluster and therefore represents a group-specific probe. This recursive process is repeated as long as the root cluster has not been reached. The optimal probe set is then determined according to probe quality criteria including cluster coverage, specificity, GC content and hairpin energy. Although this tool is not explorative, it automatically produces probes against all nodes of the clustering tree, thereby providing extensive coverage of known variants from a conserved functional gene. However, at this time, the software no longer appears to be available. ProDesign (42) uses similar clustering methods with the aim of detecting all members of a gene family in environmental samples. However, in contrast to HPD, this software uses sophisticated spaced seed hashing rather than a suffix tree algorithm to benefit from permitted mismatches between a probe and its targets, and it ensures the re-clustering of groups for which no probe was

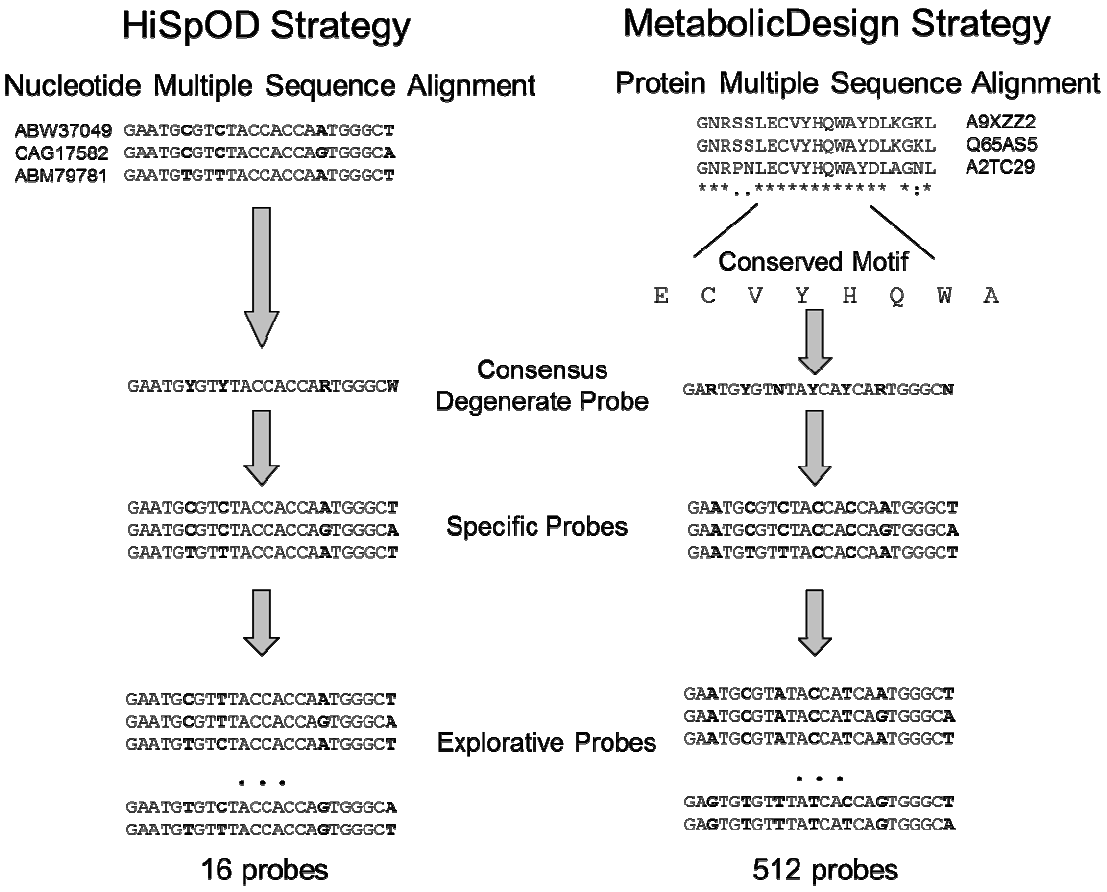

Figure 6. Comparison of the explorative probe design strategies implemented in HiSpOD and Metabolic Design software. The example shows the probe design for the *bphA1c* gene encoding the salicylate 1-hydroxylase alpha subunit involved in PAH degradation from three distinct *Sphingomonas* or *Sphingobium* species using both strategies.

found, which results in a significant improvement in sequence coverage. Although both of these strategies allow the coverage of a wider range of sequence variants, they only permit the survey of known sequences and therefore cannot be used to evaluate the unknown microbial genes present in complex environments. The main drawbacks of these strategies are thus their inability to generate explorative probes and the absence of an evaluation of specificity (*i.e.*, searching for potential cross-hybridisations) against large databases that are representative of microbial diversity.

To overcome these limitations, the HiSpOD (High Specific Oligo Design) programme was developed (39) in the context of microbial ecology. HiSpOD includes the classical parameters for the design of effective probes, including probe length, melting temperature, GC content and complexity, and adds supplemental properties that were not considered by previous programs. HiSpOD allows the design of degenerate probes for gene families after multiple alignments of nucleic sequences belonging to the same gene family and can produce consensus sequences. All combinations deduced from the degenerate probes are then divided into two groups **(Figure 6A)**. The first group corresponds to specific probes for sequences available in databases, and the second group corresponds to explorative probes that represent putative new signatures that do not correspond to any previously described microorganisms. A probe set representing the most likely gene sequence variants and a probe set representing new combinations that have not yet been deposited in databases are created based on multiple mutation events that have already been identified. Sequence-specific probes can also be designed through HiSpOD by using non-degenerate classical nucleic acid sequences. To limit cross-hybridisation, the specificity of all selected probes is checked against a large formatted database dedicated to microbial communities, i.e., the EnvExBase (Environmental Expressed sequences dataBase), which is composed of all coding DNA sequences (CDSs) from the prokaryotic (PRO), fungal (FUN) and environmental (ENV) taxonomic divisions of the EMBL databank. Specificity tests are performed using BLAST (2), and cross-hybridisation results are clustered using a single-linkage method implemented in BLASTCLUST (2).

3.2.2 Probe design for FGAs using protein sequences

In contrast to the strategies outlined above, several new strategies have been proposed to initiate probe design from conserved peptidic regions rather than from nucleic acid sequences to survey all potential nucleic acid variants.

The first strategy based on this principle was described by Bontemps *et al.* (11) and called CODEHMOP (Consensus Degenerate Hybrid Motif Oligonucleotide Probe). This strategy is derived from an adaptation of the CODEHOP (Consensus Degenerate Hybrid Oligonucleotide Primer) PCR primer design strategy, which was originally developed to identify distantly related genes encoding proteins that belong to known families (12, 125, 126). The CODEHMOP strategy aims to identify conserved amino acid motifs from multiple alignments of protein sequences. Then, the most highly conserved region (5-7 amino acids) of each protein motif is backtranslated to generate all possible nucleic combinations (15-21 nucleotides) coding for this peptide. These sequences are extended by 5' and 3' fixed ends (12-15 nucleotides each) that are derived from the most frequent nucleotide at each position flanking the conserved region in the nucleotide sequence alignment. The final probes are called "hybrids", as they comprise a variable central core with some nucleic combinations that do not correspond to any sequences yet described (to target greater diversity) combined with two fixed end sequences (available in databases) that are added to increase the probe length. This approach was used to design a prototype DNA array that included all described and undescribed *nodC* (nodulation gene) sequences in bacteria and that was applied to legume nodule samples (11). This strategy enabled the detection of new *nodC* sequences that exhibited less than 74% identity with known sequences. The application of the CODEHMOP strategy is, however, limited by its lack of implementation in a fully automated programme and its lack of probe specificity test. Nevertheless, this approach appears to be the most comprehensive way to encompass the diversity of gene sequence variants potentially found for enzymes mediating a given function.

Terrat *et al.* (149) developed a new software programme called Metabolic Design that ensures the *in silico* reconstruction of metabolic pathways, the identification of conserved motifs from multiple protein alignments, and the generation of efficient explorative probes through a simple convenient graphical interface. In this case, before the probe design stage, the user reconstructs the chosen metabolic pathway *in silico* with all substrates and products from each metabolic step. One reference enzyme for each of these steps is selected, and its

protein sequence is extracted from a curated database (by default, Swiss-Prot), which is then used to retrieve all homologous proteins from complete databases (Swiss-Prot and TrEMBL). After the most pertinent homologous sequences are selected, they are aligned to begin the probe design stage. The amino acids are backtranslated for each identified molecular site, with all redundancy of the genetic code taken into account, to produce a degenerate nucleic consensus sequence. All degenerate probes that meet the criteria defined by the user (probe length and maximal degeneracy) are retained. All of the possible specific combinations for each degenerate probe are subsequently checked for potential cross-hybridisation against a representative database (*e.g.*, EnvExBase as in the HiSpOD programme). Finally, an output file listing all of the degenerate probes selected by the user permits the deduction of all possible combinations and organises them into specific probes and exploratory probes (Figure 6B).

3.3 Whole-genome arrays (WGAs)

Many organisms that are closely related based on SSU rRNA gene sequences can exhibit remarkably different phenotypic characteristics that result from great differences in their genomes, which in turn arise from processes such as lateral gene exchange (47). Whole-genome arrays that use whole-genome sequence information of one or several closely related microorganisms provide a way of understanding such phenotypic differences (170). WGAs are divided into two main groups: whole-genome ORF arrays, which contain oligonucleotide probes for all of the open reading frames (ORFs) in a genome, and tiling arrays, which represent a complete non-repetitive tile path over the genome, irrespective of any genes that may be annotated in a particular region (9).

3.3.1 Whole-genome ORF arrays

Probe design considerations for whole-genome ORF arrays are similar to those for functional gene arrays (FGAs). However, some algorithms are specifically dedicated to the design of oligonucleotide probes for whole-genome ORF arrays. One of the most-cited software programs for designing such microarrays is PICKY (24). PICKY was initially developed for

oligonucleotide microarray design for large eukaryotic genomes and thus boasts major speed improvements when compared with other whole-genome ORF array probe design programs. Several probe design criteria are considered by PICKY to compute the optimal probe set, such as the complexity (*i.e.*, no single base should constitute more than 50% of a probe and no stretch of the same base should exceed 25% of the length of a probe), thermodynamic criteria (*i.e.*, a GC content between 30 and 70% and no secondary structures), and cross-hybridisation [*i.e.*, Kane's criteria (71)]. For the latter criterion, PICKY can handle multiple target and non-target gene sets; thus, oligonucleotide probes are defined for the target set and to prevent hybridisation with the non-target set. Most of the previously mentioned criteria are user-adjustable through a user-friendly graphical interface. Finally, to ensure the uniformity of the probe set, PICKY is able to adjust the probe length within a user-defined range.

PICKY relies on the construction of a generalised suffix array where both strands are represented. The suffix array is built using a modified Burkhardt-Kärkkaïnen algorithm (16) that allows quick and efficient construction. Using this suffix array, PICKY can first exclude low-complexity and repetitive genomic regions as well as self-similar and self-complementary regions for probe design. Such screening allows the detection of putative secondary structures without using dedicated external software. PICKY then avoids other unnecessary computations by removing regions that fail to comply with Kane's criteria, as these regions may cross-hybridise with non-target sequences. For all remaining regions, PICKY computes a score to indicate the likelihood of cross-hybridisation and then prioritises regions for oligonucleotide selection. Once all probe candidates have been computed, the melting temperatures for all possible probe/target and probe/non-target pairs are estimated according to Kane's second condition, which states that any sequence similarity over 75% identity (or a user-defined value) can potentially involve cross-hybridisation. The melting temperature of each candidate probe is assessed against its target, and all non-targets are gathered using the suffix array. As probe/non-target pairs may have mismatches, melting temperatures are not precise but are sufficient to predict whether such duplexes will potentially be present.

The calculated melting temperatures of candidate probes with all of its non-targets are then used to prioritise probes for the final processing step. This last step consists of multi-objective optimisation to compute the probe set best able to detect each gene, i.e., by avoiding

Table 3. Comparison of probe design software features for whole-genome ORF arrays (WGAs).

Software	Reference	Application	Availability	URL
Mprime	(127)	WGA-ORF	Web interface	http://kbrin.a-bldg.louisville.edu/Tools/OligoDesign/MPrime.html
OliD	(147)	WGA-ORF	Downloadable, command line (L)	Upon request
PICKY (v 2.2)	(24)	WGA-ORF	Downloadable, standalone GUI (L, W, M)	http://www.complex.iastate.edu./download/Picky/index.html
ProbeSelect	(84)	WGA-ORF	Available upon request, command line (L)	http://stormo.wustl.edu/src/probeselect-src.tar

WGA-ORF: open reading-frame oriented whole-genome array. GUI: graphical user interface. L: Linux. M: MacOS. W: Windows.

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
Mprime	Fixed by the user	Probes weighted towards 3'-end	Chosen by the user	Yes	No	Yes	Yes	No	No
OliD	Fixed by the user	Preference given to the 3'-end proximity	Chosen by the user	Yes	Yes	Yes	No	No	No
PICKY (v 2.2)	Range chosen by the user (50-90)	No localisation specified	Chosen by the user (up to 5)	Yes	Yes	Yes	Yes	No	No
ProbeSelect	Fixed by the user	No localisation specified	Chosen by the user	Yes	Yes	No	Yes	Yes	No

Software	Organism	Cross-hybridisation assessment	Database for specificity test	Input files
Mprime	Rat, mouse, human, drosophila and zebrafish	BLAST (wuBLAST)	RefSeq database for the organism	Gene name, GenBank accession number, keyword, or FASTA files. Nucleotide sequences.
OliD	No limitation	BLAST	External FASTA file (typically single organism genome)	A FASTA file containing targeted sequences. Nucleotide sequences.
PICKY (v 2.2)	No limitation	Suffix array approach, Kane's specifications and thermodynamic calculations	External FASTA file (typically single organism genome)	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.
ProbeSelect	No limitation	Suffix array approach and thermodynamic calculations	Single organism genome	A FASTA file containing targeted sequences. Nucleotide sequences.

cross-hybridisation and sharing a uniform melting temperature range. Thus, PICKY is a fast and efficient probe design software programme for whole-genome ORF arrays that addresses numerous design criteria to compute the optimal probe set. No additional external software is required to run PICKY, and it is easy to use through a graphical user interface that is available for all major computing platforms (Mac, Windows and Linux).

Several existing probe design software programs **(Table 3)** are free to use and still available for designing whole-genome ORF arrays with various strategies such as Mprime (127), OliD (147), PROBESEL (69) or ProbeSelect (84).

Targeting all genes through the use of whole-genome ORF arrays may not be sufficient for some applications, such as the identification of transcription in the antisense strand or regulatory pathway discovery (8). Consequently, PICKY proposes PERL scripts for tiling array purposes.

3.3.2 Tiling arrays

In contrast to whole-genome ORF arrays, tiling arrays aim to determine probes over the whole genome irrespective of any genes that may be annotated in the genome. The design of oligonucleotide tiling arrays is different from the selection of oligonucleotides for gene-based arrays, and additional factors should be considered, such as tiling resolution and the handling of non-unique sub-sequences.

A naïve strategy for selecting oligonucleotides for a whole-genome tiling array is to generate a tile path from the beginning of a chromosome to its end and cover the entire sequence with 25-mer probes tiled end-to-end (168). Many of the probes chosen using this approach may, however, be subject to hybridisation problems and potentially result in the misinterpretation of results. Some probes may be redundant, some may be thermodynamically unable to hybridise and some may be non-specific and thus undergo cross-hybridisation (102). Consequently, in most cases, such a naïve strategy is not the optimal approach to tiling, and the inclusion of criteria such as tiling resolution and repetitive region masking is essential to ensure the best probe design for tiling arrays.

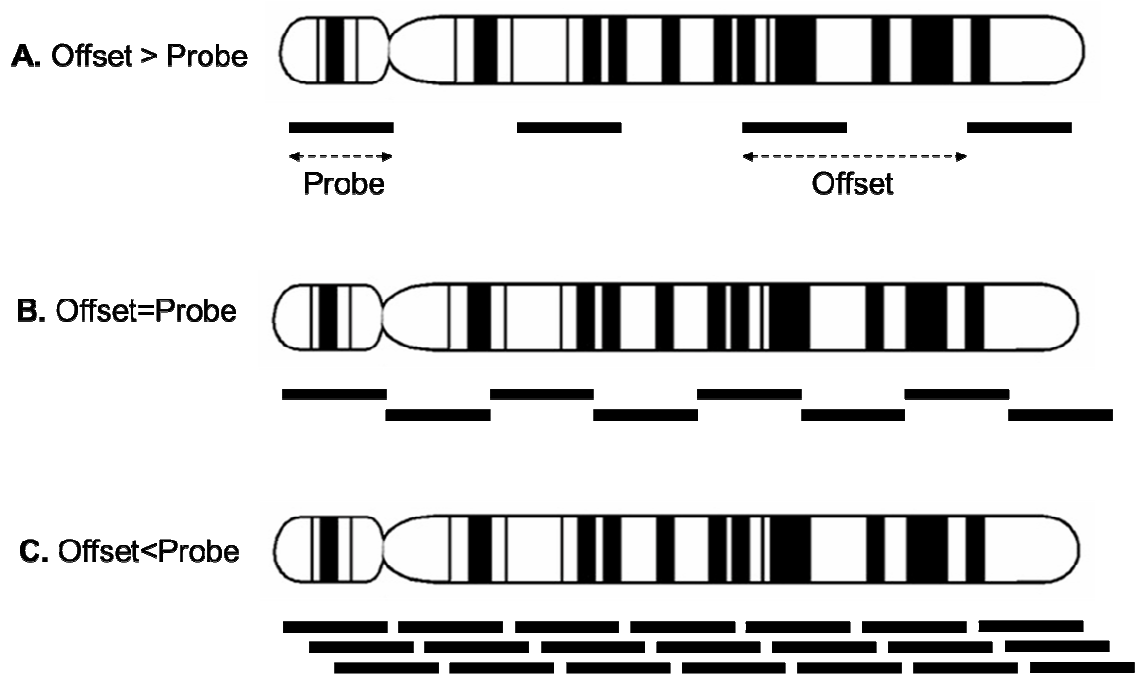

Figure 7. Tiling density. The offset between probes is the distance between the start of one probe and the start of the next. Three different tiling densities are shown: (A) illustrates gapped tiling, (B) end-to-end tiling and (C) overlapping tiling.

Tiling density **(Figure 7)** is an important factor in a tiling array design because it determines how the genome should be subdivided and how densely oligonucleotide probes are placed. Probes can be contiguous (*i.e.*, tiled end-to-end) or discontiguous, including gaps with a predetermined size range between probes for single-copy tiling. For some applications, it may be necessary to have a better resolution involving multiple feature tiling. The whole genome is covered with multiple oligonucleotide probes such that the starting position of each probe is shifted by one or several nucleotides to overlap the previous oligonucleotide's coordinates (9). Although such a strategy allows a fine-resolution analysis, the number of probes determined will eventually dramatically increase. However, advances in high-resolution microarray technology have enabled the inclusion of up to 4.2 million probes on an array.

Because genome sequences are not random, many redundant sub-sequences are scattered all along the genome. Therefore, once the tiling strategy has been determined, a common first step is to perform a genomic repeat masking prior to probe selection. To obtain an easy and relevant interpretation of a tiling array experiment, it is necessary to avoid the generation of multiple oligonucleotide probes sharing the same sequence. Because such sequences generally correspond to known repetitive sequences, algorithms such as RepeatMasker (http://www.repeatmasker.org/) are widely used to easily address this problem. RepeatMasker is capable of identifying genomic repeats in a variety of genomes using a database of well-characterised families of repetitive elements (68). It is also preferable to remove low-complexity DNA regions (*i.e.*, stretches of the same nucleotide or regions with extremely high A/T or G/C content). RepeatMasker allows the filtering of some low-complexity sequences by default, but it could be necessary in some cases to combine RepeatMasker with dedicated software that calculates entropy scores, such as NSEG (166) or DUST (52), for a more intensive filtering, especially of specific repetitive sequences of the studied genome.

The next step consists of filtering out oligonucleotide probes based on thermodynamic considerations (133). Low-binding affinity probes are useless in a tiling array experiment, as are high affinity non-specific probes, which would be uninformative because of the saturated cross-hybridisation. Therefore, probe affinity modelling and the determination of probe

Table 4. Comparison of probe design software features for tiling arrays.

Software	Reference	Application	Availability	URL
ChipD	(36)	WGA-tiling	Web interface	http://chipd.uwbacter.org/
MAMMOT (v 1.21)	(130)	WGA-tiling	Downloadable, local server	http://www.mammot.org.uk/
MOPeD	(113)	WGA-tiling	Web interface	http://moped.genetics.emory.edu/newdesign.html
OligoTiler	(9)	WGA-tiling	Web interface	http://tiling.gersteinlab.org/OligoTiler/oligotiler.cgi
PanArray (v 1.0)	(115)	WGA-tiling	Downloadable, command line (L)	http://www.cbcb.umd.edu/software/panarray/
Teolenn (v 2.0.1)	(67)	WGA-tiling	Downloadable, command line (L)	http://transcriptome.ens.fr/teolenn/

WGA-tiling: tiling whole-genome array. L: Linux.

Software	Probe length (nt)	Design orientation	Number of probes designed by gene	Secondary structure	Low complexity	GC content	T_m	ΔG	Degenerate probes
ChipD	Range chosen by the user (greater than 15)	Read input sequences from 5'-end to 3'-end	Maximum number of probes chosen by the user	No	Yes	No	Yes	Yes	No
MAMMOT (v 1.21)	Fixed by the user	Start and end locations are chosen by the user	All probes reaching selection criteria	No	Yes	Yes	Yes	No	No
MOPeD	Range chosen by the user (55-65)	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	Yes	No	No	Yes	No	No
OligoTiler	Fixed by the user	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	Yes	No	No	No	No
PanArray (v 1.0)	Fixed by the user	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	No	No	No	No	No
Teolenn (v 2.0.1)	Fixed by the user or range chosen by the user	Read input sequences from 5'-end to 3'-end	All probes reaching selection criteria	No	Yes	Yes	Yes	No	No

Software	Organism	Cross-hybridisation assessment	Database for specificity test	Input files
ChipD	No limitation	No	No	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.
MAMMOT (v 1.21)	No limitation	No	No	A FASTA file with the complete genome sequence. Nucleotide sequences.
MOPeD	Human, mouse, rhesus monkey	No	No	No input files are required.
OligoTiler	No limitation	No	No	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.
PanArray (v 1.0)	No limitation	No	No	A FASTA file containing targeted sequences (typically multiple genome sequences). Nucleotide sequences.
Teolenn (v 2.0.1)	No limitation	No	No	A FASTA file containing targeted sequences (typically a single organism genome). Nucleotide sequences.

specificity on a whole-genome level can be used to screen candidate oligonucleotides and eliminate those likely to be problematic from the microarray design (102).

For all of these reasons, it is necessary to adapt the tiling strategy and placement of oligonucleotide probes along the genome to obtain the optimal probe set. To design the oligonucleotides in each sequence window, the probe design software has to address position and hybridisation quality (80). **Table 4** summarises the available and free-to-use probe design software dedicated to tiling arrays: chipD (36), Teolenn (67), MOPeD (113), PanArray (115), Tileomatic (135), ArrayDesign (50) and MAMMOT (130).

Among these algorithms, Teolenn appears to be the most universal and flexible software to address the tiling array design problem and remains easy to use despite its command-line utilisation. Teolenn relies on a four-step workflow where each step is customisable. Thus, users are allowed to activate or deactivate each function according to their needs or available computational resources. Teolenn accepts both masked (*e.g.*, using RepeatMasker) and unmasked genome sequences in FASTA format as input.

The first step consists of generating all possible non-redundant oligonucleotide probes along the whole genome. Probe length can be fixed or may vary within a user-defined length range. In the second step, for all created probes, Teolenn assesses the oligonucleotide quality based on several criteria including melting temperature, GC content, complexity and uniqueness. Melting temperatures are computed using a nearest-neighbour method (133), complexity is measured by counting the masked bases and uniqueness within the genome is evaluated according to Gräf *et al.* (50). The third step of Teolenn is probe filtering, which strongly depends on the tiling array application. For example, if the user needs a transcriptome array, Teolenn is able to filter out all probes that are not located within an ORF as well as small RNAs based on genome annotations. However, if a homogeneous tiling path along the genome is desired, Teolenn can keep all of the possible probes without stringent quality filters. Eventually, the best probe in each genomic window that maximises a position score, where the most central probe has the best score, and a previously assessed quality score can be selected. Depending on the user's needs, all score calculations can be weighted. The designed probes can be output in plain text, FASTA format or GFF. The GFF format allows

the visualisation of the results in a genome browser such as gBrowse (143). Teolenn therefore provides a complete and flexible software programme dedicated to tiling arrays.

3.4 Other applications

3.4.1 Transcriptome arrays

Gene-oriented whole-genome arrays and tiling arrays can both be used to measure the expression of thousands of genes of an organism, thus providing a snapshot of the transcriptome in different states in tissues and cells (105). Compared with tiling arrays, gene-based WGAs appear to be relatively simple to handle, as they use fewer probes for each gene. In contrast, tiling arrays are capable of providing information regarding alternative splicing or realising the transcriptome annotation of a newly sequenced organism.

The choice between a gene-oriented WGA and a tiling platform is highly correlated with the biological question. There are also many commercial gene expression microarrays available that should be considered before undertaking a custom design.

3.4.2 Typing microarrays

Identifying microbial communities using microarrays is a common task that is generally performed using probe design software dedicated to POAs. However, for some applications such as discrimination among several strains of the same bacterial species or the classification of plasmids, it is necessary to develop dedicated tools. Such tools (100, 157) aim to design an optimal set of oligonucleotide fingerprints that is able to distinguish among similar targets. For particular applications including strain detection microarrays, a custom design is often inevitable. Two different probe design approaches are used to build these microarrays: a traditional gene-oriented approach and a strategy originally developed for sequencing by hybridisation: resequencing microarrays (82). Oligonucleotide probe design for resequencing microarrays is substantially different from previously discussed strategies. Such strategies were developed to characterise bacterial pathogens by sequencing a significant portion of their genome or, for viral pathogens, the whole genome. Multiple versions of each

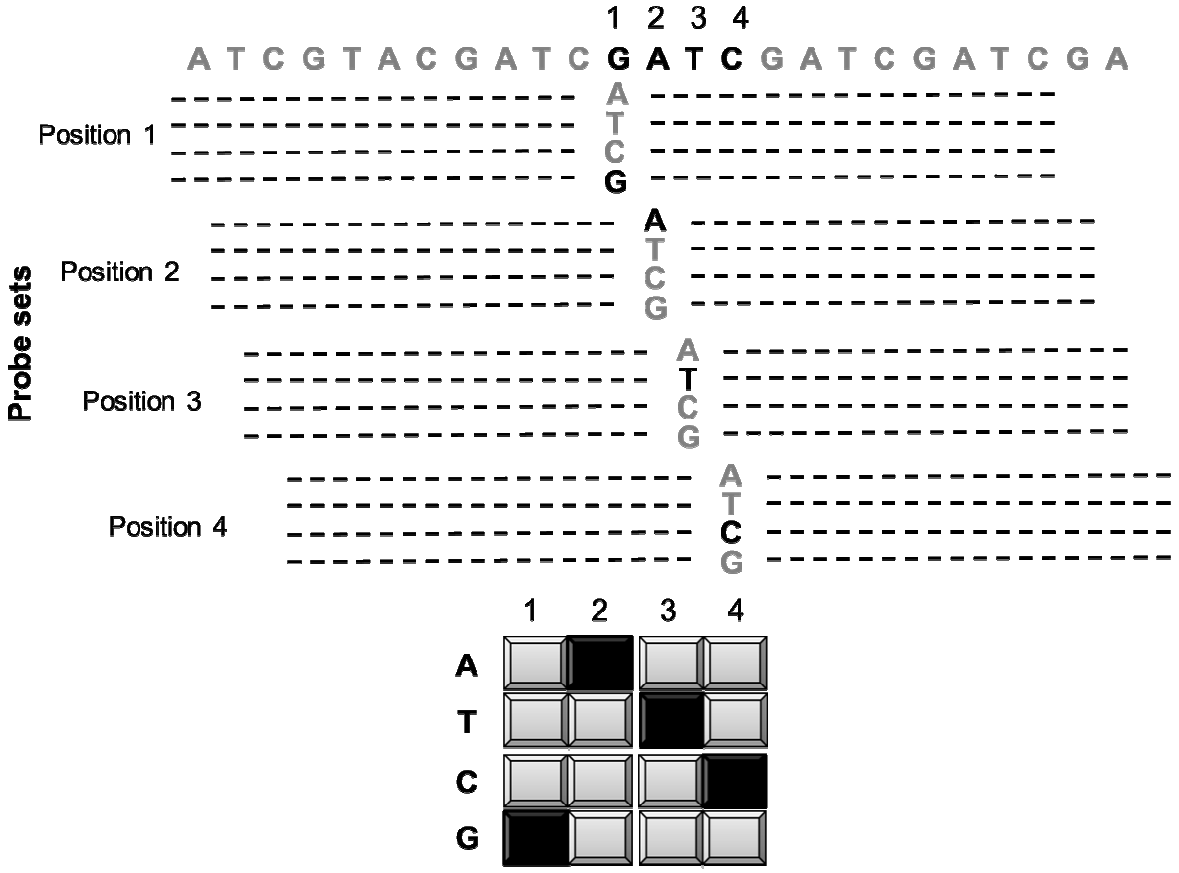

Figure 8. Resequencing microarrays. An example of a probe set construction targeting four contiguous nucleotides (1, 2, 3 and 4) of a prototype sequence using Affymetrix technology. Overlapping sets of probes shifted by one nucleotide covering the whole prototype sequence are generated. Each probe set contains one matching (black) and three mismatched (grey) probes that differ only by the central nucleotide (four additional probes could be designed for the other strand). Black lines represent sequences with the prototype sequence above. The bottom figure shows a scan of a segment of the microarray to which the labelled targets were hybridised. Black squares correspond to a positive signal. The base calls are therefore made according to the hybridisation pattern.

oligonucleotide are identified, *i.e.*, four probes in both the sense and antisense directions, for a total of eight probes per base. The four probes differ by only one central nucleotide to represent the four possible base combinations (*i.e.*, A, T, C or G). Based on hybridisation results, base-calling algorithms are able to obtain a reliable sequence that can be compared with public sequence databases **(Figure 8)**.

Fingerprints could also be generated by random probes (6). In this example, the authors created 9015 12-mers with homogeneous GC contents where each probe differed by at least four bases and 5268 13-mers that differed by at least five bases. After the hybridisation step, the authors demonstrated that their strategy produced reproducible patterns of hybridisation that distinguished among species. The use of such a probe design eliminates the need for updating the array design when new bacterial genomes become available. The primary drawback of this strategy is the need to experimentally obtain the hybridisation pattern for organisms without known sequences.

Discussion/Challenges and future trends

The success of a microarray experiment strongly depends on the determination of the best probe set while taking the biological question into account. Despite the development of numerous probe design strategies, some parameters require particular attention, as they have significant impact on probe specificity, sensitivity and quantitative capability (160, 170).

4.1 Explorative probe design

For environmental DNA microarrays including phylogenetic oligonucleotide arrays (POAs) and functional gene arrays (FGAs), explorative probe design strategies offer the opportunity to survey both known and unknown microorganisms (40). Explorative probes use the sequence variability within the targeted sequences to define new combinations that have not yet been deposited in public databases but are potentially present in the environment. One future development in microarrays will be to incorporate the original concept into probe design software, especially by offering the ability to design group-specific degenerate probes.

4.2 Probe length

Another probe design criterion that impacts both sensitivity and specificity is probe length. Short oligonucleotide probes are more specific but less sensitive than long probes (51). The building of phylogenetic oligonucleotide microarrays requires the determination of short fingerprints that are able to discriminate among microbial taxa. Existing POAs, therefore, use short probes (*i.e.*, 24-25-mers) (14, 53, 54, 109, 121), whereas FGAs or whole-genome ORF arrays may be built with either short (*i.e.*, 15-30-mers) (10, 144) or long oligonucleotides (*i.e.*, 40-70-mers) (38, 55, 57, 71, 122). The primary limitation of microarrays based on short oligonucleotide probes is the need to use, in most cases, PCR-amplified targets to ensure enrichment, which introduces an inherent PCR bias (145, 158).

A promising alternative approach to the design of oligonucleotide probes is the use of the GoArrays strategy (124) (http://g2im.u-clermont1.fr/serimour/goarrays.html). Such a strategy enables the production of oligonucleotide probes that are as specific as short probes and as sensitive as long probes, and consists of the concatenation of two short subsequences that are complementary to disjoined regions of the target, with an insertion of a short random linker (*i.e.*, 3-6-mers). This strategy has been shown to improve microarray efficiency for a wide range of applications (72, 111, 124, 171).

4.3 Databases

Most currently available probe design software programs only perform specificity tests against a reduced set of sequences, such as whole-genome data or specific sets of genes (80). Environmental DNA microarrays, however, require dedicated datasets that are as representative as possible of all of the non-target sequences potentially present in the samples. Because only a small portion of the total natural microbial diversity has been documented, it is a major challenge to design suitable probes that are specific to unique markers and do not cross-hybridise with putative and currently unknown similar sequences (20). There is a trade-off between using the largest databases and thus minimising putative cross-hybridisations and using small, dedicated databases that are less time-consuming for specificity tests. The major

public databases including GenBank (7), the European Nucleotide Archive (ENA) (79) and the DNA Data Bank of Japan (DDBJ) (70) are the most complete nucleic sequence databases with which to perform specificity tests. However, the use of such databases could drastically increase the run times of probe design software. In addition, for environmental DNA microarrays, entire public databases are not really appropriate because they contain some subsets that are not typically considered in microbial ecology, such as *Metazoa*. Furthermore, numerous erroneous annotations could negatively impact the quality of the probe design.

Within POAs, each probe must be specific with respect to all small subunit (SSU) rRNA sequences that may be present in the sample during hybridisation. Curated and dedicated secondary databases that gather all of the SSU rRNA sequences described in public databases have already been constructed [*e.g.*, Ribosomal Database Project (27), Greengenes (35) and SILVA (118)]. The differences among these databases arise from the construction and update workflows, which lead to distinct sizes: SILVA (Release 111) contains 3,194,778 16S rRNA sequences, RDP (Release 10) contains 2,578,902 16S rRNA sequences and Greengenes (10/2/2011) contains 1,049,116 16S rRNA sequences. Because they are smaller and contain no unnecessary data, these databases are well-adapted to the construction of prokaryotic POAs. PhylArray software (101) was, however, developed before these databases were publicly available, and therefore uses its own highly curated (full-length and quality-filtered) and automatically updated prokaryotic SSU rRNA database (66,076 sequences for the last release).

For environmental FGAs, the database used for specificity tests must include all known CDSs that may be encountered in natural environments. To the best of our knowledge, EnvExBase (used in the HiSpOD and Metabolic Design programs) is the first CDS database dedicated to microbial ecology (39, 149). For its construction, all annotated transcript sequences and their associated 5’ and 3’ untranslated regions (UTR) in all classes of the EMBL prokaryotic (PRO), fungal (FUN) and environmental (ENV) taxonomic divisions were extracted and curated to remove low-quality sequences. EnvExBase thus represents a 13,697,580 sequence database.

4.4 Updates and Performance

The constant increase in available sequences (26) requires that databases for specificity tests must be regularly updated. As a result, probe datasets must be re-computed as frequently as possible to include all deposited data. However, as mentioned above, assessing probe specificity against large databases can be a time-consuming task in the probe design step. A complete 16S rRNA sequence database is approximately two million sequences, and a complete CDS database such as EnvExBase represents 14 million sequences. To overcome this limitation, an interesting strategy would be to create databases specific to each ecological compartment. Usually, specificity tests are not performed against a suitable subset of sequences, primarily because of the lack of databases for microbial ecology. Depending on the environment studied, it would be more relevant to perform these tests against reduced databanks dedicated to specific ecosystems (*e.g.*, soil, marine, freshwater, and gut). However, for "universal" tool development relevant to various environments, the most complete database must be considered.

The rapid growth of datasets, particularly environmental datasets, has led to an important increase in computational requirements coupled with a fundamental change in the way that algorithms are conceived and designed [*e.g.*, mpiBLAST (30) or GPU-BLAST (159)]. Efforts to limit computation time are based on exploiting the computational resources available using specialised frameworks such as Message Passing Interface (MPI) or heterogeneous systems including General-purpose Processing on Graphics Processing Units (GPGPU). With the recent development of extremely fast broadband networks, it has become possible to distribute the calculations at increasing scales over different geographical locations (134). Cluster, grid and emerging cloud computing are all examples of shared computing resources where probe design algorithms must be deployed if specificity tests and alignments are to be performed with reasonable data processing times (46, 152)(Gardner et al 2006, Thorsen et al 2007).

4.5 Microarray formats

As mentioned above, explorative design strategies that allow the detection of unknown sequences involve the use of degenerate probes (11, 39, 101, 112, 149). The selected strategy will therefore greatly influence the choice between the two major DNA microarray types (*ex*

situ or *in situ*), the platform and the density (37, 41, 74). When *in situ* synthesis microarrays such as the Agilent, Affymetrix and NimbleGen platforms are used, all non-degenerate combinations that result from a degenerate probe have to be independently synthesised. Consequently, the final number of probes (*i.e.*, density) will exponentially increase for the array production. For instance, for the CODEHMOP (11) and Metabolic Design (149) strategies that were developed for the FGAs probe design, the degenerate probes are derived from the multiple sequence alignment of all possible nucleotide sequences that are able to code for the targeted conserved amino acid motif. Because the genetic code often involves degeneracy at the third position of each codon, a 24-mer probe (*i.e.*, targeting a seven amino acid conserved motif) will generate at least 128 combinations (assuming a minimal degeneracy rate of two for each codon). This value will reach at least 131,072 for a 51-mer probe containing 17 degenerate positions. Conversely, *ex situ* platforms allow the degenerate probes (all combinations mixed together) to be spotted in the same location on the array and consequently reduce the total number of features. However, in this latter case, the sensitivity may be affected by the complexity of the mixed oligonucleotides.

Other user choices may also affect the final number of probes per array. Replication is crucial for achieving reliable data for microarrays (142). Multiple replicates of the same probe provide some backup if a feature cannot be evaluated because of technical artifacts such as dye precipitation or dust particles. A statistical estimation has deduced that at least three replicates should be located (78). Additionally, multiple probes could be designed per gene to increase confidence in the results (23, 92)(Chou et al 2004, Loy et al 2002a) and to mask misleading signal variations whose causes (*e.g.*, target secondary structure, probe folding) are not yet fully understood (116). Third, some platforms such as Affymetrix GeneChips determine probe pairs where each probe (“match”) is accompanied by a negative control with a single differing base in a central position (“mismatch probe”) to discriminate between true signals and those arising from non-specific hybridisation (90).

To address the problem of the number of probes, several commercial companies have proposed two major types of high-density microarrays: (i) *in situ* synthesised microarrays, which are distributed by Agilent (http://www.chem.agilent.com), NimbleGen (http://www.nimblegen.com) and Affymetrix (http://www.affymetrix.com), can contain billions of probes and can be physically divided into multiple arrays per slide (up to 12) to

perform simultaneous analyses of several samples in a single experiment; and (ii) spotted microarrays [*e.g.*, Arrayit (http://www.arrayit.com)] with a current printing capacity close to 100,000 features per microarray.

4.6 Analysis

Because probe design software programs are numerous, microarray formats are heterogeneous and biological questions are different, the analysis of microarray data results can be a major challenge. For instance, if an explorative design strategy has been performed, the signals encountered for these probes must be carefully interpreted. Consequently, a future direction for this field could be to develop automatic procedures to analyse microarray data.

4.7 Other applications

Probe design is now a common task in molecular ecology. Probes can be used in several molecular techniques including microarrays as well as PCR, quantitative PCR, and FISH. In addition, a promising strategy for reducing the complexity of environmental samples by enriching the desired genomic target using probes before sequencing is being adapted for microbial ecology. The more efficient methods rely on the complementary hybridisation of nucleic acid capture probes to the targeted DNA sequences; these methods use either solid phase hybridisation (*e.g.*, using capture arrays) (1, 103, 108), or solution phase, also known as Solution Hybridisation Selection (SHS) (32, 49, 150).

The use of explorative probe design in sequence-capture approaches that couple with next generation sequencing (NGS), such as those originally developed for the direct selection of human genomic loci (1), could also improve characterisation of microbial communities in microbial ecology. In fact, sequence capture elution products should allow the full identification and characterisation of new taxa when using phylogenetic probes or new protein-coding genes with functional probes. Furthermore, the innovative approach developed by Denonfoux *et al.* (32) aims to capture large DNA fragments and allows the identification of genes flanking the targeted biomarkers. Probe design software programs remain a popular research topic and have a promising future.

Conclusions

With the availability of high-density custom microarrays, the selection of high-quality oligonucleotide probes is a crucial task. Although microarrays are particularly well-suited for the detection and quantification of genes or transcripts, accurate measurements depend on good probe design. Oligonucleotide design is an optimisation task and must take into account various parameters that influence the interaction between the probe and the target. Increasingly, the recent development of computational methods as well as the increase in the number of available sequences in databases allows the selection of large probe sets with a wide spectrum of thermodynamic properties. Several software solutions are available to help the user and solve the current bottlenecks in the choice of high-quality probe sets that must combine basic criteria such as sensibility, specificity and uniformity. Each software programme has advantages and drawbacks, and the choice of programmes must be made in total accordance with the nature of projects and the basic scientific question. Probe design strategies have evolved and hence become more easily computed over time, thus providing a foundation for more sophisticated work in bioinformatics. Although much progress has been achieved, probe design for microarray experiments remains a challenging and active research field.

References

1. **Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA.** 2007. Direct selection of human genomic loci by microarray hybridization. Nature Methods **4**:903–905.

2. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.** 1990. Basic local alignment search tool. Journal of Molecular Biology **215**:403–410.

3. **Ashelford KE, Weightman AJ, Fry JC.** 2002. PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. Nucleic Acids Research **30**:3481–3489.

4. **Auer H, Lyianarachchi S, Newsom D, Klisovic MI, Marcucci G, Marcucci U, Kornacker K. 2003.** Chipping away at the chip bias: RNA degradation in microarray analysis. Nature Genetics **35**:292–293.

5. **Bader KC, Grothoff C, Meier H. 2011.** Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. Bioinformatics **27**:1546–1554.

6. **Belosludtsev YY, Bowerman D, Weil R, Marthandan N, Balog R, Luebke K, Lawson J, Johnston SA, Lyons CR, Obrien K, Garner HR, Powdrill TF.** 2004. Organism identification using a genome sequence-independent universal microarray probe set. BioTechniques **37**:654–8– 660.

7. **Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW.** 2012. GenBank. Nucleic Acids Research **40**:D48–53.

8. **Bertone P, Gerstein M, Snyder M.** 2005. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Res. **13**:259–274.

9. **Bertone P, Trifonov V, Rozowsky JS, Schubert F, Emanuelsson O, Karro J, Kao M-Y, Snyder M, Gerstein M.** 2006. Design optimization methods for genomic DNA tiling arrays. Genome Research **16**:271–281.

10. **Bodrossy L, Stralis-Pavese N, Murrell JC, Radajewski S, Weilharter A, Sessitsch A.** 2003. Development and validation of a diagnostic microbial microarray for methanotrophs. Environmental Microbiology **5**:566–582.

11. **Bontemps C, Golfier G, Gris-Liebe C, Carrere S, Talini L, Boivin-Masson C.** 2005. Microarray-based detection and typing of the Rhizobium nodulation gene nodC: potential of DNA arrays to diagnose biological functions of interest. Applied and Environmental Microbiology **71**:8042–8048.

12. **Boyce R, Chilana P, Rose TM.** 2009. iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. Nucleic Acids Research **37**:W222–8.

13. **Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL.** 2003. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biology **4**:R9.

14. **Brodie EL, DeSantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL, Hazen TC, Richardson PM, Herman DJ, Tokunaga TK, Wan JM, Firestone MK.** 2006. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Applied and Environmental Microbiology **72**:6288–6298.

15. **Brodie EL, DeSantis TZ, Parker JPM, Zubietta IX, Piceno YM, Andersen GL.** 2007. Urban aerosols harbor diverse and dynamic bacterial populations. Proceedings of the National Academy of Sciences of the United States of America **104**:299–304.

16. **Burkhardt S, Kärkkäinen J.** 2003. Fast Lightweight Suffix Array Construction and Checking. Combinatorial Pattern Matching **2676**:55–69.

17. **Busti E, Bordoni R, Castiglioni B, Monciardini P, Sosio M, Donadio S, Consolandi C, Rossi Bernardi L, Battaglia C, De Bellis G.** 2002. Bacterial discrimination by means of a universal array approach mediated by LDR (ligase detection reaction). BMC Microbiol. **2**:27.

18. **Candela M, Consolandi C, Severgnini M, Biagi E, Castiglioni B, Vitali B, De Bellis G, Brigidi P.** 2010. High taxonomic level fingerprint of the human intestinal microbiota by ligase detection reaction--universal array approach. BMC Microbiol. **10**:116.

19. **Castiglioni B, Rizzi E, Frosini A, Sivonen K, Rajaniemi P, Rantala A, Mugnai MA, Ventura S, Wilmotte A, Boutte C, Grubisic S, Balthasart P, Consolandi C, Bordoni R, Mezzelani A, Battaglia C, De Bellis G.** 2004. Development of a universal microarray based on the ligation detection reaction and 16S rrna gene polymorphism to target diversity of cyanobacteria. Applied and Environmental Microbiology **70**:7161–7172.

20. **Chandler DP, Jarrell AE.** 2005. Taking arrays from the lab to the field: trying to make sense of the unknown. BioTechniques **38**:591–600.

21. **Chandler DP, Newton GJ, Small JA, Daly DS.** 2003. Sequence versus structure for the direct detection of 16S rRNA on planar oligonucleotide microarrays. Applied and Environmental Microbiology **69**:2950–2958.

22. **Chen S-H, Lo C-Z, Su S-Y, Kuo B-H, Hsiung CA, Lin C-Y.** 2010. UPS 2.0: unique probe selector for probe design and oligonucleotide microarrays at the pangenomic/genomic level. BMC Genomics **11 Suppl 4**:S6.

23. **Chou C-C, Chen C-H, Lee T-T, Peck K.** 2004. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. Nucleic Acids Research **32**:e99.

24. **Chou H-H, Hsia A-P, Mooney DL, Schnable PS.** 2004. Picky: oligo microarray design for large genomes. Bioinformatics **20**:2893–2902.

25. **Chung W-H, Rhee S-K, Wan X-F, Bae J-W, Quan Z-X, Park Y-H.** 2005. Design of long oligonucleotide probes for functional gene detection in a microbial community. Bioinformatics **21**:4092–4100.

26. **Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C, Jang M, Juhos S, Leinonen R, Leonard S, Lin Q, Lopez**

R, Lorenc D, McWilliam H, Mukherjee G, Plaister S, Radhakrishnan R, Robinson S, Sobhany S, Hoopen PT, Vaughan R, Zalunin V, Birney E. 2009. Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Research **37**:D19–25.

27. **Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, Tiedje JM.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Research **37**:D141–5.

28. **Curtis TP, Head IM, Lunn M, Woodcock S, Schloss PD, Sloan WT.** 2006. What is the extent of prokaryotic diversity? Philosophical transactions of the Royal Society of London Series B, Biological sciences **361**:2023–2037.

29. **Curtis TP, Sloan WT, Scannell JW.** 2002. Estimating prokaryotic diversity and its limits. Proceedings of the National Academy of Sciences of the United States of America **99**:10494–10499.

30. **Darling A, Carey L, Feng W.** 2003. The design, implementation, and evaluation of mpiBLAST. Proceedings of ClusterWorld 2003.

31. **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. Applied and Environmental Microbiology **77**:1315–1324.

32. **Denonfoux J, Parisot N, Dugat-Bony E, Biderre-Petit C, Boucher D, Morgavi DP, Le Paslier D, Peyretaillade E, Peyret P.** Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration. DNA Research (submitted)

33. **DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL.** 2007. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microbial Ecology **53**:371–383.

34. **DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL.** 2006. NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes. Nucleic Acids Research **34**:W394–9.

35. **DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology **72**:5069–5072.

36. **Dufour YS, Wesenberg GE, Tritt AJ, Glasner JD, Perna NT, Mitchell JC, Donohue TJ.** 2010. chipD: a web tool to design oligonucleotide probes for high-density tiling arrays. Nucleic Acids Research **38**:W321–5.

37. **Dufva M.** 2005. Fabrication of high quality microarrays. Biomol. Eng. **22**:173–184.

38. **Dugat-Bony E, Biderre-Petit C, Jaziri F, David MM, Denonfoux J, Lyon DY, Richard J-Y, Curvers C, Boucher D, Vogel TM, Peyretaillade E, Peyret P.** 2012. In situ TCE degradation mediated by complex dehalorespiring communities during biostimulation processes. Microb Biotechnol **5**:642–653.

39. **Dugat-Bony E, Missaoui M, Peyretaillade E, Biderre-Petit C, Bouzid O, Gouinaud C, Hill DRC, Peyret P.** 2011. HiSpOD: probe design for functional DNA microarrays. Bioinformatics **27**:641–648.

40. **Dugat-Bony E, Peyretaillade E, Parisot N, Biderre-Petit C, Jaziri F, Hill DRC, Rimour S, Peyret P.** 2012. Detecting unknown sequences with DNA microarrays: explorative probe design strategies. Environmental Microbiology **14**:356–371.

41. **Ehrenreich A.** 2006. DNA microarray technology for the microbiologist: an overview. Applied Microbiology and Biotechnology **73**:255–273.

42. **Feng S, Tillier ERM.** 2007. A fast and flexible approach to oligonucleotide probe design for genomes and gene families. Bioinformatics **23**:1195–1202.

43. **Franke-Whittle IH, Goberna M, Pfister V, Insam H.** 2009. Design and development of the ANAEROCHIP microarray for investigation of methanogenic communities. Journal of Microbiological Methods **79**:279–288.

44. **Fredrickson HL, Perkins EJ, Bridges TS.** 2001. Towards environmental toxicogenomics — development of a flow-through, high-density DNA hybridization array and its application to ecotoxicity assessment. The Science of the Total environment.

45. **Gans J, Wolinsky M, Dunbar J.** 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science **309**:1387–1390.

46. **Gardner M, Feng W-C, Archuleta J, Lin H, Ma X.** 2006. Parallel Genomic Sequence-Searching on an Ad-Hoc Grid: Experiences, Lessons Learned, and Implications, pp. 22–22. In. IEEE.

47. **Gentry TJ, Wickham GS, Schadt CW, He Z, Zhou J.** 2006. Microarray applications in microbial ecology research. Microbial Ecology **52**:159–175.

48. **Gerry NP, Witowski NE, Day J, Hammer RP, Barany G, Barany F.** 1999. Universal DNA microarray method for multiplex detection of low abundance point mutations. Journal of Molecular Biology **292**:251–262.

49. **Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C.** 2009.

Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology **27**:182–189.

50. **Gräf S, Nielsen FGG, Kurtz S, Huynen MA, Birney E, Stunnenberg H, Flicek P.** 2007. Optimized design and assessment of whole genome tiling arrays. Bioinformatics **23**:i195–204.

51. **Guschin DY, Mobarry BK, Proudnikov D, Stahl DA, Rittmann BE, Mirzabekov AD.** 1997. Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. Applied and Environmental Microbiology **63**:2397–2402.

52. **Hancock JM, Armstrong JS.** 1994. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. Comput. Appl. Biosci. **10**:67–70.

53. **Handley KM, Wrighton KC, Piceno YM, Andersen GL, DeSantis TZ, Williams KH, Wilkins MJ, N'guessan AL, Peacock A, Bargar J, Long PE, Banfield JF.** 2012. High-density PhyloChip profiling of stimulated aquifer microbial communities reveals a complex response to acetate amendment. FEMS Microbiology Ecology.

54. **Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, Jansson JK, Probst A, Borglin SE, Fortney JL, Stringfellow WT, Bill M, Conrad ME, Tom LM, Chavarria KL, Alusi TR, Lamendella R, Joyner DC, Spier C, Baelum J, Auer M, Zemla ML, Chakraborty R, Sonnenthal EL, D'haeseleer P, Holman H-YN, Osman S, Lu Z, Van Nostrand JD, Deng Y, Zhou J, Mason OU.** 2010. Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science **330**:204–208.

55. **He Z, Deng Y, Van Nostrand JD, Tu Q, Xu M, Hemme CL, Li X, Wu L, Gentry TJ, Yin Y, Liebich J, Hazen TC, Zhou J.** 2010. GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. The ISME journal **4**:1167–1179.

56. **He Z, Deng Y, Zhou J.** 2011. Development of functional gene microarrays for microbial community analysis. Curr. Opin. Biotechnol.

57. **He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, Criddle C, Zhou J. 2007.** GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. The ISME journal **1**:67–77.

58. **He Z, Van Nostrand JD, Deng Y, Zhou J.** 2011. Development and applications of functional gene microarrays in the analysis of the functional diversity, composition, and structure of microbial communities. Frontiers of Environmental Science & Engineering Science China.

59. **He Z, Van Nostrand JD, Wu L, Zhou J.** 2008. Development and application of functional gene arrays for microbial community analysis. Transactions of Nonferrous Metals Society of China **18**:1319–1327.

60. **He Z, Van Nostrand JD, Zhou J.** 2012. Applications of functional gene microarrays for profiling microbial communities. Curr. Opin. Biotechnol.

61. **He Z, Wu L, Li X, Fields MW, Zhou J.** 2005. Empirical establishment of oligonucleotide probe design criteria. Applied and Environmental Microbiology **71**:3753–3760.

62. **Huber T, Faulkner G, Hugenholtz P.** 2004. Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics **20**:2317–2319.

63. **Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS.** 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nature Biotechnology **19**:342–347.

64. **Hultman J, Ritari J, Romantschuk M, Paulin L, Auvinen P.** 2008. Universal ligation-detection-reaction microarray applied for compost microbes. BMC Microbiol. **8**:237.

65. **Huyghe A, François P, Charbonnier Y, Tangomo-Bento M, Bonetti E-J, Paster BJ, Bolivar I, Baratti-Mayer D, Pittet D, Schrenzel J, Geneva Study Group on Noma (GESNOMA).** 2008. Novel microarray design strategy to study complex bacterial communities. Applied and Environmental Microbiology **74**:1876–1885.

66. **Jacobs KA, Rudersdorf R, Neill SD, Dougherty JP, Brown EL, Fritsch EF.** 1988. The thermal stability of oligonucleotide duplexes is sequence independent in tetraalkylammonium salt solutions: application to identifying recombinant DNA clones. Nucleic Acids Research **16**:4637–4650.

67. **Jourdren L, Duclos A, Brion C, Portnoy T, Mathis H, Margeot A, Le Crom S.** 2010. Teolenn: an efficient and customizable workflow to design high-quality probes for microarray experiments. Nucleic Acids Research **38**:e117.

68. **Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J.** 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. **110**:462–467.

69. **Kaderali L, Schliep A.** 2002. Selecting signature oligonucleotides to identify organisms using DNA arrays. Bioinformatics **18**:1340–1349.

70. **Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.** 2011. DDBJ progress report. Nucleic Acids Research **39**:D22–7.

71. **Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, Madore SJ.** 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Research **28**:4552–4557.

72. **Kang S, Denman SE, Morrison M, Yu Z, Dore J, Leclerc M, McSweeney CS.** 2010. Dysbiosis of fecal microbiota in Crohn's disease patients as revealed by a custom phylogenetic microarray. Inflamm. Bowel Dis. **16**:2034–2042.

73. **Kaplinski L, Scheler O, Parkel S, Palta P, Toome K, Kurg A, Remm M.** 2010. Detection of tmRNA molecules on microarrays at low temperatures using helper oligonucleotides. BMC Biotechnol. **10**:34.

74. **Kawasaki ES.** 2006. The end of the microarray Tower of Babel: will universal standards lead the way? J Biomol Tech **17**:200–206.

75. **Kämpke T, Kieninger M, Mecklenburg M.** 2001. Efficient primer design algorithms. Bioinformatics **17**:214–225.

76. **Koltai H, Weingarten-Baror C.** 2008. Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction. Nucleic Acids Research **36**:2395–2405.

77. **Kostić T, Weilharter A, Rubino S, Delogu G, Uzzau S, Rudi K, Sessitsch A, Bodrossy L.** 2007. A microbial diagnostic microarray technique for the sensitive detection and identification of pathogenic bacteria in a background of nonpathogens. Anal. Biochem. **360**:244–254.

78. **Lee ML, Kuo FC, Whitmore GA, Sklar J.** 2000. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proceedings of the National Academy of Sciences of the United States of America **97**:9834–9839.

79. **Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Hoopen Ten P, Vaughan R, Zalunin V, Cochrane G.** 2011. The European Nucleotide Archive. Nucleic Acids Research **39**:D28–31.

80. **Lemoine S, Combes F, Le Crom S.** 2009. An evaluation of custom microarray applications: the oligonucleotide design challenge. Nucleic Acids Research **37**:1726–1739.

81. **Leparc GG, Tüchler T, Striedner G, Bayer K, Sykacek P, Hofacker IL, Kreil DP.** 2009. Model-based probe set optimization for high-performance microarrays. Nucleic Acids Research **37**:e18.

82. **Leski TA, Lin B, Malanoski AP, Stenger DA.** 2012. Application of resequencing microarrays in microbial detection and characterization. Future Microbiol **7**:625–637.

83. **Letowski J, Brousseau R, Masson L.** 2004. Designing better probes: effect of probe size, mismatch position and number on hybridization in DNA oligonucleotide microarrays. Journal of Microbiological Methods **57**:269–278.

84. **Li F, Stormo GD.** 2001. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics **17**:1067–1076.

85. **Li W, Godzik A.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**:1658–1659.

86. **Li W, Ying X.** 2006. Mprobe 2.0: computer-aided probe design for oligonucleotide microarray. Appl. Bioinformatics **5**:181–186.

87. **Li X, He Z, Zhou J.** 2005. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Research **33**:6114–6123.

88. **Liebich J, Schadt CW, Chong SC, He Z, Rhee S-K, Zhou J.** 2006. Improvement of oligonucleotide probe design criteria for functional gene microarrays in environmental applications. Applied and Environmental Microbiology **72**:1688–1691.

89. **Liles MR, Turkmen O, Manske BF, Zhang M, Rouillard J-M, George I, Balser T, Billor N, Goodman RM.** 2010. A phylogenetic microarray targeting 16S rRNA genes from the bacterial division Acidobacteria reveals a lineage-specific distribution in a soil clay fraction. Soil Biol. Biochem. **42**:739–747.

90. **Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ.** 1999. High density synthetic oligonucleotide arrays. Nature Genetics **21**:20–24.

91. **Loy A, Bodrossy L.** 2006. Highly parallel microbial diagnostics using oligonucleotide microarrays. Clin. Chim. Acta **363**:106–119.

92. **Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J, Schleifer K-H, Wagner M.** 2002. Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. Applied and Environmental Microbiology **68**:5064–5081.

93. **Loy A, Schulz C, Lücker S, Schöpfer-Wendels A, Stoecker K, Baranyi C, Lehner A, Wagner M.** 2005. 16S rRNA gene-based oligonucleotide microarray for environmental

monitoring of the betaproteobacterial order "Rhodocyclales". Applied and Environmental Microbiology **71**:1373–1386.

94. **Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer K-H.** 2004. ARB: a software environment for sequence data. Nucleic Acids Research **32**:1363–1371.

95. **Maldonado-Rodriguez R, Espinosa-Lara M, Loyola-Abitia P, Beattie WG, Beattie KL.** 1999. Mutation detection by stacking hybridization on genosensor arrays. Mol Biotechnol **11**:13–25.

96. **Manber U, Myers G.** 1993. Suffix arrays: a new method for on-line string searches. SIAM Journal on Computing 22.

97. **Marcais G, Kingsford C.** 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics **27**:764–770.

98. **McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P.** 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME journal **6**:610–618.

99. **Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen M-M, Lu G, Fang J, Liu W-M, Ryder T, Kaplan P, Kulp D, Webster TA.** 2003. Probe selection for high-density oligonucleotide arrays. Proceedings of the National Academy of Sciences of the United States of America **100**:11237–11242.

100. **Meng D, Broschat SL, Call DR.** 2008. A Java-based tool for the design of classification microarrays. BMC Bioinformatics **9**:328.

101. **Militon C, Rimour S, Missaoui M, Biderre-Petit C, Barra V, Hill DRC, Moné A, Gagne G, Meier H, Peyretaillade E, Peyret P.** 2007. PhylArray: phylogenetic probe design algorithm for microarray. Bioinformatics **23**:2550–2557.

102. **Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, Ecker JR.** 2005. Applications of DNA tiling arrays for whole-genome analysis. Genomics **85**:1–15.

103. **Mokry M, Feitsma H, Nijman IJ, de Bruijn E, van der Zaag PJ, Guryev V, Cuppen E.** 2010. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Research **38**:e116.

104. **Myers G.** 1999. A fast bit-vector algorithm for approximate string matching based on dynamic programming. J. ACM **46**:395–415.

105. **Nakaya HI, Reis EM, Verjovski-Almeida S.** 2007. Nucleic Acids Hybridization Modern Applications. Springer Netherlands, Dordrecht.

106. **Neufeld JD, Mohn WW, de Lorenzo V.** 2006. Composition of microbial communities in hexachlorocyclohexane (HCH) contaminated soils from Spain revealed with a habitat-specific microarray. Environmental Microbiology **8**:126–140.

107. **Nordberg EK.** 2005. YODA: selecting signature oligonucleotides. Bioinformatics **21**:1365–1370.

108. **Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME.** 2007. Microarray-based genomic selection for high-throughput resequencing. Nature Methods **4**:907–909.

109. **Paliy O, Agans R.** 2011. Application of phylogenetic microarrays to interrogation of human microbiota. FEMS Microbiology Ecology.

110. **Palmer C, Bik EM, Eisen MB, Eckburg PB, Sana TR, Wolber PK, Relman DA, Brown PO.** 2006. Rapid quantitative profiling of complex microbial populations. Nucleic Acids Research **34**:e5.

111. **Pariset L, Chillemi G, Bongiorni S, Romano Spica V, Valentini A.** 2009. Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. N Biotechnol **25**:272–279.

112. **Parisot N, Denonfoux J, Dugat-Bony E, Peyret P, Peyretaillade E.** 2012. KASpOD - A web service for highly specific and explorative oligonucleotide design. Bioinformatics.

113. **Patel VC, Mondal K, Shetty AC, Horner VL, Bedoyan JK, Martin D, Caspary T, Cutler DJ, Zwick ME.** 2010. Microarray oligonucleotide probe designer (MOPeD): A web service. Open Access Bioinformatics **2**:145–155.

114. **Peplies J, Glöckner FO, Amann R.** 2003. Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. Applied and Environmental Microbiology **69**:1397–1407.

115. **Phillippy AM, Deng X, Zhang W, Salzberg SL.** 2009. Efficient oligonucleotide probe selection for pan-genomic tiling arrays. BMC Bioinformatics **10**:293.

116. **Pozhitkov AE, Tautz D, Noble PA.** 2007. Oligonucleotide microarrays: widely applied--poorly understood. Brief Funct Genomic Proteomic **6**:141–148.

117. **Pozhitkov AE, Tautz D.** 2002. An algorithm and program for finding sequence specific oligonucleotide probes for species identification. BMC Bioinformatics **3**:9.

118. **Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO.** 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research **35**:7188–7196.

119. **Prüfer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J.** 2008. PatMaN: rapid alignment of short sequences to large databases. Bioinformatics **24**:1530–1531.

120. **Quince C, Curtis TP, Sloan WT.** 2008. The rational exploration of microbial diversity. The ISME journal **2**:997–1006.

121. **Rajilić-Stojanović M, Heilig HGHJ, Molenaar D, Kajander K, Surakka A, Smidt H, de Vos WM.** 2009. Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environmental Microbiology **11**:1736–1751.

122. **Relógio A, Schwager C, Richter A, Ansorge W, Valcárcel J.** 2002. Optimization of oligonucleotide-based DNA microarrays. Nucleic Acids Research **30**:e51.

123. **Reymond N, Charles H, Duret L, Calevro F, Beslon G, Fayard J-M.** 2004. ROSO: optimizing oligonucleotide probes for microarrays. Bioinformatics **20**:271–273.

124. **Rimour S, Hill DRC, Militon C, Peyret P.** 2005. GoArrays: highly dynamic and efficient microarray probe design. Bioinformatics **21**:1094–1103.

125. **Rose TM, Schultz ER, Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S.** 1998. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Research **26**:1628–1635.

126. **Rose TM, Henikoff JG, Henikoff S.** 2003. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Research **31**:3763–3766.

127. **Rouchka EC, Khalyfa A, Cooper NGF.** 2005. MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrays. BMC Bioinformatics **6**:175.

128. **Rouillard J-M, Zuker M, Gulari E.** 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Research **31**:3057–3062.

129. **Rychlik W, Spencer WJ, Rhoads RE.** 1990. Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acids Research **18**:6409–6412.

130. **Ryder E, Jackson R, Ferguson-Smith A, Russell S.** 2006. MAMMOT--a set of tools for the design, management and visualization of genomic tiling arrays. Bioinformatics **22**:883–884.

131. **Sadakane K, Shibuya T.** 2001. Indexing huge genome sequences for solving various problems. Genome Inform **12**:175–183.

132. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution **4**:406–425.

133. **SantaLucia J.** 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proceedings of the National Academy of Sciences of the United States of America **95**:1460–1465.

134. **Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP.** 2010. Computational solutions to large-scale data management and analysis. Nature Reviews Genetics **11**:647–657.

135. **Schliep A, Krause R.** 2008. Efficient algorithms for the computational design of optimal tiling arrays. IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM **5**:557–567.

136. **Schliep A, Rahmann S.** 2006. Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree. Bioinformatics **22**:e424–30.

137. **Schönmann S, Loy A, Wimmersberger C, Sobek J, Aquino C, Vandamme P, Frey B, Rehrauer H, Eberl L.** 2009. 16S rRNA gene-based phylogenetic microarray for simultaneous identification of members of the genus Burkholderia. Environmental Microbiology 11:779–800.

138. **Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM.** 2000. RNA expression analysis using a 30 base pair resolution Escherichia coli genome array. Nature Biotechnology **18**:1262–1268.

139. **Severgnini M, Cremonesi P, Consolandi C, Caredda G, De Bellis G, Castiglioni B.** 2009. ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design. Nucleic Acids Research **37**:e109.

140. **Small J, Call DR, Brockman FJ, Straub TM, Chandler DP.** 2001. Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. Applied and Environmental Microbiology **67**:4708–4716.

141. **Sokal RR, Michener CD.** 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull., Vol. 38 (1958), pp. 1409-1438 **38**:1409–1438.

142. **Spruill SE, Lu J, Hardy S, Weir B.** 2002. Assessing sources of variability in microarray gene expression data. BioTechniques **33**:916–20– 922–3.

143. **Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S.** 2002. The generic genome browser: a building block for a model organism system database. Genome Research **12**:1599–1610.

144. **Stralis-Pavese N, Sessitsch A, Weilharter A, Reichenauer T, Riesing J, Csontos J, Murrell JC, Bodrossy L.** 2004. Optimization of diagnostic microarray for application in analysing landfill methanotroph communities under different plant covers. Environmental Microbiology **6**:347–363.

145. **Suzuki MT, Giovannoni SJ.** 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Applied and Environmental Microbiology **62**:625–630.

146. **Szemes M, Bonants P, de Weerdt M, Baner J, Landegren U, Schoen CD.** 2005. Diagnostic application of padlock probes--multiplex detection of plant pathogens using universal microarrays. Nucleic Acids Research **33**:e70–e70.

147. **Talla E, Tekaia F, Brino L, Dujon B.** 2003. A novel design of whole-genome microarray probes for Saccharomyces cerevisiae which minimizes cross-hybridization. BMC Genomics **4**:38.

148. **Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB.** 2003. Oligonucleotide microarray for the study of functional gene diversity in the nitrogen cycle in the environment. Applied and Environmental Microbiology **69**:1159–1171.

149. **Terrat S, Peyretaillade E, Goncalves O, Dugat-Bony E, Gravelat F, Moné A, Biderre-Petit C, Boucher D, Troquet J, Peyret P.** 2010. Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. BMC Bioinformatics **11**:478.

150. **Tewhey R, Nakano M, Wang X, Pabón-Peña C, Novak B, Giuffre A, Lin E, Happe S, Roberts DN, LeProust EM, Topol EJ, Harismendy O, Frazer KA.** 2009. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biology **10**:R116.

151. **Thompson JD, Higgins DG, Gibson TJ.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**:4673–4680.

152. **Thorsen O, Smith B, Sosa CP, Jiang K, Lin H, Peters A, Feng W-C.** 2007. Parallel Genomic Sequence-Search on a Massively Parallel System, pp. 59–68. In. ACM Press, New York, New York, USA.

153. **Tijssen P.** 1993. Hybridization With Nucleic Acid Probes, Part I: Theory and Nucleic Acid Preparation, 1st ed. Elsevier Science Ltd.

154. **Tiquia SM, Wu L, Chong SC, Passovets S, Xu D, Xu Y, Zhou J.** 2004. Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. BioTechniques **36**:664–70– 672– 674–5.

155. **Tomiuk S, Hofmann K.** 2001. Microarray probe selection strategies. Brief Bioinform **2**:329–340.

156. **van Doorn R, Slawiak M, Szemes M, Dullemans AM, Bonants P, Kowalchuk GA, Schoen CD.** 2009. Robust detection and identification of multiple oomycetes and fungi in environmental samples by using a novel cleavable padlock probe-based ligation detection assay. Applied and Environmental Microbiology **75**:4185–4193.

157. **Vijaya Satya R, Zavaljevski N, Kumar K, Bode E, Padilla S, Wasieloski L, Geyer J, Reifman J.** 2008. In silico microarray probe design for diagnosis of multiple pathogens. BMC Genomics **9**:496.

158. **Vora GJ, Meador CE, Stenger DA, Andreadis JD.** 2004. Nucleic acid amplification strategies for DNA microarray-based pathogen detection. Applied and Environmental Microbiology **70**:3047–3054.

159. **Vouzis PD, Sahinidis NV.** 2011. GPU-BLAST: using graphics processors to accelerate protein sequence alignment. Bioinformatics **27**:182–188.

160. **Wagner M, Smidt H, Loy A, Zhou J.** 2007. Unravelling microbial communities with DNA-microarrays: challenges and future directions. Microbial Ecology **53**:498–506.

161. **Wang X, Seed B.** 2003. Selection of oligonucleotide probes for protein coding sequences. Bioinformatics **19**:796–802.

162. **Wernersson R, Juncker AS, Nielsen HB.** 2007. Probe selection for DNA microarrays using OligoWiz. Nature Protocols **2**:2677–2691.

163. **Wernersson R, Nielsen HB.** 2005. OligoWiz 2.0--integrating sequence feature annotation into the design of microarray probes. Nucleic Acids Research **33**:W611–5.

164. **Wilson KH, Wilson WJ, Radosevich JL, DeSantis TZ, Viswanathan VS, Kuczmarski TA, Andersen GL.** 2002. High-density microarray of small-subunit ribosomal DNA probes. Applied and Environmental Microbiology **68**:2535–2541.

165. **Wood WI, Gitschier J, Lasky LA, Lawn RM.** 1985. Base composition-independent hybridization in tetramethylammonium chloride: a method for oligonucleotide screening of highly complex gene libraries. Proceedings of the National Academy of Sciences of the United States of America **82**:1585–1588.

166. **Wootton JC, Federhen S.** 1993. Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry **17**:149–163.

167. **Xu D, Li G, Wu L, Zhou J, Xu Y.** 2002. PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics **18**:1432–1437.

168. **Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, Pham P, Cheuk R, Karlin-Newmann G, Liu SX, Lam B, Sakano H, Wu T, Yu G, Miranda M, Quach HL, Tripp M, Chang CH, Lee JM, Toriumi M, Chan MMH, Tang CC, Onodera CS, Deng JM, Akiyama K, Ansari Y, Arakawa T, Banh J, Banno F, Bowser L, Brooks S, Carninci P, Chao Q, Choy N, Enju A, Goldsmith AD, Gurjal M, Hansen NF, Hayashizaki Y, Johnson-Hopson C, Hsuan VW, Iida K, Karnes M, Khan S, Koesema E, Ishida J, Jiang PX, Jones T, Kawai J, Kamiya A, Meyers C, Nakajima M, Narusaka M, Seki M, Sakurai T, Satou M, Tamse R, Vaysberg M, Wallender EK, Wong C, Yamamura Y, Yuan S, Shinozaki K, Davis RW, Theologis A, Ecker JR.** 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302**:842–846.

169. **Yershov G, Barsky V, Belgovskiy A, Kirillov E, Kreindlin E, Ivanov I, Parinov S, Guschin D, Drobishev A, Dubiley S, Mirzabekov A.** 1996. DNA analysis and diagnostics on oligonucleotide microchips. Proceedings of the National Academy of Sciences of the United States of America **93**:4913–4918.

170. **Zhou J.** 2003. Microarrays for bacterial detection and microbial community analysis. Current Opinion in Microbiology **6**:288–294.

171. **Zhou Z, Dou Z-X, Zhang C, Yu H-Q, Liu Y-J, Zhang C-Z, Cao Y-J.** 2007. A strategy to optimize the oligo-probes for microarray-based detection of viruses. Virol. Sin. **22**:326–335.

172. **Ziv J, Lempel A.** 1977. A universal algorithm for sequential data compression. IEEE Trans. Inform. Theory **23**:337–343.

173. **Zuker M.** 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Research **31**:3406–3415.

3. Discussion

Les biopuces à ADN représentent des outils moléculaires à haut débit de choix pour l'étude des microorganismes, que ce soit à l'échelle d'un génome ou des communautés et continuent de rivaliser avec les approches de séquençage massif. Les différentes stratégies de détermination des sondes, alliées à la puissance des outils informatiques, permettent à l'heure actuelle d'obtenir des biopuces de qualité pour répondre aux différentes problématiques d'études des microorganismes. En effet, ces stratégies de détermination de sondes sont le fruit d'une réflexion portant sur de nombreux paramètres. Il est possible de mettre en avant certains d'entre eux comme la longueur ayant un impact direct sur la spécificité et la sensibilité de la détection des cibles. Ainsi, la discrimination phylogénétique de microorganismes sur la base de leurs séquences ADNr 16S se fera avec des sondes courtes plus spécifiques, alors que des sondes longues beaucoup plus sensibles pourront être utilisées pour l'identification par exemple de gènes métaboliques au sein d'un environnement complexe. Cependant, des stratégies comme GoArrays (Rimour et al 2005) permettent d'augmenter l'efficacité des sondes sélectionnées en combinant la spécificité à la sensibilité suite à la concaténation de deux sondes courtes à l'aide d'un linker. Dans tous les cas, une attention particulière doit être apportée à la sélection des sondes et donc aux logiciels utilisés. Outre la fiabilité de l'outil de détection, un autre défi majeur pour le développement des biopuces ADN en écologie microbienne porte sur la détection de gènes non encore identifiés et donc non répertoriés dans les bases de données. Ce concept dit exploratoire a été initialement appliqué pour la détermination d'amorces PCR et son utilisation s'est étendue aux biopuces à ADN (Bontemps et al 2005, Dugat-Bony et al 2012b).

Enfin, les logiciels doivent être suffisamment performants pour réaliser tous les tests de sélection des sondes sur des masses de données en constante croissance. Ainsi, plusieurs orientations émergent avec notamment le remplacement des étapes d'alignements, la possibilité de construire des bases de données dédiées (spécifiques de l'environnement exploré) pour les tests de spécificité et l'utilisation de nouveaux moyens de calcul. Ces derniers passent par l'utilisation de machines multi-processeurs *via* les bibliothèques de fonctions « Message Passing Interface » (MPI), ou les processeurs des cartes graphiques (General-purpose Processing on Graphics Processing Units, GPGPU). De même, il est possible désormais *via* le développement des architectures de type clusters ou grilles de calcul

de distribuer ces calculs à large échelle et à différents endroits géographiques. Ces nouvelles ressources ont d'ailleurs permis la parallèlisation d'un nouvel outil MetaExploArrays (Article partie ANNEXE) sur grille de calcul et dédié à la détermination de sondes pour POA, FGA et WGA (Jaiziri et al 2012).

Ce chapitre montre donc qu'il existe une recherche bioinformatique très active dans le domaine de la conception de sondes pour biopuces à ADN.

PARTIE 3 : Développement d'un logiciel de sélection de sondes oligonucléotidiques

1. Contexte

L'essor spectaculaire des techniques moléculaires ces dix dernières années a révolutionné le domaine de l'écologie microbienne et la manière d'explorer la diversité du monde microbien. De nombreux outils, dits à haut-débit, ont été développés permettant de générer et d'analyser d'importantes quantités de données qui étaient auparavant inaccessibles, mais qui demeurent indispensables pour comprendre le fonctionnement des écosystèmes. Les nouvelles méthodes moléculaires à haut-débit comme les biopuces à ADN ou le développement de nouvelles approches de capture de gènes couplées au séquençage haut débit, démontrent être des outils pertinents pour explorer la diversité microbienne des environnements complexes (Roh et al 2010).

Cependant, l'efficacité de ces méthodes moléculaires dépend entièrement des sondes sélectionnées. Du fait de l'augmentation constante du nombre de séquences déposées dans les bases de données qui doivent être prises en compte, les logiciels de détermination de sondes doivent être de plus en plus performants. A l'heure actuelle, la plupart de ces logiciels ne permettent pas de gérer ces grands jeux de données. En effet, l'étape d'alignement multiple permettant d'identifier des régions conservées devient rapidement irréalisable. De nouvelles stratégies basées sur l'identification de la fréquence des *k*-mers ont été développées et permettent de contourner cette contrainte (Bader et al 2011, Hysom et al 2012). De plus, la détermination de sondes exploratoires est indispensable pour pouvoir appréhender la diversité non répertoriée dans les bases de données. Actuellement, seuls les logiciels PhylArray (Militon et al 2007), HiSpOD (Dugat-Bony et al 2011) et Metabolic Design (Terrat et al 2010) permettent une détermination de sondes dédiées aux biopuces phylogénétiques et fonctionnelles intégrant ce caractère exploratoire. Toutefois, ces outils nécessitent la réalisation d'alignements multiples pouvant limiter leurs utilisations.

2. Objectif

Une nouvelle stratégie pour la détermination de sondes dédiées à l'écologie microbienne a été envisagée. L'objectif était de proposer un nouveau logiciel offrant la possibilité de sélectionner des sondes oligonucléotidiques de qualité tout en intégrant le

caractère exploratoire. De plus, ce logiciel devait permettre d'effectuer la détermination des sondes à partir d'une masse importante de données et donc de s'affranchir de l'étape d'alignement multiple. Un service web et une version téléchargeable ont donc été développés apportant une plus grande flexibilité d'utilisation de l'outil. La stratégie employée repose sur (1) l'identification de mots (*k*-mers) longs (entre 18 et 31-mers) retrouvés, de manière exacte, uniquement dans le groupe ciblé (et absents d'un groupe non ciblé également fourni par l'utilisateur) ; (2) le regroupement des *k*-mers afin d'aligner les mots provenant strictement de la même région génomique. Ainsi, à partir de l'alignement multiple de chaque *cluster*, un mot consensus peut être défini, à partir duquel pourra être déduit l'ensemble des combinaisons possibles qui représentent les signatures exploratoires ou non ; (3) le test de la couverture et de la spécificité de chacune de ces signatures, évalué par rapport aux groupes ciblés et non-ciblés, en autorisant un nombre maximal de différences entre la signature et la séquence cible ou non-ciblée. Ainsi, si cette distance est fixée à deux par l'utilisateur, l'ensemble des séquences du groupe cible présentant au maximum deux différences (mismatches ou gaps) avec la signature testée sont prises en compte pour le calcul de la couverture. De même, les séquences du groupe non-cible ayant jusqu'à deux différences sont considérées comme de potentielles hybridations croisées.

L'ensemble de la stratégie mise en place a conduit au développement du logiciel KASpOD pour « K-mer based Algorithm for high-Specific Oligonucleotide Design ». Afin d'évaluer la performance de ce nouvel outil, différentes sondes utilisées pour la construction d'une biopuce phylogénétique ont été déterminées, en ciblant l'ensemble des genres procaryotes présents au sein de la base de données Greengenes (McDonald et al 2011). Il faut également noter que cet outil peut permettre la détermination de séquences oligonucléotidiques pour d'autres applications que les biopuces à ADN ou la capture de gènes comme la PCR ou encore les approches FISH.

3. Principaux résultats

Le travail réalisé a permis de proposer un nouveau logiciel de détermination de sondes oligonucléotidiques utilisant une stratégie originale de détermination des *k*-mers qui a donné lieu à une publication sous la forme d'une « Applications Note » dans le journal « Bioinformatics ». Le logiciel KASpOD est utilisable *via* une interface web (http://g2im.u-

clermont1.fr/kaspod). Pour le traitement de grande masse de données une version téléchargeable pour une utilisation en local est également disponible sur ce même site.

La performance du logiciel KASpOD a été validée *via* la détermination de sondes oligonucléotidiques de 25-mers en utilisant en entrée 252 183 séquences d'ADNr 16S extraites la base de données Greengenes définissant 1295 genres procaryotes. Au total, 22 613 sondes non chevauchantes de 25-mers couvrant l'ensemble des séquences des genres ciblés ont été déterminées. Ces sondes sont disponibles pour les différentes applications citées précédemment. Le temps de calcul nécessaire au logiciel pour la détermination des sondes est relativement court. Les calculs soumis *via* l'interface web de KASpOD sont exécutés sur une machine multi-processeurs gérée par le CRRI (Centre Régional de Ressources Informatiques). Grâce aux 138 processeurs de 2,2Ghz et 2Go de RAM chacun, il faut compter, pour des sondes de 25-mers, et aucune différence autorisée (pour la couverture et la spécificité), environ 9min pour un genre procaryote contenant un nombre restreint de séquences (685), 36min pour un jeu moyen (4733) et 53min pour un jeu de séquences plus conséquent (9528). Toutefois ce temps de calcul est dépendant de certains paramètres comme le nombre maximal de différences autorisées entre la signature et le groupe cible ou non-cible. Ainsi, si cette distance est fixée à deux, les temps de calculs précédents sont respectivement de 55min, 4,5h et 16h. D'autres paramètres influencent le temps nécessaire à la détermination de sondes comme la disponibilité du cluster de calcul. En effet, d'autres utilisateurs ont accès à cette machine et un système de file d'attente a été mis en place. Par ailleurs, la diversité de séquence au sein du fichier contenant les séquences ciblées est également un paramètre important. Une diversité importante implique des sondes plus dégénérées pour prendre en compte toute la variabilité de séquence à chaque site moléculaire. Ainsi, le nombre de sondes se retrouve augmenté et par conséquent les temps de calculs sont plus importants.

D'une manière générale, KASpOD permet la sélection de plusieurs sondes pour chaque groupe de séquences donné en entrée. Une limite de KASpOD est qu'il ne permet pas de définir des sondes d'une taille supérieure à 31-mers. Aussi, afin de disposer de sondes longues permettant d'augmenter la sensibilité, une alternative serait d'appliquer la stratégie GoArrays (Rimour et al 2005) sur les sondes définies avec KASpOD.

Article 1 : KASpOD - A web service for highly specific and explorative oligonucleotide design

Nicolas Parisot[1,3], Jérémie Denonfoux[1,3], Eric Dugat-Bony[1,2], Pierre Peyret[1,2] and Eric Peyretaillade[1,2,*]

[1] Clermont Université, Université d'Auvergne, EA 4678 CIDAM, BP 10448, F63000 Clermont-Ferrand, France.
[2] Clermont Université, Université d'Auvergne, UFR Pharmacie, Clermont-Ferrand, France.
[3] UMR CNRS 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, France.

Associate Editor: Prof. David Posada

ABSTRACT

Summary: KASpOD is a web service dedicated to the design of signature sequences using a *k*-mer based algorithm. Such highly specific and explorative oligonucleotides are then suitable for various goals, including phylogenetic oligonucleotide arrays (POAs).

Availability: http://g2im.u-clermont1.fr/kaspod

Contact: eric.peyretaillade@udamail.fr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1. INTRODUCTION

Environmental DNA microarrays, including Phylogenetic Oligonucleotide Arrays (POAs), are key technologies which are well adapted to profiling environmental communities (Dugat-Bony, Peyretaillade, et al., 2012). The extreme diversity of microorganisms, however, means that molecular community exploration or specific analysis of microbial groups are faced with a new challenge: designing group-specific probe sets which must harbour a high coverage (i.e. being able to hybridize with all the target sequences) and a high specificity, showing no cross-hybridizations with non-target sequences (Loy et al., 2008). Sensitivity (i.e. being able to detect even low abundance targets) and uniformity (i.e. uniform thermodynamic behaviours for all the probes) are also main criteria in the selection of the best probe set (Wagner et al., 2007).

*To whom correspondence should be addressed.

The development of comprehensive POAs requires integrating large datasets produced by metagenomics projects to assess the coverage and specificity of the probe set. Unfortunately, many available probe design programmes are not suitable to deal with such data (Dugat-Bony, Peyretaillade, et al., 2012). To overcome this limitation two recent strategies have been implemented (Bader et al., 2011; Hysom et al., 2012). Despite major speed improvements both strategies are still not able to define explorative probes. They only define regular oligonucleotides found uniquely in the target group, whereas explorative probes take into account the sequence variability within the target group to define new combinations not yet deposited in public databases but potentially present in the environment.

In spite of large amounts of data, our current vision of the microbial diversity is, indeed, still incomplete. This is partially explained by the tremendous diversity of microbial species, ecological niches and technological limits: detecting 90% of the richness in some complex environments could require tens of thousands of times the current sequencing effort (Quince et al., 2008). Microarrays coupled with explorative probe design strategies are, therefore, well suited to survey complete microbial communities, including microorganisms with uncharacterised sequences (Terrat et al., 2010; Dugat-Bony, Biderre-Petit, et al., 2012).

Currently, the only software dedicated to POAs which allows the design of explorative probes, is the PhylArray programme (Militon et al., 2007) which relies on group-specific alignments prior to the probe design step to identify conserved probe-length regions. Building large multiple sequence alignments, however, represents a time-consuming task which is not compatible with high-throughput data.

Here we propose KASpOD, a fast and alignment-free algorithm to detect group-covering signature sequences allowing the design of explorative probes.

2. METHODS

2.1 Usage

KASpOD takes as input a target sequence set and a database of non-target sequences. The web interface accepts two parameters to design signatures: the oligonucleotide length (18-31-mer), and the edit distance between signatures and full-length sequences to perform

specificity and coverage evaluation steps. The edit distance is defined as the total number of differences, gaps and/or mismatches, allowed between the probe and its target.

2.2 Algorithm

KASpOD consists of three computational stages (Fig. 1).

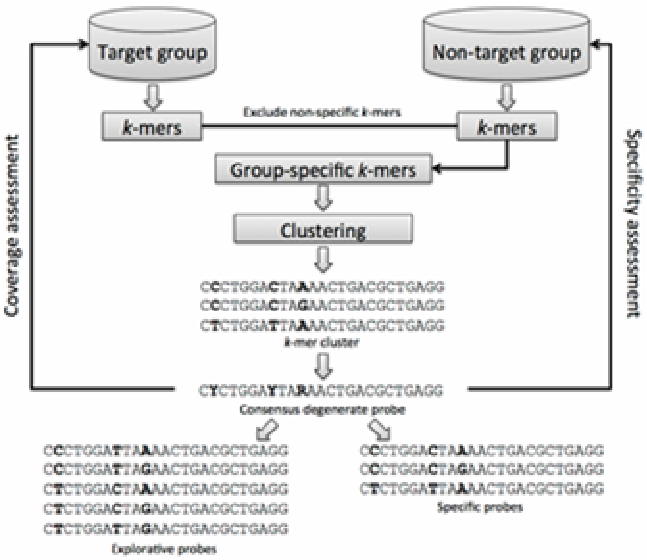

Fig. 1. The KASpOD programme workflow.

2.2.1 Search for group-specific k-mers

The first stage is the extraction of every k-mer from both the target and the non-target groups by using Jellyfish version 1.1.4 (Marcais and Kingsford, 2011). For large target groups (more than 100 sequences), a noise reduction step is performed to remove k-mers occurring only once. Every *k*-mer found in both groups is then removed from the signature candidates, as it occurs exactly in the non-target group.

2.2.2 Consensus signature sequences building

The second stage consists of clustering fully overlapping *k*-mers using CD-HIT version 4.5.4 (Li and Godzik, 2006) at an 88% identity clustering threshold. For each cluster a degenerate consensus signature is built taking into account sequence variability at each position. .

2.2.3 Coverage and specificity evaluation

The last stage performs a coverage assessment of each degenerate consensus k-mer against the target group, by using PatMaN version 1.2.2 (Prüfer et al., 2008). Coverage is computed using the number of exact or non-exact (with at most the edit distance) matches in the target group. Specificity is assessed in the same way by comparing degenerate probes against the non-target group sequences.

3. RESULTS

We used KASpOD to design 25-mer probes for 1,295 prokaryotic genera based on the recently published Greengenes taxonomy (McDonald et al., 2012) (see Supplementary Data 1 for complete procedure). Finally, 22,613 group-specific signatures were designed (Supplementary Table 2) and are freely available on the KASpOD website (http://g2im.u-clermont1.fr/kaspod/about.php). This high-quality probe set could be used to build a POA to allow monitoring of complete prokaryotic communities in complex environmental samples. The probe set was not filtered using thermodynamic calculations, in order to let the users select the entire probe set, or subset, for their own applications, such as PCR, FISH, gene capture or *in silico* for rapid sequence identification.

A runtime performance analysis of the web-service has been performed and results are available in the Supplementary Data 3.

As KASpOD does not allow the generation of probes longer than 31 nucleotides, an interesting strategy would be to combine KASpOD and GoArrays (Rimour et al., 2005) in order to concate-nate two short probes with a random linker. This approach produces oligonucleotide probes as specific as short probes and as sensitive as long ones. KASpOD could, therefore, be used for applications such as Functional Genes Array (FGAs), offering the opportunity to generate group-specific and explorative probes allowing a broad coverage of multiple sequence variants for a given gene family.

ACKNOWLEDGEMENTS

We want to thank S. Terrat and A. Mahul, for their help.

Funding: This work was supported by Direction Générale de l'Armement (DGA).

Conflict of Interest: none declared.

REFERENCES

Bader,K.C. et al. (2011) Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics*, **27**, 1546–1554.

Dugat-Bony,E., Biderre-Petit,C., et al. (2012) In situ TCE degrada-tion mediated by complex dehalorespiring communities during biostimulation processes. *Microb Biotechnol.*

Dugat-Bony,E., Peyretaillade,E., et al. (2012) Detecting unknown sequences with DNA microarrays: explorative probe design strategies. *Environmental Microbiology*, **14**, 356–371.

Hysom,D.A. et al. (2012) Skip the alignment: degenerate, multiplex primer and probe design using K-mer matching instead of alignments. *PLoS One*, **7**, e34560.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Loy,A. et al. (2008) probeCheck--a central resource for evaluating oligonucleotide probe coverage and specificity. *Environmental Microbiology*, **10**, 2894–2898.

Marcais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

McDonald,D. et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME journal*, **6**, 610–618.

Militon,C. et al. (2007) PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics*, **23**, 2550–2557.

Prüfer,K. et al. (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, **24**, 1530–1531.

Quince,C. et al. (2008) The rational exploration of microbial diversity. *The ISME journal*, **2**, 997–1006.

Rimour,S. et al. (2005) GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics*, **21**, 1094–1103.

Terrat,S. et al. (2010) Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. *BMC Bioinformatics*, **11**, 478.

Wagner,M. et al. (2007) Unravelling microbial communities with DNA-microarrays: challenges and future directions. *Microbial Ecology*, **53**, 498–506.

SUPPLEMENTARY DATA

Supplementary Data 1: Technical details about the prokaryotic oligonucleotide array (POA) construction from input data management to the probe design.

Step 1: 16S rDNA database construction

The current release of Greengenes (09-May-2011) containing 406,997 sequences was downloaded and extracted from the following URL: http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_GREENGENES_gg16S_unaligned.fasta.gz.

Then, using a PERL script, only the sequences assigned to a genus were retained for further analyses. These 310,575 sequences were then sorted by genus into different FASTA files. For each genus, a clustering step was performed at a 100% identity threshold using CD-HIT in order to remove any redundancies. Moreover, only high-quality sequences were retained:

- Sequence length greater than 1,200 nucleotides
- Less than 1% of ambiguous nucleotides (N's)

After this processing pipeline, the 16S rDNA database contained 252,250 high-quality sequences. The clustering of the whole database at high-identity thresholds (99%, 98% and 97%) coupled with manual curation, allowed us to remove potentially badly assigned sequences. Furthermore, some microbial genera were clustered together, as they were hardly distinguishable on the basis of their sequences.

Eventually, 252,183 16S rDNA sequences were fed to KASpOD to perform the probe design.

Step 2: Probe design

Each genus was then used to perform a probe design with a stand-alone version of the KASpOD software. The non-target group was composed of the 252,183 16S rDNA sequences minus the target group (*i.e.* the genus being processed).

The smallest target groups contained only one sequence (*Arhodomonas*, *Methylosphaera*, *Roseisalinus*, *Subtercola* and *Thermopallium*) with a file size of 2KB, whereas the largest was the *Corynebacterium* genus with 20,093 sequences and a file size of 33MB.

Concerning the non-target groups, the largest was composed of 252,182 sequences with a file size of 401MB and the smallest contained 232,090 sequences and had a file size of 368MB.

Each genus represents one job and computations were distributed on a multi-processor computer (40CPUs). The whole design for the 1,295 microbial genera lasted 10 days.

Step 3: Probe selection

The last stage consists of the probe set selection from the 3,242,105 candidate probes previously generated. Using a PERL script, probes were selected in order to build a probe set where each of the selected 252,183 16S rDNA sequences were covered by at least three different probes.

First step: the non-overlapping probes are selected within the probes showing no cross-hybridisations.

Second step: while there are some 16S rDNA sequences which are not covered by at least three probes, the programme selects additional probes with increasing numbers of cross-hybridisations. During this step, the programme ensures that no more than two probes show significant cross-hybridisation with the same non-targeted genus, thereby avoiding misleading interpretations of hybridisation data.

Finally, 22,613 probes were selected which could be used to build a phylogenetic oligonucleotide microarray, or for other applications (PCR, qPCR, FISH, gene capture, in silico sequence identification).

Supplementary Table 2: List of the 22,613 16S group-specific signatures designed.

http://g2im.u-clermont1.fr/kaspod/16S_Greengenes_ProbeSet.xls

Supplementary Table 3: Runtime performance analysis of the KASpOD's web-service.

Oligo length	Edit distance	Target File	Non-Target File	Time (minutes)
18	0	Small	Large	6
		Medium		32
		Large		53
			Small	39
		Large	Medium	69
			Large	53
	2	Small	Large	78
		Medium		360
		Large		778
			Small	733
		Large	Medium	832
			Large	778
25	0	Small	Large	9
		Medium		36
		Large		53
			Small	53
		Large	Medium	52
			Large	53
	2	Small	Large	55
		Medium		271
		Large		958
			Small	549
		Large	Medium	883
			Large	958
31	0	Small	Large	9
		Medium		18
		Large		53
			Small	52
		Large	Medium	54
			Large	53
	2	Small	Large	25
		Medium		201
		Large		789
			Small	575
		Large	Medium	656
			Large	789

Target Files: The small target file was composed of 685 16S rDNA sequences (1MB) belonging to the *Stenotrophomonas* genus. The medium target file was composed of 4,733 16S rDNA sequences (7.4MB) belonging to the *Faecalibacterium* genus. The large target file was composed of 9,528 16S rDNA sequences (15MB) belonging to the *Pseudomonas* genus.

Non-Target Files: The non-target files were built using a reduced personal 16S rDNA sequences database without the *Stenotrophomonas*, *Faecalibacterium* and *Pseudomonas* genera. The small non-target file was constructed by randomly taking 500 sequences out of the database (740KB). The medium non-target file was constructed by randomly taking 5,000 sequences out of the database (7.2MB). The large non-target file was constructed by randomly taking 10,000 sequences out of the database (14MB).

Nevertheless, the authors would like to emphasize that the run times are given for guidance and are dependent on many parameters (*e.g.* number of jobs on the cluster queue, heterogeneity of the target file or number of cross-hybridizations). The job status is therefore important for the user to know whether or not the job is running.

4. Discussion

A l'heure actuelle, de nombreux logiciels de détermination de sondes pour biopuces à ADN sont disponibles, mais peu d'entre eux sont appliqués pour des études environnementales (Dugat-Bony et al 2012b, Lemoine et al 2009). Le développement de nouveaux logiciels, proposant des nouvelles stratégies permettant de répondre aux exigences et aux contraintes de l'écologie microbienne, apparaît donc nécessaire. C'est avec cet objectif que le logiciel KASpOD a été développé. Il offre de nouvelles opportunités en combinant les atouts de logiciels dédiés pour la détermination de sondes pour POA comme PhylArray (Militon et al 2007), ou pour FGA comme HiSpOD (Dugat-Bony et al 2011) et Metabolic Design (Terrat et al 2010), qui intègrent le caractère exploratoire des sondes tout en optimisant la recherche des hybridations croisées potentielles. Cependant, la stratégie développée au travers de KASpOD est différente de celles proposées pour les autres logiciels. En effet, la détermination des sondes ne se fait pas à partir du résultat d'alignements multiples de séquences, limitants pour des jeux de données importants, mais à partir de la recherche de motifs nucléiques (ou *k*-mers) spécifiques des séquences ciblées. Ces k-mers sont recherchés et extraits des séquences cibles données en entrée en utilisant l'outil Jellyfish (Marçais and Kingsford 2011). Une telle approche permet de fortement réduire les temps de calcul et l'usage de mémoire. KASpOD permet également de déterminer plusieurs sondes spécifiques pour chaque groupe de séquences en entrée. L'utilisateur ayant plusieurs sondes à sa disposition, il a la possibilité de choisir la ou les meilleures sondes sur des critères thermodynamiques (Tm, %GC…), la position sur le gène ou encore la confiance accordée aux résultats obtenus (hybridations croisées) notamment pour une application de type biopuces à ADN.

Contrairement aux autres logiciels, les différents tests de couverture et de spécificité n'utilisent pas l'approche BLAST, mais PatMaN (Pattern Matching in Nucleotide databases) (Prufer et al 2008) qui est capable de rechercher rapidement et de manière exhaustive toutes les occurrences, exactes ou non, de courtes séquences nucléiques au sein d'un large jeu de données de séquences. En effet, PatMan peut récupérer les occurrences non exactes en permettant à l'utilisateur de fixer un nombre maximal de différences (gaps ou mismatches) qu'il autorise entre la séquence testée et la séquence de la base de données. De plus, cet outil est capable d'utiliser une séquence dégénérée comme requête contrairement au BLAST, qui

impose une analyse par combinaison non dégénérée. Actuellement déployable en local ou sur un cluster de calcul, KASpOD peut tout à fait évoluer vers un déploiement sur une grille de calcul permettant de réduire encore plus les temps de calcul. Le logiciel KASpOD se présente donc comme un nouvel outil performant pour disposer de sondes présentant une très bonne couverture, une grande spécificité et possédant le caractère exploratoire. Dans le cadre de l'écologie microbienne, et plus particulièrement de l'étude des environnements complexes, cette nouvelle stratégie de détermination de sondes présente toutes les qualités pour définir des sondes de qualité pour biopuces à ADN, ou pour assurer la capture de gènes.

PARTIE 4 : Développement d'une méthode innovante de capture de gènes en solution couplée à du séquençage haut-débit pour l'exploration métagénomique ciblée des environnements complexes

1. Contexte

L'émergence des nouvelles techniques de séquençage (NGS) permet à l'heure actuelle d'étudier directement l'ADN total extrait d'un environnement (métagénome) sans passer par la construction de banques de clones nécessaire au séquençage par la méthode de Sanger (Edwards et al 2006). Ces NGS (Ansorge 2009, Glenn 2011, Metzker 2010, Shendure and Ji 2008) offrent de nouvelles opportunités pour explorer et étudier les communautés microbiennes jusqu'alors non cultivées et non caractérisées au sein des environnements complexes (Eisen 2007, Eisen et al 2008, Sogin et al 2006, Venter 2004).

Cependant, une exploration des environnements complexes dans leur globalité, nécessitent un effort de séquençage très important, dépassant les capacités actuelles des NGS (Quince et al 2008). De plus, la quantité importante de données générées, la longueur des lectures encore limitée (de 20 bases à 1kb avec le développement récent des NGS de troisième génération) ou le taux d'erreur de séquençage restent des problèmes majeurs notamment pour assurer l'assemblage des séquences et permettre la reconstruction de génomes ou de grandes régions d'ADN (Hoff 2009). A l'heure actuelle, l'utilisation de ces nouvelles technologies reste encore limitée pour explorer finement les environnements complexes et coûteuse pour de nombreuses structures de recherche (Bentley 2006, Roh et al 2010). Une alternative intéressante serait donc de pouvoir réduire la complexité des échantillons métagénomiques sans passer par la PCR, source importante de biais, en enrichissant spécifiquement les séquences nucléiques d'intérêt. Suite à cet enrichissement, les approches NGS pourront alors être mises en œuvre.

2. Objectif

Afin de proposer une nouvelle alternative en écologie microbienne pour l'étude des environnements complexes en lien avec l'essor des nouvelles méthodes de séquençage, l'objectif de ce travail a été de développer une nouvelle méthode de capture de gènes, utilisant des sondes sensibles, spécifiques et exploratoires, combinée au séquençage de deuxième génération. Actuellement, aucune approche de capture de gènes utilisant des sondes n'a été

appliquée sur des échantillons métagénomiques. Cette méthode, basée sur la capture de gènes en solution, représente une nouvelle approche moléculaire en écologie microbienne permettant de réduire la complexité des métagénomes étudiés et donc d'assurer une exploration ciblée des communautés microbiennes d'intérêt. Cette approche, tout en permettant d'explorer de manière exhaustive la diversité génétique de gènes d'intérêt, présente l'avantage d'assurer l'identification des régions flanquantes associées aux séquences ciblées. Il est alors possible, par la caractérisation de grandes régions d'ADN, de mettre en évidence de nouvelles organisations génomiques voire de reconstruire de nouveaux opérons et donc d'identifier de nouveaux gènes pouvant avoir un rôle dans une voie métabolique donnée. Il faut également noter que contrairement à la PCR qui nécessite l'identification de deux régions conservées pour définir deux séquences oligonucléotidiques, une seule peut être suffisante pour cette approche.

Afin d'évaluer l'efficacité de cette nouvelle méthode, elle a tout d'abord été appliquée en ciblant et en enrichissant le gène codant pour la méthyl-coenzyme M réductase (*mcrA*) directement à partir du génome de la souche *Methanosarcina acetivorans* C2A. Par la suite, elle a été utilisée pour explorer la diversité des communautés méthanogènes impliquées dans la production de méthane au niveau de la zone anoxique d'un lac méromictique.

3. Principaux résultats

Les travaux ont conduit à la rédaction d'une publication soumise dans la revue « DNA research ». Cette étude s'est inscrite dans une problématique méthodologique, c'est-à-dire proposer un outil efficace et pertinent pour l'étude de la diversité des microorganismes des environnements complexes, et également biologique en relation avec la production de méthane.

La validation de la méthode a été réalisée premièrement en enrichissant spécifiquement le gène *mcrA* (~1,6 kb) au sein du génome de la souche *Methanosarcina acetivorans* C2A (~5,8 Mb) en utilisant un jeu de six sondes déterminées par le logiciel HiSpOD (Dugat-Bony et al 2011) et ciblant différentes régions du gène. Suite à la capture, les cibles ont été clonées puis séquencées et analysées par qPCR. Suite aux deux cycles de capture 100% des séquences piégées et séquencées correspondent au gène *mcrA*. Ces résultats sont confirmés par l'approche de PCR quantitative qui montre un enrichissement de 461 et de 175 365 respectivement pour le premier et le deuxième cycle de capture. Ces résultats traduisent l'efficacité de la méthode pour enrichir spécifiquement le gène *mcrA* à partir du génome de la

souche étudiée. Une deuxième validation a été réalisée en utilisant un jeu de 26 sondes ciblant toute la diversité du gène *mcrA*/*mrtA* présente dans les bases de données et permettant l'identification de variants géniques encore non identifiés. Ces sondes ont été utilisées pour étudier l'ADN métagénomique extrait de la zone anoxique du lac Pavin abritant des communautés d'archées méthanogènes. Sur les dix fragments capturés et clonés, cinq correspondent au gène *mcrA* avec des similarités significatives (99%) avec des séquences isolées auparavant par PCR au sein de ce même écosystème. De plus, ces séquences ont permis d'avoir accès aux régions flanquantes du gène *mcrA*, avec des portions couvrant les gènes *mcrG* et *mcrC* (gènes de l'opéron codant pour la méthyl-coenzyme M réductase) ou mettant en évidence un gène (*fmdC*) adjacent à l'opéron *mcr* et impliqué dans la méthanogénèse hydrogénotrophe. Ces résultats mettent en avant le potentiel de l'approche pour enrichir significativement l'ADN métagénomique en séquence *mcrA*, mais également pour capturer de grandes régions génomiques permettant d'explorer les régions adjacentes du gène ciblé.

Afin de d'évaluer la pertinence de l'approche capture de gènes, celle-ci a été comparée à une approche métagénomique directe et à une approche PCR utilisant des amorces universelles du gène *mcrA*. Environ 100 000 lectures pour chaque approche ont été générées puis traitées pour être au final regroupées au sein d'OTUs à un seuil de 91% au niveau protéique. Une diversité totale de 58 OTUs a été observée avec seulement 1 OTU provenant de l'approche métagénomique directe, 40 OTUs de l'approche PCR et 44 OTUs de l'approche capture. L'analyse phylogénétique a montré que toutes les séquences identifiées par l'approche PCR étaient affiliées au niveau de trois ordres différents alors que l'approche capture a permis de caractériser des séquences correspondant à ces trois mêmes ordres, mais également à celui des *Methanobacteriales*. Ces résultats montrent donc une évaluation plus exhaustive de la diversité par l'approche capture en comparaison avec l'approche PCR l'approche métagénomique directe. De plus, une approche d'assemblage des séquences issues de la capture a permis de reconstruire des contigs permettant d'explorer les régions flanquantes du gène *mcrA*. Il a ainsi été possible d'identifier la séquence de gènes adjacents mais également de mettre en évidence une organisation génétique encore jamais décrite chez les archées méthanogènes. Ces résultats soulignent la pertinence de l'approche capture de gènes pour explorer la diversité des communautés microbiennes d'intérêt, et ceci de manière plus complète que ne le permettent les approches moléculaires classiques comme la PCR. Grâce à cette étude, il a pu aussi être montré que l'approche peut être facilement couplée à des

approches de séquençage massif pour évaluer la diversité totale d'un écosystème et/ou pour assurer la reconstruction de grandes régions génomiques. Celles-ci peuvent mettre en lumière de nouvelles organisations génétiques traduisant éventuellement l'existence d'adaptations métaboliques particulières chez les microorganismes étudiés ou d'identifier de nouveau gènes potentiellement impliqués dans les voies métaboliques ciblées.

Article 2 : Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration

Jérémie Denonfoux[1,3,5,†], Nicolas Parisot[1,3,5,†], Eric Dugat-Bony[1,2], Corinne Biderre-Petit[4,5], Delphine Boucher[1,2], Diego P. Morgavi[6], Denis Le Paslier[7,8,9], Eric Peyretaillade[1,2] and Pierre Peyret[1,2, *]

[1] *Clermont Université, Université d'Auvergne, Centre de Recherche en Nutrition Humaine Auvergne, EA 4678, Conception, Ingénierie et Développement de l'Aliment et du Médicament, BP 10448, 63000 Clermont-Ferrand, France.*

[2] *Clermont Université, Université d'Auvergne, UFR Pharmacie, 63000 Clermont-Ferrand, France*

[3] *Clermont Université, Université Blaise pascal, 63000 Clermont-Ferrand, France*

[4] *Clermont Université, Université Blaise Pascal, Laboratoire Microorganismes : Génome et Environnement, BP 10448, 63000 Clermont-Ferrand, France*

[5] *UMR CNRS 6023, Université Blaise Pascal, 63000 Clermont-Ferrand, France*

[6] *INRA, UR1213 Herbivores, Site de Theix, 63122 St-Genès-Champanelle, France*

[7] *CEA, DSV, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, 91057 Evry, France*

[8] *CNRS, UMR8030, 91057 Evry, France*

[9] *UEVE, Université d'Evry, 91057 Evry, France*

[†] These authors contributed equally to this study

[*] To whom correspondence should be addressed: Pierre Peyret, EA4678 CIDAM, 28 place Henri Dunant, 63001 Clermont-Ferrand; Email: pierre.peyret@udamail.fr; Tel: +33 473 178 308; Fax: +33 473 275 624

Running title: Sequence capture method for metagenome exploration

Abstract

Next-Generation Sequencing (NGS) provides faster acquisition of metagenomics data but complete exploration of complex ecosystems remains difficult due to the extraordinary diversity of microorganisms. To reduce the environmental complexity, we present a method based on the Solution Hybrid Selection (SHS) principle combined with NGS to characterise DNA fragments harbouring biomarkers of interest. Enrichment performance was evaluated both on a Methanosarcina strain and a metagenomic sample from a meromictic lake, by capturing fragments containing the methyl coenzyme M reductase subunit A gene (mcrA), the biomarker of the methanogenesis. Methanogen diversity was compared with random-shotgun and mcrA-based amplicons pyrosequencing strategies. The SHS approach allowed the capture of DNA fragments reaching 2.5kb with an enrichment efficiency of between 41 and 100%, depending on the sample complexity. With the same sequencing effort, SHS detected a broader mcrA diversity and allowed efficient detection of the rare biosphere. By reconstructing contigs, we identified novel genetic organisations around the mcrA biomarker. This method will be helpful to explore phylogenetic and functional diversity in metagenomic samples and reveal gene linkage useful for better understanding of adaptive processes.

Keywords: α-subunit of the methyl-coenzyme M reductase / metagenomics / sequence capture / 454 pyrosequencing / microbial diversity

1. Introduction

Microorganisms play a crucial role in biosphere functioning. They represent a very diverse group of organisms living on earth [1,2]. Although the study of isolated species over the last century has delivered large and interesting results about microbial genetics, physiology, biotechnology and molecular biology, the diversity and structure of complex microbial communities are still poorly understood. This is historically due to our inability to culture most microorganisms using standard microbiological techniques [1,3]. Consequently, while there are probably millions of bacterial species on the planet, to date only a few thousand have been formally described [4].

The emergence of culture-independent techniques, such as metagenomics [5], circumvents the problem of unculturability, and transcends previous studies on individual organisms to focus on microbial communities present in an environment. Metagenomics was applied to enrich our knowledge of environmental microbiology by exploring the structural (gene/species richness and distribution) [6] and functional (metabolic) [7] profiling of complex environmental microbial communities. Based on unselective (shotgun analysis) or targeted (activity-driven and sequence-driven studies) methods, metagenomics tries to link genome information with structure, function and relationships among microbial populations [8,9].

The recent development of NGS technologies allows researchers to study genetic materials recovered from environmental samples without the preparation of a metagenomic clone library [10]. Furthermore, they allow the analysis of a greater amount of sequence information because they have a higher throughput and lower cost than other methods [11]. Nevertheless, Quince et al. [12] highlight that covering 90% of the richness in some hyper-diverse environments by NGS metagenomics, could require tens of thousands of times the current sequencing effort. In addition, the short and massive amount of metagenomic sequence reads (between 20 and 700 bases depending on the platform) can be problematic for assembling and identifying complete Coding DNA Sequence (CDS) and/or operon structure [13]. One promising alternative could be to reduce the environmental sample complexity by enriching the desired genomic target before sequencing.

Currently, several strategies of genomic-scale sequence enrichment have been reported [14]. The more efficient methods rely on complementary hybridisation of nucleic acid capture probes to the targeted DNA sequences. They use either solid phase hybridisation [15-17], or solution phase,

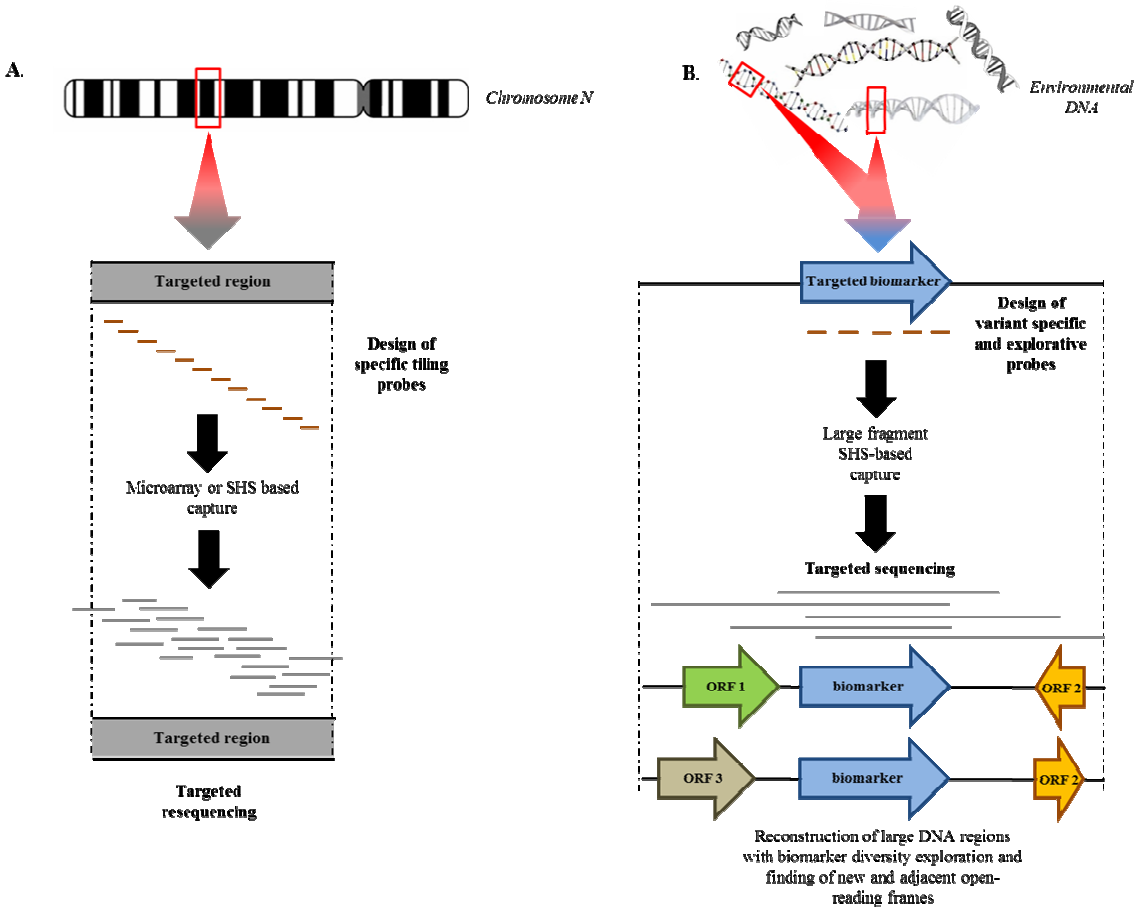

Figure 1. Schematic comparison of targeted capture methods applied to classical direct selection of individual genomic loci (human for instance) (A) and our new approach for metagenomics targeting (B). The enrichment through microarray and the SHS of large genomic regions within complex eukaryotic genomes, as described in A, uses specific tiling probes to targeted resequencing genomic loci for copy number variation (CNV) and single nucleotide polymorphism (SNP) detection. Our SHS method **(B)** uses the design of specific variants and explorative probes across a targeted biomarker to enrich specifically large DNA fragments from complex metagenomic DNA. Captured DNA fragments are sequenced in order to explore biomarker diversity and adjacent flanking regions. The red rectangles indicate the targeted regions

also known as Solution Hybrid Selection SHS [18,19], with the aim of ascertaining genetic variation by specifically enriching and resequencing regions from complex eukaryotic genomes.

To the best of our knowledge, only high throughput enrichment methods based on Polymerase Chain Reaction (PCR) products have been applied to target functional genes in complex environments [20]. But none are using oligonucleotides capture probes to specifically enrich targeted genes from a complex environmental genomic DNA: therefore, we propose a new application of this methodology in the context of microbial ecology **(Fig. 1A)** to specifically capture DNA fragments harbouring known or unknown genetic biomarkers of interest **(Fig. 1B)**. We hypothesised that the use of variant specific and explorative probes [21,22] would allow a better description of the overall diversity of biomarkers (including the rare biosphere and unknown sequences), and would facilitate the discovery of new genes associated to the targeted ones via the reconstruction of adjacent DNA regions. This should lead to better diversity coverage not influenced by PCR biases, as generally encountered in amplicons sequencing [23,24]. This strategy will yield an increased sequence coverage over target regions (not limited to a specific DNA region as for PCR enrichment), and an overall lowering of the cost per target, which is not the case in a shotgun sequencing.

In the present study, we describe the first adaptation of the SHS capture method for the selective enrichment of a target-specific biomarker from a complex environmental metagenomic DNA. We chose to develop this method by targeting the alpha subunit of the methyl coenzyme M reductase (*mcrA*) gene coding for the enzyme involved in the final step of methanogenesis, arranged in a single transcriptional unit designating the mcr operon which is highly conserved among all methanogens [25,26]. We surveyed its diversity in a permanently stratified crater lake located in the French Massif Central (Lake Pavin), where both the sediments and the anoxic water column contribute to methane production [27]. To highlight the broad benefit of this new method, compared to the more classical ones, the pyrosequencing of products from three methods was assessed: the present SHS method, a classical random-shotgun metagenomic approach, and *mcrA*-targeted amplicon sequencing survey, were applied to the same environmental sample.

	Primers / Probes (5' - 3')	Name	Size (bases)	Target	Reference
PCR	TAYATGTCNGGYGGTGTHGG	MM_01	20	*mcrA*	Mihajlovski et al. 2008
	ACRTTCATNGCRTAGTTNGG	MM_02			
	CCATCTCATCCCTGCGTGTC	454 Ti-A	20	Pyrosequencing adaptators	Roche Applied Science
	CCTATCCCCTGTGTGCCTTG	454 Ti-B			
	CCATCTCATCCCTGCGTGTCTCCGACGACTACACGACGACTTAYATGTCNGGYGGTGTHGG	Fusion primers	61	adapter A – RL001 MID – MM_01	This study
	CCTATCCCCTGTGTGCCTTGGCAGTCGACTACRTTCATNGCRTAGTTNGG		50	adapter B– MM_02	
	CGATGTCCATCAGGCCCGA	50-68-Forwrad	19	*mcrA-find* spanning fragment	This study
	AGCTCGAAGTGGAAGGCACAA	97-117-Reverse	21		
Capture	TCTGGCTCGGATCCTACATGTCCGGTGGTGTCGGGTTCACCCAGTATGCA	P1	50	*mcrA*	This study
	CTGGTCTCTCCGGCTGGTACCTCTCCATGTATGTCCACAAGGAAGCATGG	P2			
	TGAAGACCACTTCGGTGGATCCCAGAGAGCAACCGTGCTCGCAGCTGCAT	P3			
	TCGGTCACCTCTCAGACATCGTCCAGACAAGCCGTGTATCCAAAGACCCC	P4			
	AAATTCCTGAGACTCGGCCCTGAACAGGATGCAAGAAAGCAGGAAATGAT	P5			
	CGATGATGCACATGGGTGCCCCTCTCGGTCAGCGTGCAATCACTCCTTAC	P6			
Pyrosequencing	ACACGACGACT	RL001 MID	11	-	Roche Applied Science
	ACACGTAGTAT	RL002 MID			
	ACACTACTCGT	RL003 MID			
	CCATCTCATCCCTGCGTGTCTCCGACGACT	A adapter-key	30	-	Roche Applied Science
	CCTATCCCCTGTGTGCCTTGGCAGTCGACT	B adapter-key			

Table S1. Primer and probes sets used for *mcrA* gene surveys

Probe name	Sequence (5'-3')
mcrA M. kandleri (1)	TCTACGACCAGATCTGGCTAGGATCCTACATGTCAGGAGGTGTCGGTTTC
mcrA M. paludicola (1)	TGTATGACCAGATCTGGCTCGGCTCCTACATGTCCGGTGGTGTCGGCTTC
mrtA M. smithii (1)	TATATGATCAGGTTTGGTTAGGTTCTTACATGTCAGGAGGTGTAGGTTTC
mrtA_M. bryantii (1)	TATACGATCAGATCTGGCTCGGATCTTACATGTCTGGTGGTGTTGGATTC
mcrA M_arboriphilus (1)	TATACGACCAAATTTGGTTAGGTTCTTACATGTCTGGTGGTGTTGGATTT
mcrA_M. bryantii (1)	TTTACGACCAAATCTGGCTTGGTTCATACATGTCAGGTGGTGTAGGATTC
mcrA cluster 1 (1)	CAGTGTGGTGCATCCAACGTCTTCTCAATAAGGGGCGACGAGGGACTGCC
mcrA cluster 2 (2)	ACTGGAAATGATGAAATCGCTGATGAAATYGACCAGAGATACGTCCTTAA
mcrA cluster 3 (2)	GCTGCAGCATCTGCATGTTCCACTGGATTTGCAACTGGAAACGCMCAAAC
mcrA cluster 4 (4)	GCAGGTGAAGCAGCAATYGCTGACTTCTCATACGCWGCAAAACACGCCGA
mcrA cluster 5 (4)	GGTAGAGTATGTGACGGYGGTACAATYTCAAGATGGTCTGCAATGCAGAT
mcrA cluster 6 (4)	TCAGTATGTATGGCAACAGGAAACTCAAATGCWGGRGTTAATGGATGGTA
mcrA cluster 7 (4)	ACAATAGCAAGATGGAGTGCWATGCAGATWGGAATGTCATTCATTACAGC
mcrA cluster 8 (8)	TGCACAAGGAAGGMTGGTCACGTCTCGGMTTCTTCGGMTACGACCTGCAG
mcrA cluster 9 (1)	CAGTATGAACAGTTCCCGACCATGATGGAAGACCACTTCGGCGGTTCCCA
mcrA cluster 10 (4)	CAGTACGAGCAGTTCCCGACSATGATGGARGACCACTTCGGCGGGTCCCA
mcrA cluster 11 (1)	ATGGCTGATATCATTCAGACAAGCCGTGTTGACCGCAGAAGATCCAGCACA
mcrA cluster 12 (2)	CCCTTGAGGTAGTCGGTGCAGGMTGTATGCTCTACGACCAGATCTGGCT
mcrA cluster 13 (1)	GTTCTGTCCTACCAGGGCGACGAAGGTCTCCCAGACGAACTCCGTGGTCC
mcrA cluster 14 (8)	TAGCAACCGAAGTTACACTTTAYRGTCTTGAMCAATATGAAGAATATCCA
mcrA cluster 15 (2)	CATTAGAACAATACGAAGAATACCCAGCTTTACTYGAAACTCACTTCGGT
mcrA cluster 16 (8)	TGTGATGGTGGTACMACWTCCCGATGGTCTGCTATGCAGATYGGTATGTC
mcrA cluster 17 (2)	GCAATGCAGATAGGGATGTCATTCATTACAGYATACAAACTCTGTGCTGG
mcrA cluster 18 (8)	TATACGATCAGATCTGGCTAGGTTCWTACATGTCAGGTGGWGTAGGWTTC
mcrA cluster 19 (4)	CGGTGGTGTCGGTTTCACCCAGTATGCAACMGCWGCATACACCGACAACA
mcrA cluster 20 (2)	ATCCGAACTACGCSATGAACGTCGGCCACCAGGGCGAGTATGCAGGCATC

Table S2. Oligonucleotide probes sequence targeting the Methyl Coenzyme M reductase subunit A gene (*mcrA*).

The 49 and 50-mers probes designed could be specific (1 oligonucleotide) or degenerated (2, 4, 8 oligonucleotides) as indicated in brackets. Probes were designed from the most conserved regions of each group determined after a clustering using ClustalW2 (http://www.ebi.ac.uk/Tools/msa/clustalw2/)

2. Materials and methods

2.1. Capture probe design and synthesis

Two sets of capture probes were designed: the first for targeting the *mcrA* gene from the *Methanosarcina acetivorans* C2A genome (GenBank accession no. AE010299) and the second for targeting the *mcrA* sequences pool from environmental samples. The first set of capture probes consisted of six high specific 50-mer probes named P1 to P6 targeting six distinct zones of the *M. acetivorans* C2A *mcrA* gene **(Table S1)**. They were designed using HiSpOD software [28]. Adaptor sequences were added at each end, resulting in 80-mer hybrid probes consisting of 5′-ATCGCACCAGCGTGT(X)50CACTGCGGCTCCTCA-3′ with X50 indicating the specific capture probe.

The second set of capture probes consisted of 26 probes (one 49-mers and twenty-five 50-mers) designed to target all *mcrA* and *mrtA* (encoding the alpha subunit of the methyl coenzyme M reductase isoform II, MCRII) **(Table S2)**.

Oligonucleotides were purchased from Eurogentec S.A. (Belgium). The RNA probe preparation was done as described by Gnirke et al. [19].

2.2. Preparation of biological samples and libraries

Two biological models were used in this study: the *M. acetivorans* C2A strain (DSM 2834) and the Lake Pavin located in the French Massif Central (45°29′74″N, 2°53′28″E). The *M. acetivorans* C2A strain was cultivated using the medium 304 (http://www.dsmz.de/microorganisms/medium/pdf/DSMZ_Medium304.pdf) according to the manufacturer's instructions. Genomic DNA (gDNA) from the strain was extracted using the Easy DNA kit (Invitrogen) whereas environmental DNA was extracted from 350mL of freshwater collected from Lake Pavin at 90m depth as described by Dugat-Bony et al. [28]

Libraries were prepared using Roche's GS FLX Titanium General Library Preparation Kit (Roche Applied science) according to the manufacturer's instructions, starting with 5µg of DNA sheared by nebulisation. DNA fragments were selected by size using AMPure beads (Beckman Coulter genomics). After purification, fragment end polishing, adaptor ligation (A and B adapter-keys; **Table S1**) and fill-in reaction according to the standard procedure, the

libraries were PCR-amplified with the 454 Ti-A and 454 Ti-B primers **(Table S1**). The cycle conditions were 3min at 93°C followed by 20 cycles of 15sec at 93°C, 1min at 58°C, 8min at 68°C and a final elongation step at 68°C for 6min. The amplified libraries were then purified with AMPure Beads (Beckman Coulter Genomics) and stored at -20°C until use.

For the amplicons library, *mcrA* fragments were PCR-amplified on total community DNA by using the *mcrA*-specific primer pair MM_01 / MM_02 [29] **(Table S1)**. Obtained amplicon was run on a 2% (wt/vol) agarose gel and the ~500bp-sized product was purified using a QIAquick gel extraction kit (Qiagen) followed by an AMPure Beads purification (Beckman Coulter Genomics) according to the manufacturer's instructions. DNA was quantified by fluorometry for both metagenomics and amplicon libraries, using a Quant-iT PicoGreen dsDNA assay kit (Invitrogen). The DNA quality and size distribution was assessed on an Agilent Bioanalyzer High Sensivity DNA chip (Agilent Technologies).

2.3. Hybridisation capture and elution

For each SHS-capture method library a mix of 2.5µg of salmon sperm DNA (Ambion) and 500ng of DNA library in a 7µL final volume was denatured 5 min at 95°C and held 5min at 65°C before adding of 13µL prewarmed (65°C) hybridisation buffer (10X SSPE, 10X Denhardt's Solution, 10mM EDTA and 0.2% SDS) and 6µL freshly prepared of biotinylated RNA probes (500ng). After 24h at 65°C, 500ng of washed M-280 Dynabeads coated with streptavidin (Invitrogen) were added to the hybridisation mix and incubated 30min at room temperature (RT). Beads were magnetically pulled down using a magnetic stand (Ambion) and washed once 15min at RT with 500µL of 1X SSC/0.1% SDS, followed by three 10min washes at 65°C with 500µL of prewarmed 0.1X SSC/0.1% SDS. Captured DNA was eluted with 50µL of 0.1 M NaOH for 10min at RT and purified on a QIAquick column (Qiagen) in a final volume of 20µL. A 2.5µL aliquot was subjected to 15 cycles of PCR amplification using 454 Ti-A and Ti-B primers as described above.

2.4. Sanger sequencing and data analysis

PCR products were cloned using the TOPO TA cloning kit (Invitrogen). Plasmids were screened for high-size inserts by digestion with EcoRI and then sequenced using the Sanger

method on MWG DNA sequencing services (Ebersberg, Germany). Sequences were processed and joined using the Staden package programme [30], and primer sequences were removed from paired-end consensus sequences. The *mcr* sequence data retrieved from environmental application of the SHS method to Lake Pavin, were deposited in the GenBank database under accession numbers JQ404494, JQ404495, JQ404496, JQ404497 and JQ404498, as well as the sequence of the *mcrA-fmd* spanning region fragment, under accession number JQ425691.

2.5. 454 GS FLX Titanium sequencing and data analysis

DNA samples were sequenced using the GS FLX Titanium platform at the Centre Jean Perrin, on the "GINA" platform (part of GENTYANE platform, labeled IBISA since 2009: BP 392, 63011 Clermont-Ferrand, France), according to the manufacturer's specifications. Pyrosequences were trimmed using the PRINSEQ-lite PERL script [31] for quality filtering and de-replication of reads using parameters described in the preprocessing chart (http://prinseq.sourceforge.net/Preprocessing_454_SFF_chart.pdf).

Functional assignment and enrichment performance was assessed by performing a BLASTX query [32] against a database containing 12,603 McrA protein sequences downloaded from the Genbank database (http://www.ncbi.nlm.nih.gov/) using WWW-Query (http://pbil.univ-lyon1.fr/search/query_fam.php) to perform an advanced keyword search. Reads showing >40% identity over 100 or more amino acids were considered as McrA sequences. Chimera detection was done using the UCHIME program [33] with a stringent threshold score of 5. Sequences containing possible frameshifts were identified by using the "–w 20" BLAST option and disabling low complexity filters. Amino acid sequences without frameshifts were extracted from BLAST results and only the sequences which passed this filter were chosen for further phylogenetic analysis.

The sequence data from pyrosequencing strategies were deposited in the NCBI as a Short Read Archive (SRA) project under accession no. SRA049219.

2.6. Phylogenetic analysis and tree construction

	Primers (5' - 3')	Name	Size	Target	Reference
qPCR (Enrichment calculation)	TGCAAGGGCACATGCAACAC	346-365-Forward	20	*mcrA*	This study
	TGCTGCAAATCTGGGCACTG	516-535-Reverse			
	GCTGCTTATGTGGCCTGGAT	55-174-Forward	20	*fmdA*	This study
	GCATACCGAGGCGTTCGTT	326-344-Reverse	19		
qPCR (Methanogen abundance)	CAGGCTGTTCAACCGCATTTGC	1F45_1_212-233	22	OTU78	This study
	TCAGACCTTCATCGCTTCTGAT	1R30_1_364-385	22		
	GCTTCCCGGCCGCAATGGA	1F67_1_181-199	19	OTU13	This study
	TTGACACCAGCGTTCGCGT	1R57_1_277-295	19		
	CCCAGAGAGCATCCGTTCTG	1F101_1_229-248	20	OTU9	This study
	CAAGCGTCCCCAGCCTTCC	1R29_1_338-356	19		
	TGCAACTGAAGTCACGCTCTACG	1F8_1_136-158	23	OTU2	This study
	GGACAGACCTGATGCGGCT	1R54_1_235-253	19		
	CGAGAGCCACTTCGGCGGA	1F68_1_211-229	19	OTU7	This study
	TTCCTTGTGGGCGAGCATGG	1R59_1_324-343	20		
	CTTCGGTGGTTCCCAGCGTGCAT	1F93_1_224-246	23	All	This study
	TGCAGGTCGTAGCCGAAGAAGC	1R24_1_364-385	22		

Table S3. Primers sets used for qPCR experiments

All SHS-method and metagenomics sequences related to McrA were compared to a sequence obtained from the amplicon approach based on the amino acid alignment using the ClustalW2 alignment method [34] driven by the Seaview version 4 programme [35] to select those showing at least 100 amino acid in common with this reference sequence. Only overlapping regions of remaining amino acid sequences, plus all amplicon pyrosequences and 29 McrA sequences previously identified from the same sampling depth and downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genbank/), were fed to CD-HIT [36] to assign them to Operational Taxonomic Units (OTUs) using a complete linkage clustering method at 91% cut-off value [27,37].

One representative sequence of each OTU was subsequently chosen using CD-HIT output to build a phylogenetic tree under Seaview 4 [35] using the neighbor-joining method [38,39] and then bootstrapped with 1,000 trials. Closely related sequences available from GenBank (http://www. ncbi.nlm.nih.gov/) were included in the phylogenetic trees to decipher microbial community diversity. A final tree was drawn in MEGA version 5 [40].

2.7. qPCR experiments for enrichment and methanogen abundance calculation

The assays conducted in 20µL consisted of 5µL of DNA sample or standard *mcrA* PCR product (from 5×107 copies to 50 copies, covering 8 log of dynamic range for each gene), 10µL of 2X MESA Green qPCR for SYBR assay mixture (Eurogentec S.A) and 0.2µM forward and reverse primers. They were carried out using a thermo cycling protocol with an initial step of 95°C for 5min, followed by 40 cycles of denaturation at 95°C for 15sec, annealing at melting temperature of each primer sets for 15sec and elongation at 68°C for 30sec. The samples and each point of the standard curve were quantified in triplicate. The primer sets are described in the **Table S3**. Data analysis was achieved with Realplex software version 1.5 (Eppendorf Inc.) and MxPro qPCR software 4.10d (Agilent technologies). Based on the ΔΔCt method [41], relative enrichments (R) were calculated according to $R=2^{-\Delta\Delta Ct}$. This relative quantification method established a mean Ct value comparison of the *mcrA* (target gene) to the *fmdA* (non-target gene distant from 500kbp upstream the *mcrA* gene) for the 0.5kb fragment size clone library as a calibration (ΔCt). ΔCt comparison of samples prior and after capture, referred as ΔΔCt, gave a resulting unit showing a fold change describing relative capture enrichment.

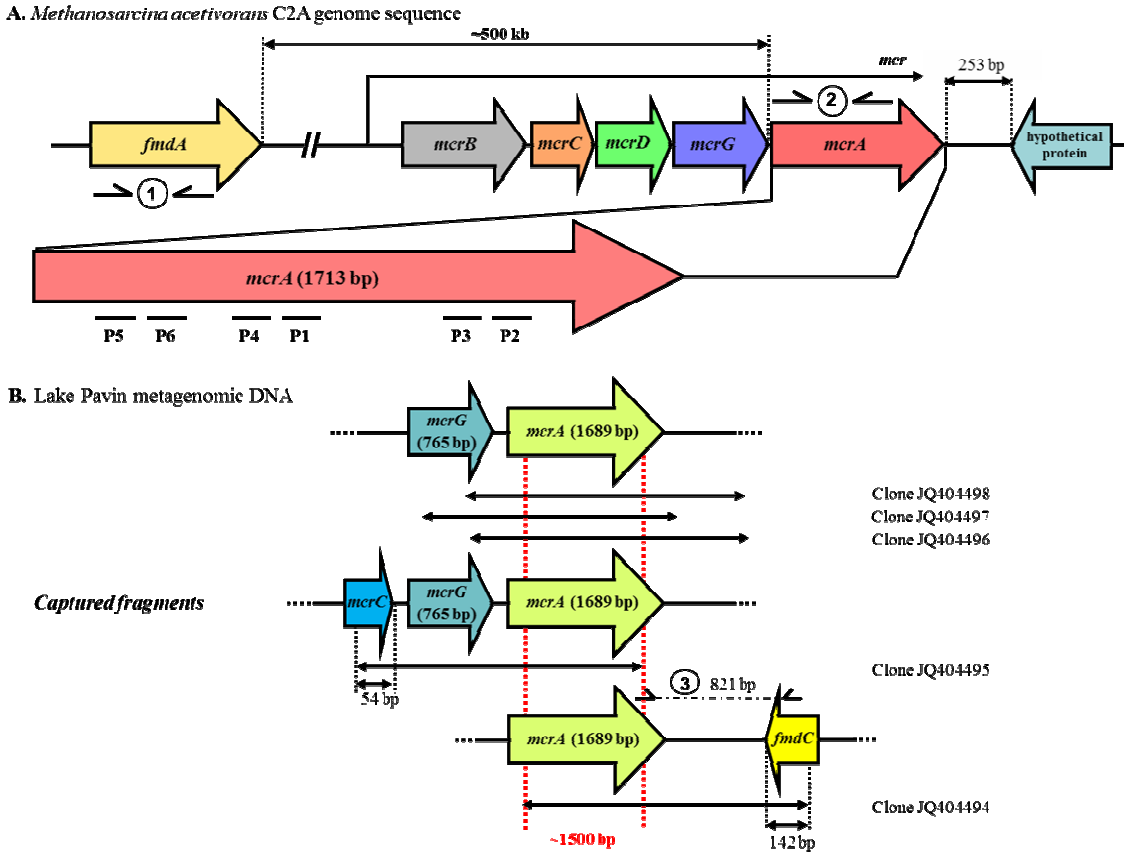

Figure 2. Schematic representation of *mcr* operon fragments on (A) *Methanosarcina acetivorans* C2A gDNA and (B) Lake Pavin metagenomic DNA. Primer pairs used for *fmdA* (1) and *mcrA* (2) quantification as well as *mcrA-fmdC* region (3) amplification are symbolized. Dashed arrows indicate the sequence coverage of each of the five clones retrieved from the environmental sample (B). P1 to P6: Positions of the six capture probes in *mcrA* gene of *M. acetivorans* (see Supplementary Table S1 for probes sequence).

2.8. SHS *de novo* read assembly

The filtered reads from the SHS were assembled with Newbler version 2.6 (Roche Applied Science) using the stringent assembly parameters of 60 bases overlap and 95% overlap identity, and using the '-rip' option to force Newbler to place each read uniquely into one contig. Functional assignment of contigs and singletons was performed by a BLASTX query [32] against our database containing the 12,603 McrA protein sequences. Prediction of *mcrA* gene location within contigs and singletons was performed using a BLASTN query [42] against the genome of *Candidatus Methanoregula boonei* 6A8 (*Methanomicrobiales* order, accession no. NC_009712), *Methanosaeta concilii* GP-6 (*Methanosarcinales* order, accession no. CP002565) and *Methanosphaera stadtmanae* DSM 3091 (*Methanobacteriales* order, accession no. CP000102) as references. Only contigs extending beyond *mcrA* of at least 100 nucleotides were kept for a new BLASTX [32] analysis against the non-redundant protein sequences database (nr) in order to identify putative open-reading frames within the flanking regions.

3. Results

3.1. Development of an SHS method for genomic-scale sequence enrichment

3.1.1. Method validation: mcrA gene enrichment from Methanosarcina acetivorans C2A genomic DNA. We performed the initial validation of our enrichment strategy by capturing the *mcrA* gene from a 1 to 3kb fragment size clone library of the completely sequenced methanogenic *Methanosarcina. acetivorans* C2A strain, using a minimal probe set spanning different non-overlapping regions of the gene **(Fig. 2A)**. The qPCR reactions revealed a relative enrichment in *mcrA* sequences by at least a factor of 461 times after the first cycle of capture and of at least 175,365 times after the second. Furthermore, as the *M. acetivorans* C2A genome has a size close to 6Mbp (5,751kbp) and hosts a single *mcrA* gene copy, the probability of randomly sequencing this gene from a 1 to 3kb fragment size clone library, should be approximately in the range 0.02 to 0.05%. By using our solution-based DNA capture-enrichment method, and working on this isolated species, the likelihood could increase to 7.8 to 23% after the first cycle, and should reach 100% after the second.

The DNA sequence of fragments retrieved after the second cycle of capture was controlled by the cloning-sequencing method. Six clones were sequenced and all had a perfect correspondence to the *mcrA* gene from *M. acetivorans* C2A, reinforcing the efficiency of the two iterative cycles of capture. Sequence assembly of captured fragments yielded a contig of 1,834bp containing the nearly complete *mcrA* gene (1,645bp) and its 3' non-coding region (189bp). After validating this approach, we further tested the performance of the method by enriching *mcrA* sequences from a complex methanogenic freshwater environment.

3.1.2. Environmental application: mcrA sequence enrichment from a methanogenic lacustrine environment (Lake Pavin). The freshwater sample used for the construction of the environmental DNA library was collected in the anoxic zone at 90m depth, where it encountered the highest methanogen diversity in such a lacustrine environment [27]. The *mcrA* sequence enrichment was conducted using an improved probe set, covering all known *mcrA* sequences and able to target new variants with explorative probes **(Table S2)**. The efficiency of the *mcrA* enrichment was controlled by the cloning and sequencing of the second capture product. Five out of the ten sequenced clones with the largest inserts (size ranging from 2,041 to 2,493bp) included *mcrA* sequences. All positive clones showed a ~1,500bp common zone corresponding to the *mcrA* gene, but also harboured upstream or downstream regions containing other genes **(Fig. 2B)**. BLAST analysis of *mcrA* sequences against the NCBI nr database revealed that they are very similar (99% similarity) to *mcrA* sequences previously retrieved from this ecosystem (accession nos. GQ389949, GQ389912 and GQ389806) [27]. The closest relative to *mcrA*, *mcrG* and partial *mcrC* sequences from cultured methanogen, belonged to *Candidatus Methanoregula boonei* 6A8 (>85%, 84% and 81% similarity respectively): this is an hydrogenotrophic species belonging to the Methanomicrobiales order and first isolated from an acidic peat bog [43]. Furthermore, the *fmdC* gene fragment identified 821 bp downstream the target gene **(Fig. 2B)** shared 77% identity with the subunit C of the formyl methanofuran dehydrogenase gene of this species. This gene has been located on the reference genome (GenBank: CP000780.1) at almost 300kbp from the *mcr* operon. It should be noted that such an organisation with the *fmd* operon located just downstream from the *mcr* operon, has never yet been described in methanogen genomes. In order to exclude the possibility of chimera formation during metagenomic library amplification, a PCR fragment spanning the *mcrA–fmdC* region was obtained directly from the initial metagenomic DNA sample, using two specific primers **(Fig. 2B, Table S1)**. The sequencing of the 821bp PCR

Table 1. Summary statistics from 454 pyrosequencing

	Metagenome	Amplicons	SHS
Total number of raw reads	136,256	121,665	177,977
Number of reads after pre-processing	116,365	119,437	122,772
Average length of cleaned reads (bases)	471	414	454
mcrA homologous sequences[1]	3	119,409	50,727
Enrichment performance (%)	0.003	99.98	41.32
Number of chimera	0	150	30
Number of reads containing frameshifts	1	80,390	21,855
Number of high-quality *mcrA* homologous sequences (without chimera and frameshifts)	2	38,869	28,842
McrA sequences used for methanogenic diversity and abundance (comparison of a common region)	1	38,807	11,442
Number of OTUs	1	40	44
McrA sequences related to OTUs	1	38,784[2]	11,324[2]
Relative abundance of *mcrA* sequences affiliated to *Methanomicrobiales* order (%)	0	98.57	98.82
Relative abundance of *mcrA* sequences affiliated to *Methanosarcinales* order (%)	0	0.005	0.86
Relative abundance of *mcrA* sequences affiliated to Novel order (%)	100	1.43	0.13
Relative abundance of *mcrA* sequences affiliated to *Methanobacteriales* order (%)	0	0	0.19

[1]BLASTX parameters: percentage of identity: 40 %; E-value cut-off: 10
[2]McrA sequences related to OTUs containing more than one sequence

product (JQ425691) confirmed the organisation revealed by the SHS method (100% identity with the captured DNA fragment).

Our results highlighted the efficiency of the capture method to enrich targets out of a complex environmental genomic mixture, and the capacity to extend beyond the initial targeted biomarker gene sequence. Additionally, the SHS method coupled with NGS technologies was used to assess the depth of coverage of the archaeal *mcrA* diversity from a complex ecosystem.

3.2. Metagenome exploration using large genomic-scale sequence enrichment coupled to NGS

The benefit of the SHS method in terms of diversity coverage, compared to more classical approaches, was further examined by sequencing the capture product of our SHS method. A new random-shotgun DNA metagenomic library adapted for pyrosequencing (fragment sizes around 500bp) was prepared for the SHS and for direct sequencing (shotgun metagenomics approach). From the same metagenomic DNA sample, *mcrA* PCR products were also amplified with the primer set MM_01-MM_02 [29]. Sequencing (captured DNA fragments, metagenome and amplicons) was performed using the 454 GS FLX Titanium technology, generating a slightly different amount of raw data with an average length ranging from 414 to 471 bases. After pre-processing, their number was almost the same for all three approaches **(Table 1)**.

3.2.1. Functional assignment and enrichment performance. Only three reads (0.003% of total reads) from the random-shotgun sequencing approach corresponded to the *mcrA* gene, whereas for the SHS method 50,727 reads were identified as *mcrA* sequences (41.32% of reads) and almost all sequences for the amplicons approach (119,409 reads, being 99.98%).

For *mcrA* diversity evaluation, however, we only analysed high-quality sequences (no chimera, no frameshift), and all the problematic reads were subsequently excluded.

3.2.2. Methanogen diversity and abundance. In total, this concerned 1 read from metagenomics, 11,442 reads from the SHS method and 38,807 reads from amplicons. Furthermore, 29 additional sequences (referred as Pavin90m) from a previous study isolated at 90m depth and produced with the same PCR primer set 27 were included in the analysis.

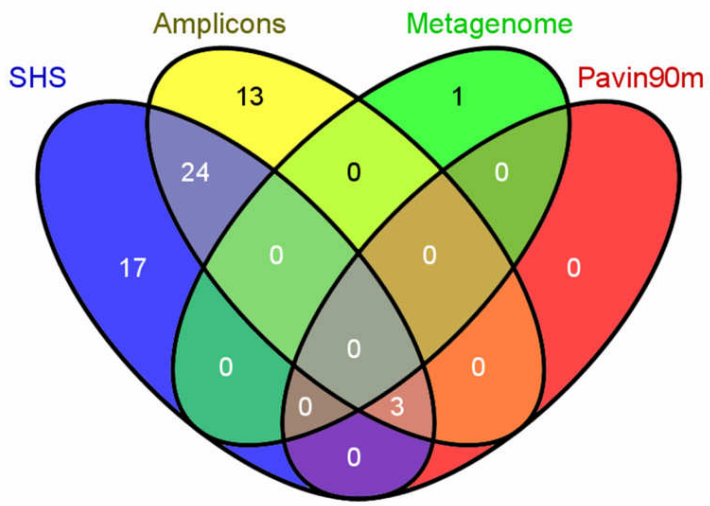

Figure 3. Venn diagram showing the number of unique and shared OTUs for in-solution capture method (SHS), PCR based strategy (Amplicons), and sequences isolated at 90m depth from a previous PCR-based study on Lake Pavin (Pavin90m) [27]. Venn diagram was processed using Venny (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

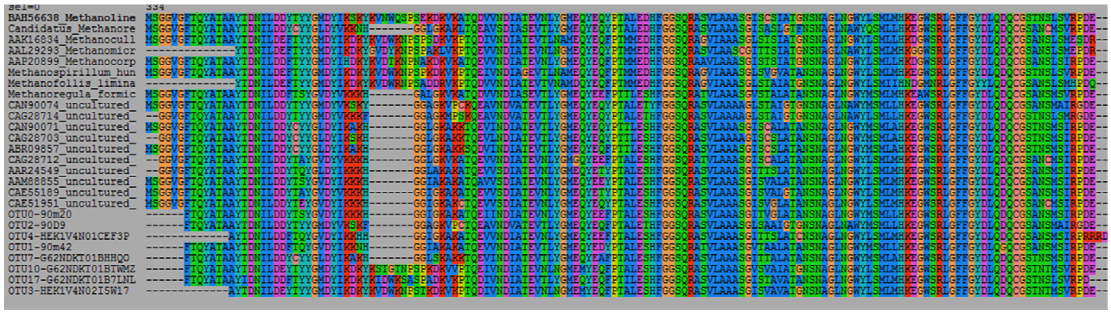

Fig. S1. Proteic alignment showing insertions events within the *mcrA* gene between cultured methanogen species and OTUs 10, 17 and 3 belonging to *Methanomicrobiales* order

Following the clustering method, 127 distinct OTUs were observed, of which 58 were kept for further detailed phylogenetic analysis as they contained more than one sequence, including also the metagenome which contained a single final read. Among these 58 OTUs, 44 were detected from the SHS method, 40 from the amplicons approach, 1 from metagenomics and 3 from Pavin90m sequences. The SHS method and amplicons shared 27 OTUs, including the 3 determined from Pavin90m sequences **(Fig. 3)**. In contrast, the remaining 31 OTUs were specific to a single method: 1 for the metagenome, 17 for the SHS and 13 for the amplicons **(Fig. 3)**.

The 58 OTUs covered four lineages encompassing *Methanobacteriales*, *Methanomicrobiales*, *Methanosarcinales*, and a putative fourth lineage referred as Novel Order. Most were closely related to the *Methanomicrobiales* order (48 OTUs, 98.6% of the total input sequences). Of the 48 OTUs, the OTU3, OTU10 and OTU17 formed a distinct branch within this cluster **(Fig. 4)**, and they were closely related to cultured methanogens species which also show an insertion event within their McrA protein sequence **(Fig. S1)**. Both the SHS and amplicons strategies clustered sequences in the most abundant OTUs **(Fig. 5)**. These abundant OTUs represented respectively 94% and 98% of the total sequences for each approach. The *Methanosarcinales* (2 OTUs; **Fig. 6A**) grouped into two distinct branches related to the reference acetoclastic species *Methanosaeta concilii* GP6 (85 and 87% similarity with OTU9 and OTU18 respectively). The most abundant was the OTU9 clustering, with 0.83% of the total SHS reads versus 0.005% for the total amplicons reads **(Fig. 5)**. In contrast, the putative Novel Order (5 OTUs; **Fig. 6B**) was dominated by OTU13 clustering with 1.39% of the total amplicons sequences, versus 0.13% for the total SHS reads **(Fig. 5)**. Despite the substantial sequencing effort for amplicons, no sequences belonging to the Methanobacteriales order were recovered from this approach. These sequences were obtained only from the SHS sample **(Fig. 6C)** where they were clustered in 3 OTUs with one showing 90% similarity to MrtA sequences (MCR isoenzyme encoded by the mrt operon) of *Methanosphaera stadtmanae* DSM 3091 [44], and the remaining two showing 77 to 79% identity to MrtA sequences of *Methanobacterium lacus* affiliated to the *Methanobacteriales* order and isolated from the sediments of Lake Pavin [45]. These sequences represented 0.19% of total SHS *mcrA*-related sequences, with the most abundant OTU78 clustering 0.11% of the total SHS reads **(Fig. 5)**.

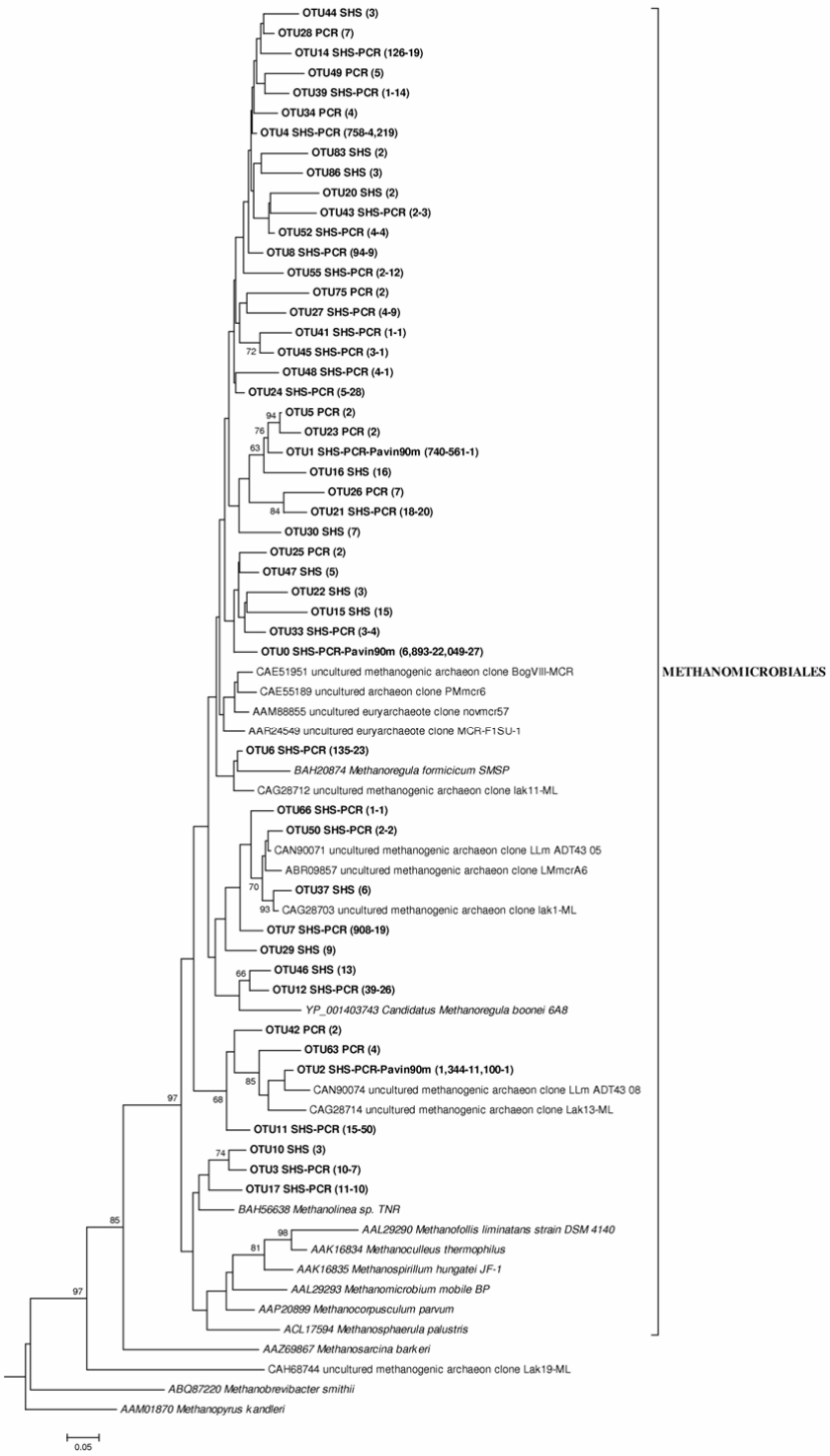

Figure 4. Phylogenetic analysis of deduced McrA amino acid sequences obtained from PCR, SHS and Pavin90m strategies showing evolutionary distance within the order *Methanomicrobiales.* Evolutionary history was inferred using the neighbour-joining method [38,39] (NJ, Poisson distance model) using Seaview software [35]. The final tree was drawn in MEGA 5 [40]. The bars represent a 5% sequence divergence. Numbers at the nodes represent bootstrap values >60% (1,000 resamplings). The number of amino acid sequences assigned to each OTU is given in brackets, together with the name of the strategies for obtaining them. McrA amino acid sequence from *Methanosarcina barkeri* (AAZ69867), uncultured methanogenic archaeon clone Lak19-ML (CAH68744), *Methanobrevibacter smithii* (ABQ87220) were used as outgroups and *Methanopyrus kandleri* (AAM01870) as an outgroup for rooting the tree. Bold arrows indicate dominant OTUs.

The GC content on the *mcrA* genes ranging from 50.4 to 61.1% for amplicons and from 37 to 63.2% for SHS, in comparison with the *mcrA* database ranging from 36.2 to 67.2% indicates that SHS is probably less affected by GC composition than PCR approach.

In parallel, qPCR based experiments were performed to precisely describe the methanogen abundance in Lake Pavin in relation with the most abundant OTUs from the different orders when specific primers could be determined **(Table S3)**. Results were compared against the relative sequences abundance obtained previously for the selected OTUs with amplicons and SHS. The OTU2 selected encompassing the *Methanomicrobiales* order showed a similar abundance pattern between qPCR and amplicons (33.5% and. 28.62%) in contrast to SHS (11.87%). In contrast, for the second Methanomicrobiales OTU (OTU7), SHS relative abundance (8.02%) was closest to qPCR result (3.6%) while amplicons sequences clustered within this OTU were detected at only 0.05%. The same trend is observed for the OTU9 corresponding to the *Methanosarcinales*. No significant difference could be observed for OTU13 (Novel Order). Finally, no amplification during qPCR experiment could be obtained for the OTU78 belonging to *Methanobacteriales*. However, we succeeded to validate the presence of this OTU in lake Pavin by successive PCR cycles followed by PCR products cloning and sequencing (100% identity). This result indicates that *Methanobacteriales* could be rare in this ecosystem.

3.2.3. De novo assembly of SHS reads. In order to reconstruct contigs and demonstrate the ability of our method to explore the genetic organisation adjacent to the targeted *mcrA* gene, de novo assembly was performed using pyrosequencing reads obtained by the SHS method **(Table 2)**. We identified 693 contigs harbouring *mcrA* genes ranging from 301 to 1639 bases. Diversity analysis showed a similar distribution to the previously described analysis obtained for unassembled sequences (data not shown). By mapping the sequences on complete reference genomes belonging to *Methanomicrobiales*, *Methanosarcinales* and *Methanobacteriales* orders (no genome was available for the Novel Order), we identified contigs extending in the *mcrA* flanking regions **(Fig. 7)**. Upstream sequences were all identified as a part of the *mcrG* gene. We also characterized two adjacent ORFs located at 200 bases downstream from the *mcrA* gene on the same orientation coding respectively for a DtxR family iron (metal) dependent repressor and a DOMON domain-containing protein. The DtxR sequences were closely related (76 to 83% identity) to *Methanosphaerula palustris* E1-9C (accession no. ACL16981) belonging to the *Methanomicrobiales* order. On the reference

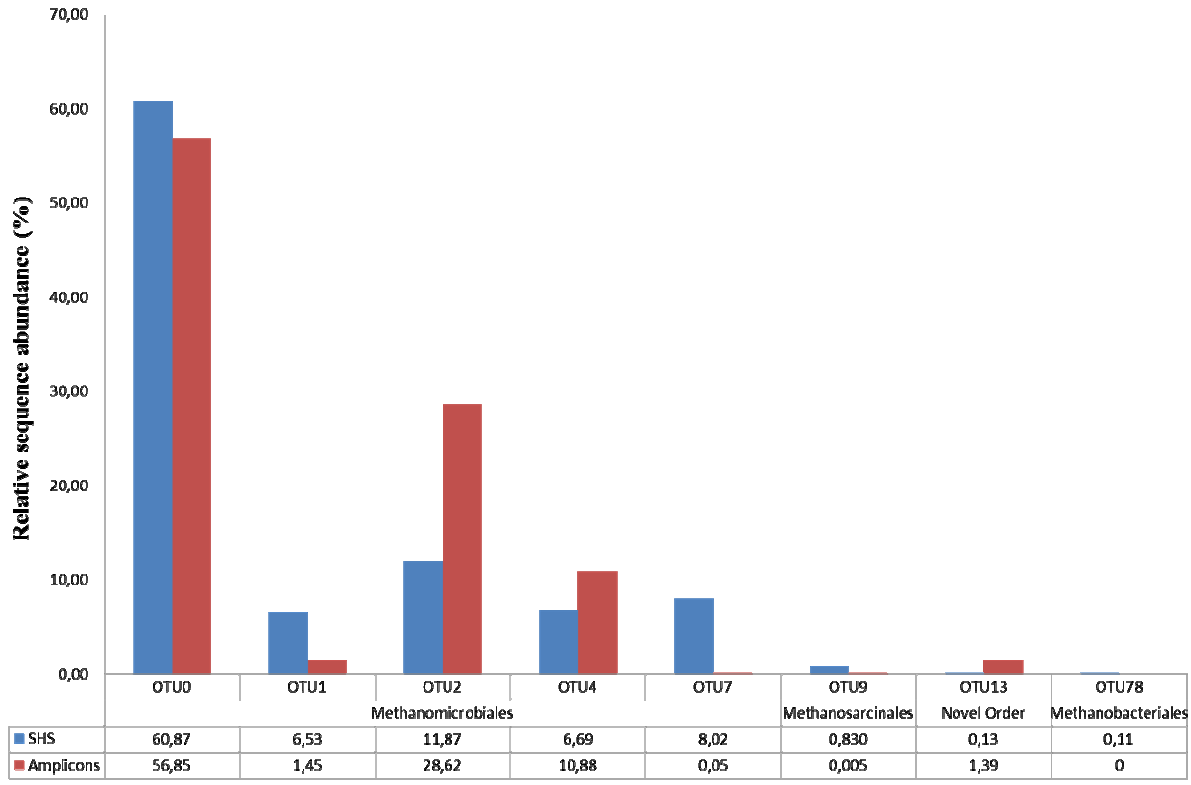

Figure 5. Graphic representation associated to numerical values illustrating the relative abundances pattern related to the dominant OTUs within the four methanogen orders for targeted capture method (SHS) and PCR-based strategy (amplicons).

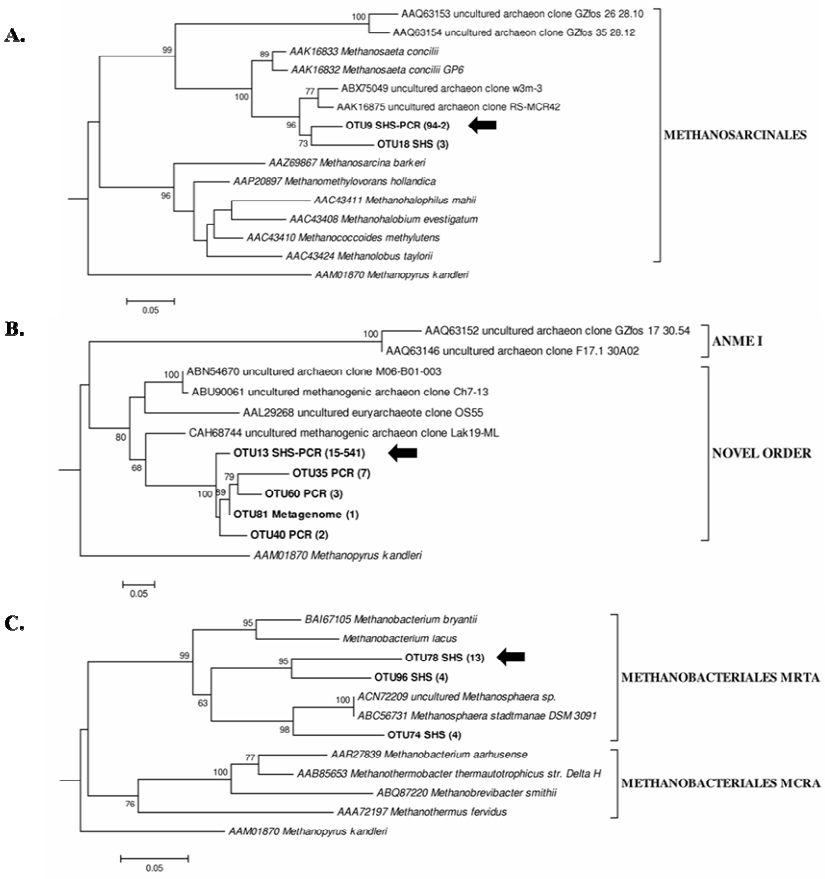

Figure 6. Phylogenetic analysis of deduced McrA amino acid sequences obtained from PCR, SHS and Pavin90m strategies, showing evolutionary distance within the order *Methanosarcinales* (A), Novel Order (B) and *Methanobacteriales* (C). Evolutionary history was inferred using the neighbour-joining method [38,39] (NJ, Poisson distance model) using Seaview software [35]. The final tree was drawn in MEGA 5 [40]. The bars represent 5% sequence divergence. Numbers at the nodes represent bootstrap values >60% (1,000 resamplings). The number of amino acid sequences assigned to each OTU is given into brackets, together with the name of the strategies for obtaining them. McrA amino acid sequence from *Methanopyrus kandleri* (AAM01870) was used as an outgroup for rooting the tree. Bold arrows indicate dominant OTUs.

genome of this species, the gene has been located ~700kbp downstream to the *mcr* operon. The sequences of DOMON domain-containing protein are closely related (74 to 80% identity) to *Methanosaeta concilii* GP-6 (accession no. AEB67518), belonging to *Methanosarcinales* order. On the reference genome of this species, the gene has been located ~50kbp downstream to the *mcr* operon.

4. Discussion

We presented the first attempt to capture specific-target DNA from a complex environmental metagenome using a modified SHS capture method coupled to NGS. Our data showed that the relative enrichment factor can be multiplied by applying two cycles of capture, to reach 175,365 times, which was superior compared to previous studies using a single cycle [18,19] and more efficient than that developed by Summerer et al. [46] using microarray-based capture. When applied to the anoxic layer of Lake Pavin, where Archaea accounted for 17% of DAPI-stained cells [47] of which only a fraction corresponded to methanogens, our SHS strategy demonstrated its ability to specifically enrich *mcrA* sequences. The SHS strategy, allowed us to identify changes in genomic organisation which are not easily accessed by other approaches, except strain isolation or perhaps with deep metagenomic sequencing.

The random-shotgun metagenomics approach clearly demonstrates that many hundreds of thousands of additional single reads would have been necessary in order to estimate the biodiversity of the methanogen community inhabiting this environment. The SHS experiment contained a much higher level of *mcrA* data thus providing a solid taxonomic basis for studying the diversity of methanogens. Finally, PCR is the most effective enrichment approach with approximately 100% of amplicons corresponding to the biomarker, demonstrating the specificity of the primers used [29].

The comparison of the overall methanogen communities retrieved with both SHS and amplicon strategies revealed similar patterns with a high abundance and high diversity of *Methanomicrobiales* sequences (more than 98% of the total sequences representing 48 OTUs). This data is in accordance with the study realised by Biderre-Petit et al. [27]. High throughput sequencing, however, reveals methanogen diversity to be much more important than previously described by amplicons library and Sanger sequencing [27]. It is important to notice that the amplicons sequencing approach, even with high throughput sequencing, missed all taxonomic groups assigned to *Methanobacteriales*, as well as certain assigned to

Table 2. Summary statistics from *de novo* assembly

Newbler version 2.6	SHS
No. of reads used for assembly	122,772
No. of reads assembled into contigs	53,307
No. of singletons	56,834
Outliers[1]	12,631
No. of contigs assembled	1916
N_{50} contig size (bases)	820
No. of *mcrA* homologous contigs	693
No. of *mcrA* homologous singletons	1142
Average *mcrA* homologous contig length (bases)	590
Largest *mcrA* homologous contig length (bases)	1639

[1]Reads were discarded due to quality control by Newbler

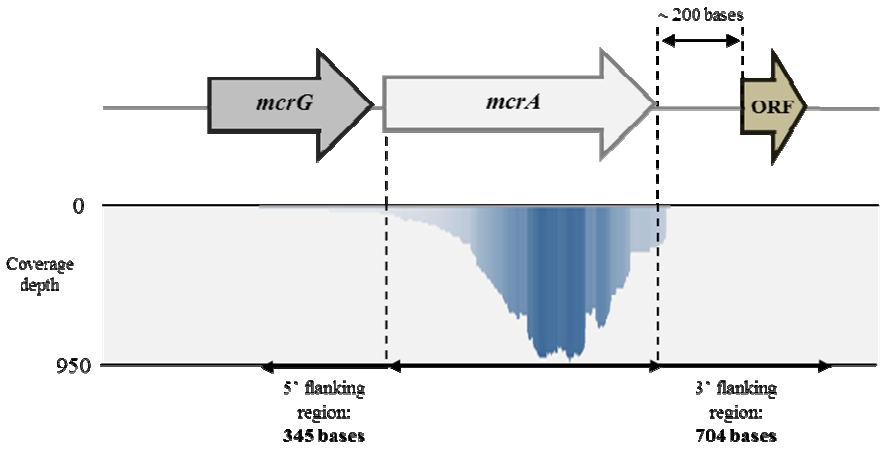

Figure 7. Graphic overview of sequence coverage analysis of *de novo* SHS read assembly against the reference genomes of *Candidatus Methanoregula boonei* 6A8, *Methanosaeta concilii* GP-6 and *Methanosphaera stadtmanae* DSM 3091. Shown is the *mcrA* gene as well as the adjacent genomic organisation (*top*) and coverage depth distribution (*bottom*). Coverage on the *mcrA* gene as well as on the flanking regions is shown in black arrow.

Methanosarcinales indicating a possible bias of *mcrA* primers. This often leads to significant underestimation of true community diversity [24,48]. SHS seems very efficient to target rare sequences as demonstrated for *Methanobacteriales* and does not appear to be influenced by genes GC content. As already demonstrated for microarrays approaches [21,22,49], the use of more extensive explorative capture probe sets in future work could avoid the limitation of sequence availability, and make possible the detection of a large number of previously uncharacterized microbial populations. Moreover, results retrieved with SHS and amplicons were correlated by the qPCR analysis.

We also used a *de novo* SHS read assembly to explore the flanking regions of the targeted gene, and we identified ORFs (dtxR and DOMON domain) downstream from the *mcrA* which were not normally described adjacent to the *mcr* operon in nucleotide databases. The particular genomic organisation observed, probably linking methanogenesis to electron transfer process and Fe homeostasis within the anoxic layer of the Lake Pavin, should reflect a particular adaptation to this environment. More experiments are needed, however, to validate this hypothesis.

With the emergence of third generation sequencing platforms and the possibility to sequence longer DNA sequences without library construction [50,51], the SHS strategy should provide real benefits to link genomic structure and function in microbial communities.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Acknowledgements

We would like to thank Yannick Bidet and Maud Privat from the Centre Jean Perrin for their help regarding samples processing on the 454 GS FLX pyrosequencing platform. We also thank Sarah Orlhac and Nicolas Gallois for their efficient technical assistance and David Tottey for reviewing the English version of the manuscript.

Funding

This work was supported by the ANR-09-EBIO-009 project (Agence Nationale de la Recherche). JD was supported by a studentship from the Centre National de la Recherche

Scientifique (CNRS, grant number 163588) and the Région Auvergne. NP was funded by Direction Générale de l'Armement (DGA).

References

1. Whitman, W. B., Coleman, D. C. and Wiebe, W. J. 1998, Prokaryotes: The unseen majority. Proc. Natl. Acad. Sci. U.S.A., 95, 6578-6583.

2. Curtis, T. P., Head, I. M., Lunn, M., Woodcock, S., Schloss, P. D. and Sloan, W. T. 2006, What is the extent of prokaryotic diversity? Philos. Trans. R. Soc. Lond., B, Biol. Sci., 361, 2023-2037.

3. Amann, R., Ludwig, W. and Schleifer, K.-H. 1995, Phylogenetic Identification and In Situ Detection of Individual Microbial Cells without Cultivation. Microbiol. Rev., 59, 143-169.

4. Eisen, J. A. 2007, Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol., 5, e82.

5. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M. 1998, Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol., 5, R245-249.

6. Biddle, J. F., Fitz-Gibbon, S., Schuster, S. C., Brenchley, J. E. and House, C. H. 2008, Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc. Natl. Acad. Sci. U.S.A., 105, 10583-10588.

7. Tringe, S. G., von Mering, C., Kobayashi, A., et al. 2005, Comparative metagenomics of microbial communities. Science, 308, 554-557.

8. Riesenfeld, C. S., Schloss, P. D. and Handelsman, J. 2004, Metagenomics: genomic analysis of microbial communities. Annu. Rev. Genet., 38, 525-552.

9. Suenaga, H. 2011, Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. Environ. Microbiol., 14, 13-22.

10. Edwards, R. A., Rodriguez-Brito, B., Wegley, L., et al. 2006, Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics, 7, 57.

11. Mardis, E. R. 2008, The impact of next-generation sequencing technology on genetics. Trends Genet., 24, 133-141.

12. Quince, C., Curtis, T. P. and Sloan, W. T. 2008, The rational exploration of microbial diversity. Isme J, 2, 997-1006.

13. Hoff, K. J. 2009, The effect of sequencing errors on metagenomic gene prediction. BMC Genomics, 10, 520.

14. Summerer, D. 2009, Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. Genomics, 94, 363-368.

15. Albert, T. J., Molla, M. N., Muzny, D. M., et al. 2007, Direct selection of human genomic loci by microarray hybridization. Nat. Methods, 4, 903-905.

16. Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J. and Zwick, M. E. 2007, Microarray-based genomic selection for high-throughput resequencing. Nat. Methods, 4, 907-909.

17. Mokry, M., Feitsma, H., Nijman, I. J., et al. 2010, Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res., 38, e116.

18. Tewhey, R., Nakano, M., Wang, X., et al. 2009, Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol., 10, R116.

19. Gnirke, A., Melnikov, A., Maguire, J., et al. 2009, Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol., 27, 182-189.

20. Iwai, S., Chai, B., Sul, W. J., Cole, J. R., Hashsham, S. A. and Tiedje, J. M. 2010, Gene-targeted-metagenomics reveals extensive diversity of aromatic dioxygenase genes in the environment. Isme J, 4, 279-285.

21. Terrat, S., Peyretaillade, E., Goncalves, O., et al. 2010, Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. BMC Bioinformatics, 11, 478.

22. Dugat-Bony, E., Peyretaillade, E., Parisot, N., et al. 2011, Detecting unknown sequences with DNA microarrays: explorative probe design strategies. Environ. Microbiol., doi: 10.1111/j.1462-2920.2011.02559.x.

23. Suzuki, M. and Giovannoni, S. 1996, Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol., 62, 625-630.

24. Hong, S., Bunge, J., Leslin, C., Jeon, S. and Epstein, S. S. 2009, Polymerase chain reaction primers miss half of rRNA microbial diversity. Isme J, 3, 1365-1373.

25. Reeve, J. N. 1992, Molecular biology of methanogens. Annu. Rev. Microbiol., 46, 165-191.

26. Klein, A., Allmansberger, R., Bokranz, M., Knaub, S., Müller, B. and Muth, E. 1988, Comparative analysis of genes encoding methyl coenzyme M reductase in methanogenic bacteria. Mol. Gen. Genet., 213, 409-420.

27. Biderre-Petit, C., Jezequel, D., Dugat-Bony, E., et al. 2011, Identification of microbial communities involved in the methane cycle of a freshwater meromictic lake. FEMS Microbiol. Ecol., 77, 533-545.

28. Dugat-Bony, E., Missaoui, M., Peyretaillade, E., et al. 2011, HiSpOD: probe design for functional DNA microarrays. Bioinformatics, 27, 641-648.

29. Mihajlovski, A., Alric, M. and Brugere, J. F. 2008, A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of the mcrA gene. Res. Microbiol., 159, 516-521.

30. Staden, R. 1996, The Staden sequence analysis package. Mol. Biotechnol., 5, 233-241.

31. Schmieder, R. and Edwards, R. 2011, Quality control and preprocessing of metagenomic datasets. Bioinformatics, 27, 863-864.

32. Altschul, S. F., Madden, T. L., Schäffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389-3402.

33. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. and Knight, R. 2011, UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27, 2194-2200.

34. Larkin, M. A., Blackshields, G., Brown, N. P., et al. 2007, Clustal W and Clustal X version 2.0. Bioinformatics, 23, 2947-2948.

35. Gouy, M., Guindon, S. and Gascuel, O. 2009, SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol. Biol. Evol., 27, 221-224.

36. Li, W. and Godzik, A. 2006, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22, 1658-1659.

37. Luton, P. E., Wayne, J. M., Sharp, R. J. and Riley, P. W. 2002, The mcrA gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. Microbiology, 148, 3521–3530.

38. Studier, J. and Keppler, K. 1988, A note on the neighbor-joining algorithm of Saitou and Nei. Mol. Biol. Evol., 5, 729-731.

39. Saitou, N. and Nei, M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol., 4, 406-425.

40. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011, MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol., 28, 2731-2739.

41. Livak, K. J. and Schmittgen, T. D. 2001, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods, 25, 402-408.

42. Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. 1990, Basic local alignment search tool. J. Mol. Biol., 215, 403-410.

43. Brauer, S. L., Cadillo-Quiroz, H., Yashiro, E., Yavitt, J. B. and Zinder, S. H. 2006, Isolation of a novel acidiphilic methanogen from an acidic peat bog. Nature, 442, 192-194.

44. Fricke, W. F., Seedorf, H., Henne, A., et al. 2005, The Genome Sequence of Methanosphaera stadtmanae Reveals Why This Human Intestinal Archaeon Is Restricted to Methanol and H2 for Methane Formation and ATP Synthesis. J. Bacteriol., 188, 642-658.

45. Borrel, G., Joblin, K., Guedon, A., et al. 2011, Methanobacterium lacus sp. nov., a novel hydrogenotrophic methanogen from the deep cold sediment of a meromictic lake. Int. J. Syst. Evol. Microbiol., DOI: 10.1099/ijs.1090.034538-034530.

46. Summerer, D., Wu, H., Haase, B., et al. 2009, Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. Genome Res., 19, 1616-1621.

47. Lehours, A. C., Bardot, C., Thenot, A., Debroas, D. and Fonty, G. 2005, Anaerobic microbial communities in Lake Pavin, a unique meromictic lake in France. Appl. Environ. Microbiol., 71, 7389-7400.

48. Jeon, S., Bunge, J., Leslin, C., Stoeck, T., Hong, S. and Epstein, S. S. 2008, Environmental rRNA inventories miss over half of protistan diversity. BMC Microbiol., 8, 222.

49. Militon, C., Rimour, S., Missaoui, M., et al. 2007, PhylArray: phylogenetic probe design algorithm for microarray. Bioinformatics, 23, 2550-2557.

50. McCarthy, A. 2010, Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. Chem. Biol., 17, 675-676.

51. Schadt, E. E., Turner, S. and Kasarskis, A. 2010, A window into third-generation sequencing. Hum. Mol. Genet., 19, R227-240.

4. Discussion

L'approche moléculaire de capture de gènes a démontré sa pertinence pour assurer un enrichissement significatif des séquences ciblées. En effet, dans notre étude en partant d'un échantillon métagénomique complexe nous avons pu obtenir une efficacité d'enrichissement supérieur à 40% du biomarqueur de la méthanogénèse. Cette étude comparative a clairement démontré les difficultés à décrire la diversité de façon exhaustive par le séquençage direct de l'ADN métagénomique. Afin d'avoir une vision globale de la diversité des communautés méthanogènes au sein de cet environnement, des millions de lectures supplémentaires auraient été nécessaires, en accord avec les conclusions tirées des travaux initiés par Quince et al. (2008) sur les efforts de séquençage à fournir pour explorer la diversité microbienne.

Cette stratégie a aussi permis, contrairement aux approches classiques utilisant la PCR, d'explorer de manière plus exhaustive les communautés méthanogènes. Un tel résultat a été rendu possible grâce au fait que cette approche s'affranchit des biais occasionnés par les méthodes basées sur la PCR. En effet, l'efficacité de ces dernières est intimement liée au choix des amorces et à la part relative de chaque communauté à identifier au sein de l'écosystème. Ainsi, l'utilisation de la capture a permis d'identifier de nouvelles séquences non répertoriées dans les bases de données mais également d'accéder à des méthanogènes rares. Une autre limitation des approches PCR, levée par l'approche capture, est la taille des séquences pouvant être identifiées. En effet, la longueur des amplicons obtenus n'est pas toujours suffisante pour caractériser précisément les communautés microbiennes (Wommack et al 2008). L'approche capture, permettant quant à elle d'obtenir des fragments de grande taille, un nombre de sites moléculaires du biomarqueur ciblé plus important est donc disponible pour entreprendre l'identification phylogénétique des communautés.

Enfin, l'exploitation des régions flanquantes au gène *mcrA* a permis la caractérisation de gènes impliqués dans des processus métaboliques en lien avec les conditions physico-chimiques du lac Pavin. La possibilité d'identifier de grandes régions d'ADNg représente l'autre atout majeur de cette approche de capture. En effet, ces régions peuvent inclure plusieurs gènes pouvant faire partie de la même unité transcriptionnelle et donc permettre de donner des pistes pour l'annotation fonctionnelle de nouveaux gènes à séquences inconnues mais associés à des gènes codant pour des protéines à fonction connue (Korbel et al 2004, Overbeek et al 1999). Une telle approche représente une alternative pour la prédiction de la fonction des gènes. Actuellement, la prédiction de fonction des gènes est généralement basée

sur une recherche d'homologie de séquences dans les bases de données en utilisant notamment les outils BLAST (Altschul et al 1990) mais en gardant à l'esprit qu'un grand nombre de séquences sont mal annotées (Valencia et al 2009). De même, du fait de la connaissance partielle de l'extraordinaire diversité des microorganismes dans les environnements, il a été montré que suite au séquençage direct de métagénomes, environ 30 à 60% des protéines ne pouvaient être clairement identifiées avec une fonction connue en utilisant les bases de données actuelles (Vieites et al 2009).

Cette méthode peut être facilement multiplexable et automatisable pour assurer la capture de plusieurs échantillons en même temps. Cette technique a démontré son potentiel en enrichissant spécifiquement un gène d'intérêt au sein d'un échantillon complexe et elle s'impose comme une stratégie de choix en écologie microbienne pour l'étude des communautés microbiennes en s'affranchissant des biais PCR. Il est également possible de l'appliquer pour la capture de biomarqueurs phylogénétiques comme par exemple l'ARNr 16S, permettant ainsi d'explorer finement la structure des communautés microbiennes. De même, l'utilisation de KASpOD permettra d'affiner la sélection des sondes de capture intégrant le caractère exploratoire, en terme de couverture et de spécificité, de rapidité et sur le choix de différents critères thermodynamiques.

Enfin, l'émergence du séquençage de troisième génération devrait faciliter l'obtention de données sur de très grandes régions d'ADN capturées et améliorer encore nos connaissances sur le monde microbien.

Conclusions et perspectives

L'immense réservoir génétique des communautés microbiennes renferme des capacités métaboliques uniques leur permettant de s'adapter à tous types d'environnement. Les microorganismes participent activement aux grands cycles biogéochimiques assurant ainsi le bon fonctionnement des écosystèmes. Ils jouent également un rôle dans les changements globaux notamment par le biais de la production de gaz à effet de serre, qui vient s'ajouter à celle générée par les activités anthropiques intensives. Du fait de l'intérêt grandissant des pouvoirs publics pour le changement global et les émissions de gaz à effet de serre, de nombreuses recherches sur des stratégies permettant de réduire les émissions de méthane voient le jour. L'objectif est de mieux comprendre les mécanismes mis en jeu au niveau de la production de méthane par les méthanogènes et leurs implications dans le cycle du carbone.

De plus, les méthanogènes, du fait de leurs associations syntrophiques, pourraient participer à l'élimination de polluants issus de l'utilisation intensive des énergies fossiles. L'exploitation des capacités microbiennes, notamment de biodégradation des hydrocarbures aromatiques polycycliques en anaérobiose, représente donc un potentiel très intéressant pour la mise en place de stratégies de bioremédiation pour la restauration des environnements contaminés. Néanmoins, le développement de ces stratégies nécessite une meilleure connaissance du monde microbien passant par l'utilisation de techniques adaptées en raison de son extrême diversité. Au cours de ce travail de thèse, les avantages et les inconvénients de chacune des techniques actuellement disponibles en écologie microbienne ont pu être décrites. Il est apparu important de pouvoir tirer parti du potentiel même de ces méthodes comme les biopuces à ADN ou le séquençage massif, afin de proposer de nouvelles stratégies d'exploration de la structure et de la fonction des communautés microbiennes. Il faut également noter qu'un grand nombre d'approches de biologie moléculaire nécessite l'utilisation d'oligonucléotides comme sondes ou comme amorces. Il apparaît essentiel de disposer de logiciels performants pour une détermination efficace de ces séquences oligonucléotidiques tel que nous l'avons présenté au travers du chapitre du livre. Toujours dans le but d'améliorer la détermination de ces séquences, nous avons développé un nouveau logiciel de sélection de sondes adapté aux problématiques environnementales. Ce logiciel, nommé KASpOD, combine les critères de sensibilité, de spécificité et le caractère

exploratoire pour la détermination d'oligonucléotides de qualité, et cette détermination peut être réalisée à partir de grands jeux de données.

Tirant partie de cette expertise sur la détermination de sondes, les travaux menés au cours de cette thèse ont conduit au développement d'une nouvelle méthode d'étude des échantillons métagénomiques. Cette approche, appelée capture de gènes en solution, présente de nombreux avantages pour étudier la diversité des communautés microbiennes à partir d'échantillons environnementaux complexes. La capture offre la possibilité de cibler spécifiquement des populations microbiennes difficilement accessibles par d'autres approches et de décrire des organisations géniques pouvant être à la base de mécanismes d'adaptations. La méthode a été validée en utilisant un jeu de 26 sondes de 50-mers ciblant toutes les séquences codant pour la méthyl coenzyme M réductase. Le couplage avec le séquençage haut débit a permis une exploration exhaustive des populations de méthanogènes au sein d'un environnement lacustre, le lac Pavin. Cette approche a révélé une très grande diversité de méthanogènes (plus de 40 OTU) appartenant à quatre ordres différents. En comparaison à une approche métagénomique directe, qui n'a permis d'identifier qu'un seul OTU, ou à une approche amplicons ne révélant la présence que de trois ordres, la capture a clairement montré son efficacité pour assurer une meilleure évaluation de la diversité microbienne, notamment des populations peu représentées. En effet, les conditions physico-chimiques particulières du lac Pavin au niveau de sa zone anoxique et proche des sédiments (pH, salinité, température) peuvent être propices à l'existence de microniches abritant la biosphère rare. Les faibles températures prévalant à 90m de profondeur (~5°C), qui sont largement en dessous des optimums de croissance des *Methanobacteriales* (Borrel et al 2012), peuvent expliquer une sous-représentation de ce type de méthanogènes au sein de cet environnement. Ces méthanogènes seront donc confinés à des microniches. De même, il est possible que d'autres communautés de méthanogènes comme les *Methanomicrobiales*, présentes à cette profondeur dans le lac Pavin, soient plus compétitives que les *Methanobacteriales* en présence des faibles concentrations en H_2, généralement rencontrées à cette profondeur.

Outre la possibilité de pouvoir explorer rapidement et efficacement la diversité des communautés microbiennes à un très haut-débit, l'utilisation de la capture de gènes pour identifier de larges régions d'ADN génomique peut favoriser l'identification de nouveaux gènes mais aussi la compréhension de processus adaptatifs liés à l'environnement. Ainsi, il a pu être mis en évidence pour certaines populations de méthanogènes du lac Pavin, des

associations potentielles entre la métathanogénèse et l'homéostasie du Fer. La proximité relative entre les gènes impliqués dans la régulation du métabolisme du fer et l'opéron *mcr*, ainsi que les conditions physico-chimiques particulières du lac Pavin (très riche en Fer dissous), permettent de supposer une adaptation particulière de la méthanogénèse aux conditions ferrugineuses de cet environnement complexe. Dans la colonne d'eau du lac Pavin, le cycle du fer implique la formation de particules de fer oxydé et réduit. Il a été proposé que les particules de fer ferrique, riche en PO_4 et formées dans la partie inférieure de l'interface entre la zone oxique et anoxique, sédimentent jusqu'à la surface des sédiments où elles seraient réduites et re-solubilisées (Michard et al 1994). Les particules de fer réduit (pyrite, FeS_2), protovivianite ($Fe_3(PO_4)_2$) et sidérite ($FeCO_3$), ne peuvent pas être solubilisées avec les conditions présentes dans les zones profondes du lac Pavin et ces dernières s'accumulent donc dans les sédiments. La zone profonde anoxique du lac Pavin est donc très riche en fer dissous Fe(II) avec des concentrations supérieures à 1mM à 90m (Bura-Nakić et al 2009). Les communautés méthanogènes (notamment les hydrogénotrophes) pourraient donc tirer bénéfice de ces concentrations en fer en rivalisant avec des communautés ferro-réductrices, compétitrices directes des méthanogènes pour l'acquisition des substrats issus de la fermentation (Lehours et al 2009)

Les résultats obtenus au cours de cette thèse ouvrent de nombreuses perspectives tant en bioinformatique qu'en écologie microbienne. La détermination de sondes reste au cœur de nombreuses techniques (FISH, PCR, biopuces ADN, criblage de banques métagénomique et capture de gènes) et nécessite des logiciels pouvant gérer les flux massifs de données issues du séquençage nouvelles générations. KASpOD a apporté de nombreuses optimisations en terme de temps de calcul et pourrait évoluer vers le déploiement sur des architectures plus puissantes de type grilles de calcul. Actuellement, KASpOD est déployé sur un cluster composé de 140 CPUs hébergé au Centre Régional des Ressources Informatiques (CRRI) de Clermont-Ferrand. Une perspective intéressante serait de pouvoir bénéficier du potentiel des architectures parallèles comme par exemple la grille de calcul « Enabling Grid for E-sciences » (EGEE), réunissant 250 partenaires dans le monde entier générant une capacité totale de 40 000 CPUs et plusieurs Petabits de stockage. Cette augmentation des capacités de calculs offre la possibilité de considérer un plus grand nombre de critères assurant la sélection de sondes performantes. Parmi ces critères, il est possible de citer les paramètres thermodynamiques ou les structures secondaires des sondes et des cibles qui, pour être évaluées, demandent de longs temps de calcul. La prise en compte d'un maximum de critères par le logiciel pourrait donc

limiter le nombre de sondes nécessaires à la détection de chaque gène ciblé. Il faut cependant garder à l'esprit que la thermodynamique des hybridations des acides nucléiques reste mal connue, notamment s'agissant des réactions au niveau de l'interface liquide/solide (Pozhitkov et al 2007). Malgré des caractéristiques thermodynamiques définies comme étant de bonne qualité, une sonde peut donc conduire à des résultats erronés (*e.g.* absence d'hybridation avec la cible). Pour s'affranchir de cette limite, la stratégie actuelle consiste à sélectionner un groupe de sondes permettant de cibler différentes régions de chaque gène (Chou 2004). Cependant, la mise en œuvre de cette stratégie est dépendante du gène ciblé et peut dans certains cas s'avérer difficile en raison, par exemple, de la taille et de la diversité des séquences au sein du groupe ciblé ou des critères (*e.g.* taille, paramètres thermodynamiques) auxquels doivent répondre chaque sonde. Une autre amélioration, liée à l'évolution constante des bases de données internationales comme GenBank (http://www.ncbi.nlm.nih.gov/genbank/), EMBL (http://www.ebi.ac.uk/embl/) et DDBJ (http://www.ddbj.nig.ac.jp/), serait d'intégrer au logiciel la recherche automatique des séquences utilisées pour la détermination des sondes. Malheureusement, la mauvaise qualité des bases de données internationales ne les rend pas directement exploitable. Il est donc nécessaire de constituer des bases de données de qualité en termes de séquences mais aussi d'annotation (taxonomique et/ou fonctionnelle). Nous avons pu montrer que la détermination de sondes de qualité pouvait permettre la mise en place d'une nouvelle approche d'exploration de la diversité microbienne à travers la capture de gènes. Cette approche s'est montrée plus efficace que la PCR (amplicons) et l'approche directe de séquençage de métagénome. Cette méthode généraliste peut être appliquée à n'importe quel gène cible afin d'assurer l'exploration taxonomique et/ou fonctionnelle de tous les types d'environnements. De plus, les sondes produites pour cibler un type de gène peuvent être réemployées pour d'autres études. Ainsi, les sondes de capture *mcrA* sont actuellement utilisées pour explorer la diversité des méthanogènes chez le ruminant. En effet, les activités agricoles et particulièrement l'élevage des ruminants, contribuent à la production annuelle de méthane à hauteur de 17% (Conrad 2009). Il est estimé que la production globale de produits alimentaires issus de l'élevage des ruminants va continuer à augmenter, et notamment la production laitière qui pourrait doubler d'ici 2050. Ces prévisions laissent présager de fortes répercussions environnementales liées à une demande imortante de la part des pays développés en produits agricoles issus de l'élevage. De nouvelles stratégies sont en train d'émerger pour promouvoir des solutions alternatives afin de limiter la production de CH_4

chez les ruminants. Ce projet, nommé Crédit, est financé par l'Agence Nationale de la Recherche (ANR) et est porté par le Dr. Diego MORGAVI de l'INRA de Clermont-Ferrand/ Theix. Outre son impact sur l'environnement, la production de méthane dans le rumen est connue depuis longtemps comme une perte d'énergie pour l'animal (entre 4 et 6% de l'énergie alimentaire). Il a été montré que la synthèse de méthane peut être réduite en modifiant le processus fermentaire au sein du rumen. C'est avec cet objectif que des stratégies nutritionnelles commencent à être employées dans le domaine de l'élevage (Beauchemin et al 2008). En effet, le remplacement des fourrages par des aliments concentrés, l'utilisation de lipides, ou encore l'utilisation d'extraits de plantes ont démontré une réduction de la production de méthane (Eugène et al 2008, Machmüller 2006, Martin et al 2006). L'objectif du projet consiste à diminuer la production de méthane chez le ruminant grâce à l'utilisation d'une alimentation supplémentée avec un principe actif produit par un champignon. Une partie du projet consiste donc à étudier, par la méthode de capture de gènes ciblée sur le gène *mcrA*, les communautés méthanogènes du rumen pour comprendre les mécanismes mis en jeu lors de ce régime alimentaire et qui permettent la réduction de la production de méthane.

Une autre perspective intéressante de la capture de gènes serait de l'appliquer pour cibler le biomarqueur le plus étudié, le gène codant l'ARNr 16S. C'est avec cet objectif qu'un premier jeu de sondes généralistes est actuellement testé sur divers environnements (symbiome microbien d'arthropodes : collaboration Sylvain Charlat, Laboratoire de Biométrie et Biologie Evolutive (LBBE) à Villeurbanne ; rumen collaboration Diego Morgavi, INRA de Clermont-Ferrand Theix ; stations d'épurations : collaboration Denis Le Paslier Genoscope Evry ; environnements lacustres : collaboration Corinne PETIT LMGE Clermont-Ferrand).

Enfin, au-delà des questions biologiques pour lesquelles le développement de l'approche capture de gènes en solution a été entrepris, différentes considérations d'ordre technologique sont à considérer afin d'améliorer encore cette nouvelle méthode. La première concerne le multiplexage de la technique, de manière à pouvoir disposer d'une méthode à haut débit permettant de traiter simultanément plusieurs échantillons. Ce multiplexage passe par une automatisation de la méthode en plaque 96 puits pour la préparation en amont des banques d'ADN utilisées pour la capture mais également pour les différentes étapes de phase de capture (hybridation en solution, sédimentation magnétique…). Différents protocoles optimisés commencent à voir le jour et permettent la production de banques d'ADN à moindre coût, comme celui assurant l'obtention de 192 banques en une seule journée pour un

prix de 15$ (Rohland and Reich 2012). Ces méthodes, combinant des aspects biologiques à des développements en robotique et en automatisme, ouvrent de réelles perspectives pour l'automatisation de la capture de gènes en solution. Pour optimiser cette nouvelle approche, il est également nécessaire d'améliorer la spécificité de l'approche en jouant sur les différentes séquences d'adaptateurs associées aux sondes et nécessaires à la préparation des banques. Il s'agit notamment de limiter les phénomènes d'hybridations aspécifiques dues à ces séquences (Rohland and Reich 2012). Pour cela, deux stratégies alternatives sont envisagées. La première consiste à préparer des banques d'ADN à l'aide d'adaptateurs plus courts comportant une séquence étiquette pour un éventuel multiplexage, et des sites moléculaires d'ancrage assurant l'incorporation par PCR des séquences utilisées pour le séquençage. La seconde alternative consiste à utiliser des oligonucléotides qualifiés de « bloquants », qui s'hybrident de manière complémentaire au niveau des adaptateurs au cours de l'hybridation empêchant ainsi les hybridations aspécifiques au niveau de ces adaptateurs. Ces oligonucléotides ont également la particularité de posséder un didésoxynucléotide en leur extrémité 3' empêchant l'initiation de toute polymérisation. Cette optimisation de la spécificité d'hybridation est particulièrement utile dans le cadre des études environnementales, qui emploient des échantillons métagénomiques très complexes. Enfin, la capture, en permettant de piéger de grands fragments d'ADN ainsi que l'émergence des technologies de séquençage de troisième génération, devrait permettre de séquencer en une seule fois ces longues molécules d'ADN et donc de s'affranchir de l'étape d'assemblage.

En résumé, les résultats obtenus au cours de cette thèse lient le développement d'outils innovants à l'acquisition de données massives sur les populations microbiennes productrices de méthane. Ces populations montrent une très forte diversité avec probablement l'occupation de microniches écologiques insoupçonnées jusqu'à présent. D'autre part, la compréhension du développement de ces méthanogènes dans différents environnements pourrait permettre la mise en œuvre de nouvelles stratégies visant à réduire production de méthane dans le cadre du changement global, ainsi qu'à utiliser leurs potentialités métaboliques pour la réhabilitation de sites pollués.

Références

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005). PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Applied and Environmental Microbiology* **71:** 8966-8969.

Adessi C, Matton G, Ayala G, Turcatti G, Mermod J, Mayer P *et al* (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic acids research* **28:** E87.

Ahmadian A, Ehn M, Hober S (2006). Pyrosequencing: History, biochemistry and future. *Clinica Chimica Acta* **363:** 83-94.

Ahmed N, Claesson MJ, O'Sullivan O, Wang Q, Nikkilä J, Marchesi JR *et al* (2009). Comparative Analysis of Pyrosequencing and a Phylogenetic Microarray for Exploring Microbial Community Structures in the Human Distal Intestine. *PLoS One* **4:** e6669.

Ahmed N, Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR *et al* (2011). Prospective Genomic Characterization of the German Enterohemorrhagic Escherichia coli O104:H4 Outbreak by Rapid Next Generation Sequencing Technology. *PLoS One* **6:** e22751.

Ahn J, Yang L, Paster BJ, Ganly I, Morris L, Pei Z *et al* (2011). Oral Microbiome Profiles: 16S rRNA Pyrosequencing and Microarray Assay Comparison. *PLoS One* **6:** e22788.

Aitken C, Jones D, Larter S (2004). Anaerobic hydrocarbon biodegradation in deep subsurface oil reservoirs. *Nature* **431:** 291-294.

Alain K, Holler T, Musat F, Elvert M, Treude T, Kruger M (2006). Microbiological investigation of methane- and hydrocarbon-discharging mud volcanoes in the Carpathian Mountains, Romania. *Environmental microbiology* **8:** 574-590.

Albert FW, Hodges E, Jensen JD, Besnier F, Xuan Z, Rooks M *et al* (2011). Targeted resequencing of a genomic region influencing tameness and aggression reveals multiple signals of positive selection. *Heredity* **107:** 205-214.

Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X *et al* (2007). Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4:** 903-905.

Almomani R, van der Heijden J, Ariyurek Y, Lai Y, Bakker E, van Galen M *et al* (2010). Experiences with array-based sequence capture; toward clinical applications. *European Journal of Human Genetics* **19:** 50-55.

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990). Basic local alignment search tool. *Journal of molecular biology* **215:** 403-410.

Amann R, Krumholz L, Stahl D (1990). Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *Journal of bacteriology* **172:** 762-770.

Amann R, Ludwig W, Schleifer K (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews* **59:** 143-169.

Amstutz U, Andrey-Zurcher G, Suciu D, Jaggi R, Haberle J, Largiader CR (2010). Sequence Capture and Next-Generation Resequencing of Multiple Tagged Nucleic Acid Samples for Mutation Screening of Urea Cycle Disorders. *Clinical Chemistry* **57:** 102-111.

Ansorge WJ (2009). Next-generation DNA sequencing techniques. *New Biotechnology* **25:** 195-203.

Antipova A, Sokolsky T, Clouser C, Dimalanta E, Hendrickson C, Kosnopo C *et al* (2009). Polymorphism discovery in high-throughput resequenced microarray-enriched human genomic loci. *Journal of biomolecular techniques : JBT* **20:** 253-257.

Ariesyady H, Ito T, Okabe S (2007). Functional bacterial and archaeal community structures of major trophic groups in a full-scale anaerobic sludge digester. *Water Research* **41:** 1554-1568.

Aufhammer SW, Warkentin E, Berk H, Shima S, Thauer RK, Ermler U (2004). Coenzyme Binding in F420-Dependent Secondary Alcohol Dehydrogenase, a Member of the Bacterial Luciferase Family. *Structure* **12:** 361-370.

Bader KC, Grothoff C, Meier H (2011). Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics* **27:** 1546-1554.

Banning N, Brock F, Fry JC, Parkes RJ, Hornibrook ER, Weightman AJ (2005). Investigation of the methanogen population structure and activity in a brackish lake sediment. *Environmental microbiology* **7:** 947-960.

Bapteste E, Brochier C, Boucher Y (2005). Higher-level classification of the Archaea: evolution of methanogenesis and methanogens. *Archaea* **1:** 353-363.

Barnes R, Goldberg E (1976). Methane production and consumption in anoxic marine sediments. *Geology* **4:** 297-300.

Bau S, Schracke N, Kränzle M, Wu H, Stähler PF, Hoheisel JD *et al* (2008). Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. *Analytical and Bioanalytical Chemistry* **393:** 171-175.

Baumard P, Budzinski H, Garrigues P, Dizer H, Hansen PD (1999). Polycyclic

aromatic hydrocarbons in recent sediments and mussels (Mytilus edulis) from the Western Baltic Sea. *Marine Environmental Research* **47:** 17-47.

Beal EJ, House CH, Orphan VJ (2009). Manganese- and Iron-Dependent Marine Methane Oxidation. *Science* **325:** 184-187.

Beauchemin KA, Kreuzer M, O'Mara F, McAllister TA (2008). Nutritional management for enteric methane abatement: A review. *Australian Journal of Experimental Agriculture* **48:** 21-27.

Beazley M, Martinez R, Rajan S, Powell J, Piceno Y, Tom L *et al* (2012). Microbial community analysis of a coastal salt marsh affected by the deepwater horizon oil spill. *PLoS One* **7:** e41305.

Belova SE, Baani M, Suzina NE, Bodelier PLE, Liesack W, Dedysh SN (2011). Acetate utilization as a survival strategy of peat-inhabiting Methylocystis spp. *Environmental Microbiology Reports* **3:** 36-46.

Bentley DR (2006). Whole-genome re-sequencing. *Current Opinion in Genetics & Development* **16:** 545-552.

Berdugo-Clavijo C, Dong X, Soh J, Sensen CW, Gieg LM (2012). Methanogenic biodegradation of two-ringed polycyclic aromatic hydrocarbons. *FEMS Microbiology Ecology* **81:** 124-133.

Berk H, Thauer R (1997). Function of coenzyme F420-dependent NADP reductase in methanogenic archaea containing an NADP-dependent alcohol dehydrogenase. *Archives Of Microbiology* **168:** 396-402.

Berner EK, Berner RA (1996). Global Environment: Water, Air, and Geochemical Cycles. *Prentice Hall, Upper Saddle River, NJ*.

Bertrand JC, Caumette P, Lebaron P, Matheron R, Normand P (2012). Ecologie Microbienne : Microbiologie des milieux naturels et anthropisés. *Presses Universitaires de Pau et des Pays de l'Adour*.

Besombesa J, Maîtreb A, Patissiera O, Marchanda N, Chevrona N, Stoklovb M *et al* (2001). Particulate PAHs observed in the surrounding of a municipal incinerator. *Atmospheric Environment* **35:** 6093-6104.

Bianchin M, Smith L, Barker JF, Beckie R (2006). Anaerobic degradation of naphthalene in a fluvial aquifer: A radiotracer study. *Journal of Contaminant Hydrology* **84:** 178-196.

Biderre-Petit C, Jézéquel D, Dugat-Bony E, Lopes F, Kuever J, Borrel G *et al* (2011). Identification of microbial communities involved in the methane cycle of a freshwater meromictic lake. *FEMS Microbiology Ecology* **77:** 533-545.

Bleicher K, Zellner G, Winter J (1989). Growth of methanogens on cyclopentanol/CO2 and specificity of alcohol dehydrogenase. *FEMS Microbiology Letters* **59:** 307-312.

Blow N (2008). DNA sequencing: generation next-next. *Nature Methods* **5:** 267-274.

Bodrossy L, Stralis-Pavese N, Murrell J, Radajewski S, Weilharter A, Sessitsch A (2003). Development and validation of a diagnostic microbial microarray for methanotrophs. *Environmental microbiology* **5:** 566-582.

Boetius A, Ravenschlag K, Schubert C, Rickert D, Widdel F, Gieseke A *et al* (2000). A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407:** 623-626.

Bonacker L, Baudner S, Thauer R (1992). Differential expression of the two methyl-coenzyme M reductases in *Methanobacterium thermoautotrophicum* as determined immunochemically via isoenzyme-specific antisera. *European journal of biochemistry / FEBS* **206:** 87-92.

Bontemps C, Golfier G, Gris-Liebe C, Carrere S, Talini L, Boivin-Masson C (2005). Microarray-Based Detection and Typing of the Rhizobium Nodulation Gene nodC: Potential of DNA Arrays To Diagnose Biological Functions of Interest. *Applied and Environmental Microbiology* **71:** 8042-8048.

Borrel G, Joblin K, Guedon A, Colombet J, Tardy V, Lehours AC *et al* (2012). Methanobacterium lacus sp. nov., isolated from the profundal sediment of a freshwater meromictic lake. *International Journal of Systematic and Evolutionary Microbiology* **62:** 1625-1629.

Bottari B, Ercolini D, Gatti M, Neviani E (2006). Application of FISH technology for microbiological analysis: current state and prospects. *Applied Microbiology and Biotechnology* **73:** 485-494.

Bourne D, McDonald I, Murrell J (2001). Comparison of pmoA PCR primer sets as tools for investigating methanotroph diversity in three Danish soils. *Applied and Environmental Microbiology* **67**.

Brauer SL, Cadillo-Quiroz H, Ward RJ, Yavitt JB, Zinder SH (2010). Methanoregula boonei gen. nov., sp. nov., an acidiphilic methanogen isolated from an acidic peat bog. *International Journal of Systematic and Evolutionary Microbiology* **61:** 45-52.

Bräuer SL, Cadillo-Quiroz H, Yashiro E, Yavitt JB, Zinder SH (2006). Isolation of a novel acidiphilic methanogen from an acidic peat bog. *Nature* **442:** 192-194.

Brenner M, Zhang H, Scott R (1993). Nature of the low activity of S-methyl-coenzyme M reductase as determined by active site titrations. *The Journal of biological chemistry* **268:** 18491-18495.

Brodie EL, DeSantis TZ, Joyner DC, Baek SM, Larsen JT, Andersen GL *et al* (2006). Application of a High-Density Oligonucleotide Microarray Approach To Study Bacterial Population Dynamics during Uranium Reduction and Reoxidation. *Applied and Environmental Microbiology* **72:** 6288-6298.

Bura-Nakić E, Viollier E, Jézéquel D, Thiam A, Ciglenečki I (2009). Reduced sulfur and iron species in anoxic water column of meromictic crater Lake Pavin (Massif Central, France). *Chemical Geology* **266:** 311-317.

Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M *et al* (2010). Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *Science* **328:** 723-725.

Burrows K, Cornish A, Scott D, Higgins I (1984). Substrate specificities of the soluble and particulate methane mono-oxygenases of *Methylosinus trichosporium* Ob3B. *Journal of General Microbiology* **130:** 3327-3333.

Cahnmann H (1955). Detection and qantitative determination of benzo(a)pyrene in American shale oil. *Analytical chemistry* **27:** 1235-1240.

Call DR, Bakko MK, Krug MJ, Roberts MC (2003). Identifying Antimicrobial Resistance Genes with DNA Microarrays. *Antimicrobial Agents and Chemotherapy* **47:** 3290-3295.

Casamayor EO, Muyzer G, Pedrós-Alió C (2001). Composition and temporal dynamics of planktonic archaeal assemblages from anaerobic sulfurous environments studied by 16S rDNA denaturing gradient gel electrophoresis and sequencing. *Aquatic Microbial Ecology* **25:** 237-246.

Casamayor EO, Massana R, Benlloch S, Øvreås L, Díez B, Goddard V *et al* (2002). Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environmental microbiology* **4:** 338-348.

Castro H, Ogram A, Reddy KR (2004). Phylogenetic Characterization of Methanogenic Assemblages in Eutrophic and Oligotrophic Areas of the Florida Everglades. *Applied and Environmental Microbiology* **70:** 6559-6568.

Cerniglia C (1992). Biodegradation of polycyclic aromatic hydrocarbons. *Biodegradation* **3:** 351-368.

Chaban B, Ng SYM, Jarrell KF (2006). Archaeal habitats — from the extreme to the ordinary. *Canadian Journal of Microbiology* **52:** 73-116.

Chang H, Jackson D, Kayne P, Ross-Macdonald P, Ryseck R, Siemers N (2011). Exome sequencing reveals comprehensive genomic alterations across eight cancer cell lines. *PLoS One* **6:** e21097.

Chang W, Um Y, Holoman TRP (2006). Polycyclic Aromatic Hydrocarbon (PAH) Degradation Coupled to Methanogenesis. *Biotechnology Letters* **28:** 425-430.

Cheng W, Yagi K, Sakai H, Xu H, Kobayashi K (2005). Changes in concentration and δ13C value of dissolved CH4, CO2 and organic carbon in rice paddies under ambient and elevated concentrations of atmospheric CO2. *Organic Geochemistry* **36:** 813-823.

Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology* **30:** 344-348.

Chew YV, Holmes AJ (2009). Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest. *Journal of Microbiological Methods* **78:** 136-143.

Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR *et al* (2011). The Origin of the Haitian Cholera Outbreak Strain. *New England Journal of Medicine* **364:** 33-42.

Chin KJ, Lueders T, Friedrich MW, Klose M, Conrad R (2004). Archaeal Community Structure and Pathway of Methane Formation on Rice Roots. *Microbial Ecology* **47:** 59-67.

Chong S, Liu Y, Cummins M, Valentine D, Boone D (2002). Methanogenium marinum sp. nov., a H2-using methanogen from Skan Bay, Alaska, and kinetics of H2 utilization. *Antonie Van Leeuwenhoek* **81:** 263-270.

Chou CC (2004). Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic acids research* **32:** e99-e99.

Chou CC, Chen CH, Lee TT, Peck K (2004). Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Res* **32:** e99.

Chou LS, Liu CSJ, Boese B, Zhang X, Mao R (2009). DNA Sequence Capture and Enrichment by Microarray Followed by Next-Generation Sequencing for Targeted Resequencing: Neurofibromatosis Type 1 Gene as a Model. *Clinical Chemistry* **56:** 62-72.

Chouari R, Le Paslier D, Daegelen P, Ginestet P, Weissenbach J, Sghir A (2005). Novel predominant archaeal and bacterial groups revealed by molecular analysis of an anaerobic sludge digester. *Environmental microbiology* **7:** 1104-1115.

Christensen N, Batstone D, He Z, Angelidaki I, Schmidt J (2004). Removal of polycyclic aromatic hydrocarbons (PAHs) from sewage sludge by anaerobic degradation. *Water science and technology : a journal of the International Association on Water Pollution Research* **50:** 237-244.

Clarke J, Wu H, Jayasinghe L, Patel A, Reid S, Bayley H (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature nanotechnology* **4:** 265-270.

Colby J, Stirling D, Dalton H (1977). The soluble methane mono-oxygenase of *Methylococcus capsulatus* (Bath). Its ability to oxygenate n-alkanes, n-alkenes, ethers, and alicyclic, aromatic and heterocyclic compounds. *The Biochemical journal* **165:** 395-402.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ *et al* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic acids research* **37:** D141-D145.

Conrad R (1999). Contribution of hydrogen to methane production and control of hydrogen concentrations in methanogenic soils and sediments. *FEMS Microbiology Ecology* **28:** 193-202.

Conrad R, Erkel C, Liesack W (2006). Rice Cluster I methanogens, an important group of Archaea producing greenhouse gas in soil. *Current Opinion in Biotechnology* **17:** 262-267.

Conrad R (2009). The global methane cycle: recent advances in understanding the microbial processes involved. *Environmental Microbiology Reports* **1:** 285-292.

Cook KL, Sayler GS (2003). Environmental application of array technology: promise, problems and practicalities. *Current Opinion in Biotechnology* **14:** 311-318.

Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, Luikart G (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12:** 347.

Costello A, Lidstrom M (1999). Molecular characterization of functional and phylogenetic genes from natural populations of methanotrophs in lake sediments. *Applied and Environmental Microbiology* **65:** 5066-5074.

Cruz-Martínez K, Suttle KB, Brodie EL, Power ME, Andersen GL, Banfield JF (2009). Despite strong seasonal responses, soil microbial consortia are more resilient to long-term changes in rainfall than overlying grassland. *The ISME Journal* **3:** 738-744.

Curtis TP, Sloan WT, Scannell JW (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 10494-10499.

D'Ascenzo M, Meacham C, Kitzman J, Middle C, Knight J, Winer R *et al* (2009). Mutation discovery in the mouse using genetically guided array capture and resequencing. *Mammalian Genome* **20:** 424-436.

Dahl F (2005). Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic acids research* **33:** e71-e71.

Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, Nilsson M *et al* (2007). Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proceedings of the National Academy of Sciences* **104:** 9387-9392.

Daiger SP, Sullivan LS, Bowne SJ, Birch DG, Heckenlively JR, Pierce EA *et al* (2010). Targeted High-Throughput DNA Sequencing for Gene Discovery in Retinitis Pigmentosa **664:** 325-331.

Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG (2006). The Affymetrix GeneChip® Platform: An Overview. *Methods in enzymology* **410:** 3-28.

Daniels L, Fuchs G, Thauer R, Zeikus J (1977). Carbon monoxide oxidation by methanogenic bacteria. *Journal of bacteriology* **132:** 118-126.

Davidova IA, Gieg LM, Duncan KE, Suflita JM (2007). Anaerobic phenanthrene mineralization by a carboxylating sulfate-reducing bacterial enrichment. *The ISME Journal* **1:** 436-442.

Davies J, Evans W (1964). Oxidative metabolism of naphthalene by soil pseudomonads. The ring-fission mechanism. *The Biochemical journal* **91:** 251-261.

De Long SK, Kinney KA, Kirisits MJ (2007). Prokaryotic Suppression Subtractive Hybridization PCR cDNA Subtraction, a Targeted Method To Identify Differentially Expressed Genes. *Applied and Environmental Microbiology* **74:** 225-232.

DeAngelis KM, Brodie EL, DeSantis TZ, Andersen GL, Lindow SE, Firestone MK (2008). Selective progressive response of soil microbial community to wild oat roots. *The ISME Journal* **3:** 168-178.

Dedysh S, Liesack W, Khmelenina V, Suzina N, Trotsenko Y, Semrau J *et al* (2000). *Methylocella palustris* gen. nov., sp. nov., a new methane-oxidizing acidophilic bacterium from peat bogs, representing a novel subtype of serine-pathway methanotrophs. *International journal of systematic and evolutionary microbiology* **50:** 955-969.

Dedysh S, Khmelenina V, Suzina N, Trotsenko Y, Semrau J, Liesack W *et al* (2002). *Methylocapsa acidiphila* gen. nov., sp. nov., a novel methane-oxidizing and dinitrogen-fixing acidophilic bacterium from Sphagnum bog. *International journal of systematic and evolutionary microbiology* **52:** 251-261.

Dedysh SN, Knief C, Dunfield PF (2005). Methylocella Species Are Facultatively Methanotrophic. *Journal of bacteriology* **187:** 4665-4670.

Dedysh SN, Belova SE, Bodelier PLE, Smirnova KV, Khmelenina VN, Chidthaisong A *et al* (2007). Methylocystis heyeri sp. nov., a novel type II methanotrophic bacterium possessing 'signature' fatty acids of type I methanotrophs. *International Journal of Systematic and Evolutionary Microbiology* **57:** 472-479.

Delmont TO, Robe P, Cecillon S, Clark IM, Constancias F, Simonet P *et al* (2010). Accessing the Soil Metagenome for Studies of Microbial Diversity. *Applied and Environmental Microbiology* **77:** 1315-1324.

Delmont TO, Prestat E, Keegan KP, Faubladier M, Robe P, Clark IM *et al* (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *The ISME Journal* **6:** 1677-1687.

DeLong E, Wickham G, Pace N (1989). Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science* **243:** 1360-1363.

Denef V, Park J, Rodrigues J, Tsoi T, Hashsham S, Tiedje J (2003). Validation of a more sensitive method for using spotted oligonucleotide DNA microarrays for functional genomics studies on bacterial communities. *Environmental microbiology* **5:** 933-943.

Denman SE, Tomkins NW, McSweeney CS (2007). Quantitation and diversity analysis of ruminal methanogenic populations in response to the antimethanogenic compound bromochloromethane. *FEMS Microbiology Ecology* **62:** 313-322.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al* (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology* **72:** 5069-5072.

DeSantis TZ, Brodie EL, Moberg JP, Zubieta IX, Piceno YM, Andersen GL (2007). High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library When Sampling the Environment. *Microbial Ecology* **53:** 371-383.

Dhillon A, Lever M, Lloyd KG, Albert DB, Sogin ML, Teske A (2005). Methanogen Diversity Evidenced by Molecular Characterization of Methyl Coenzyme M Reductase A (mcrA) Genes in Hydrothermal Sediments of the Guaymas Basin. *Applied and Environmental Microbiology* **71:** 4592-4601.

Diatchenko L, Lau Y, Campbell A, Chenchik A, Moqadam F, Huang B *et al* (1996). Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences of the United States of America* **93:** 6025-6030.

Díaz E (2004). Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility. *International microbiology : the official journal of the Spanish Society for Microbiology* **7:** 173-180.

Ding G-C, Heuer H, He Z, Xie J, Zhou J, Smalla K (2012). More functional genes and convergent overall functional patterns detected by geochip in phenanthrene-spiked soils. *FEMS Microbiology Ecology***:** n/a-n/a.

Dlugokencky EJ, Bruhwiler L, White JWC, Emmons LK, Novelli PC, Montzka SA *et al* (2009). Observational constraints on recent increases in the atmospheric CH4burden. *Geophysical Research Letters* **36:** doi:10.1029/2009GL039780.

Doerfert SN, Reichlen M, Iyer P, Wang M, Ferry JG (2009). Methanolobus zinderi sp. nov., a methylotrophic methanogen isolated from a deep subsurface coal seam. *International Journal of Systematic and Evolutionary Microbiology* **59:** 1064-1069.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research* **36:** e105-e105.

Dolfing J, Xu A, Gray N, Larter S, Head I (2009). The thermodynamic landscape of methanogenic PAH degradation. *Microbial biotechnology* **2:** 566-574.

Doyle E, Muckian L, Hickey AM, Clipson N (2008). Chapter 2 Microbial PAH Degradation **65:** 27-66.

Duc L, Neuenschwander S, Rehrauer H, Wagner U, Sobek J, Schlapbach R *et al* (2009). Development and experimental validation of anifHoligonucleotide microarray to study diazotrophic communities in a glacier forefield. *Environmental microbiology* **11:** 2179-2189.

Dufva M (2005). Fabrication of high quality microarrays. *Biomolecular Engineering* **22:** 173-184.

Dufva M (2009). Fabrication of DNA Microarray. *Methods in molecular biology* **529:** 63-79.

Dugat-Bony E, Missaoui M, Peyretaillade E, Biderre-Petit C, Bouzid O, Gouinaud C *et al* (2011). HiSpOD: probe design for functional DNA microarrays. *Bioinformatics* **27:** 641-648.

Dugat-Bony E, Biderre-Petit C, Jaziri F, David M, Denonfoux J, Lyon D *et al* (2012a). In situ TCE degradation mediated by complex dehalorespiring communities during biostimulation processes. *Microbial biotechnology***:** doi: 10.1111/j.1751-7915.2012.00339.x.

Dugat-Bony E, Peyretaillade E, Parisot N, Biderre-Petit C, Jaziri F, Hill D *et al* (2012b). Detecting unknown sequences with DNA microarrays: explorative probe design strategies. *Environmental microbiology* **14:** 356-371.

Dunbar J, Barns SM, Ticknor LO, Kuske CR (2002). Empirical and Theoretical Bacterial Diversity in Four Arizona Soils. *Applied and Environmental Microbiology* **68:** 3035-3045.

Dunfield PF, Yuryev A, Senin P, Smirnova AV, Stott MB, Hou S *et al* (2007). Methane oxidation by an extremely acidophilic bacterium of the phylum Verrucomicrobia. *Nature* **450:** 879-882.

Dunfield PF, Belova SE, Vorob'ev AV, Cornish SL, Dedysh SN (2010). Methylocapsa aurea sp. nov., a facultative methanotroph possessing a particulate methane monooxygenase, and emended description of the genus Methylocapsa. *International Journal of Systematic and Evolutionary Microbiology* **60:** 2659-2664.

Earl J, Hall G, Pickup R, Ritchie D, Edwards C (2003). Analysis of methanogen diversity in a hypereutrophic lake using PCR-RFLP analysis of *mcr* sequences. *Microbial ecology* **46:** 270-278.

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM *et al* (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics* **7:** 57.

Ehrenreich A (2006). DNA microarray technology for the microbiologist: an overview. *Applied Microbiology and Biotechnology* **73:** 255-273.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G *et al* (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323:** 133-138.

Eisen J (2007). Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biology* **5:** e82.

Eisen JA, Dethlefsen L, Huse S, Sogin ML, Relman DA (2008). The Pervasive Effects of an Antibiotic on the Human Gut Microbiota, as Revealed by Deep 16S rRNA Sequencing. *PLoS Biology* **6:** e280.

Eisenstein M (2012). Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology* **30:** 295-296.

Epelde L, Becerril JM, Kowalchuk GA, Deng Y, Zhou J, Garbisu C (2010). Impact of Metal Pollution and Thlaspi caerulescens Growth on Soil Microbial Communities. *Applied and Environmental Microbiology* **76:** 7843-7853.

Ermler U (1997). Crystal Structure of Methyl-Coenzyme M Reductase: The Key Enzyme of Biological Methane Formation. *Science* **278:** 1457-1462.

Essaid H, Bekins B, Herkelrath W, Delin G (2011). Crude oil at the Bemidji site: 25 years of monitoring, modeling, and understanding. *Ground Water* **49:** 706-726.

Ettwig KF, Butler MK, Le Paslier D, Pelletier E, Mangenot S, Kuypers MMM *et al* (2010). Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. *Nature* **464:** 543-548.

Eugène M, Massé DI, Chiquette J, Benchaar C (2008). Meta-analysis on the effects of lipid supplementation on methane production in lactating dairy cows. *Canadian Journal of Animal Science* **88:** 331-334.

Fedurco M (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research* **34:** e22-e22.

Fisher M, Triplett E (1999). Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Applied and environmental microbiology* **65:** 4630-4636.

Forster P, Ramaswamy V, Artaxo P, Berntsen T, Betts R, Fahey D *et al* (2007). Changes in Atmospheric Constituents and in Radiative Forcing. *Cambridge***:** Cambridge University Press.

Franke-Whittle IH, Goberna M, Insam H (2009a). Design and testing of real-time PCR primers for the quantification ofMethanoculleus,Methanosarcina,Methanothermobacter, and a group of uncultured methanogens. *Canadian Journal of Microbiology* **55:** 611-616.

Franke-Whittle IH, Goberna M, Pfister V, Insam H (2009b). Design and development of the ANAEROCHIP microarray for investigation of methanogenic communities. *Journal of Microbiological Methods* **79:** 279-288.

Franzmann PD, Liu Y, Balkwill DL, Aldrich HC, Conway de Macario E, Boone DR (1997). Methanogenium frigidum sp. nov., a psychrophilic, H2-using methanogen from Ace Lake, Antarctica. *International journal of systematic bacteriology* **47:** 1068-1072.

Friedrich M (2005). Methyl-Coenzyme M Reductase Genes: Unique Functional Markers for Methanogenic and Anaerobic Methane-Oxidizing Archaea. *Methods in enzymology* **397:** 428-442.

Galand P, Saarnio S, Fritze H, Yrjälä K (2002). Depth related diversity of methanogen Archaea in Finnish oligotrophic fen. *FEMS microbiology ecology* **42:** 441-449.

Galbraith EA, Antonopoulos DA, White BA (2004). Suppressive subtractive hybridization as a tool for identifying genetic diversity in an environmental metagenome: the rumen as a model. *Environmental microbiology* **6:** 928-937.

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

Garber K (2008). Fixing the front end. *Nature Biotechnology* **26:** 1101-1104.

Garcia J, Patel B, Ollivier B (2000). Taxonomic, phylogenetic, and ecological diversity of methanogenic Archaea. *Anaerobe* **6:** 205-226.

Gardner MK, Feng W-c, Archuleta J, Lin H, Mal X: Parallel genomic sequence-searching on an ad-hoc grid: experiences, lessons learned, and implications. *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*; Tampa, Florida. ACM: 2006.

Gelsomino A, Keijzer-Wolters A, Cacco G, van Elsas J (1999). Assessment of bacterial community structure in soil by polymerase chain reaction and denaturing gradient gel electrophoresis. *Journal of microbiological methods* **38:** 1-15.

Gentry TJ, Wickham GS, Schadt CW, He Z, Zhou J (2006). Microarray Applications in Microbial Ecology Research. *Microbial Ecology* **52:** 159-175.

Ghebremedhin B, Layer F, Konig W, Konig B (2008). Genetic Classification and Distinguishing of Staphylococcus Species Based on Different Partial gap, 16S rRNA, hsp60, rpoB, sodA, and tuf Gene Sequences. *Journal of Clinical Microbiology* **46:** 1019-1025.

Gibson J, S. Harwood C (2002). Metabolic diversity In aromatic compound utilization By anaerobic microbes. *Annual Review of Microbiology* **56:** 345-369.

Gieg L, Suflita J (2002). Detection of anaerobic metabolites of saturated and aromatic hydrocarbons in petroleum-contaminated aquifers. *Environmental science & technology* **36:** 3755-3762.

Gieg LM, Davidova IA, Duncan KE, Suflita JM (2010). Methanogenesis, sulfate reduction and crude oil biodegradation in hot Alaskan oilfields. *Environmental microbiology* **12:** 3074-3086.

Glenn TC (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11:** 759-769.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W *et al* (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* **27:** 182-189.

Goberna M, Gadermaier M, Garcia C, Wett B, Insam H (2010). Adaptation of Methanogenic Communities to the Cofermentation of Cattle Excreta and Olive Mill Wastes at 37 C and 55 C. *Applied and Environmental Microbiology* **76:** 6564-6571.

Good J (2011). Reduced representation methods for subgenomic enrichment and next-generation sequencing. *Methods in molecular biology* **772:** 85-103.

Gray ND, Sherry A, Hubert C, Dolfing J, Head IM (2010). Methanogenic Degradation of Petroleum Hydrocarbons in Subsurface Environments **72:** 137-161.

Griebler C, Safinowski M, Vieth A, Richnow H, Meckenstock R (2004). Combined application of stable carbon isotope analysis and specific metabolites determination for assessing in situ degradation of aromatic hydrocarbons in a tar oil-contaminated aquifer. *Environmental science & technology* **38:** 617-631.

Gundel L, Lee V, Mahanama K, Stevens R, Daisey J (1995). Direct Determination of the Phase Distributions of Semi-Volatile Polycyclic Aromatic Hydrocarbons Using Annular Denuders. *Atmospheric Environment* **29:** 1719-1733.

Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH *et al* (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences* **105:** 9145-9150.

Guschin D, Mobarry B, Proudnikov D, Stahl D, Rittmann B, Mirzabekov A (1997). Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. *Applied and environmental microbiology* **63:** 2397-2402.

Hales B, Edwards C, Ritchie D, Hall G, Pickup R, Saunders J (1996). Isolation and identification of methanogen-specific DNA from blanket bog peat by PCR amplification and sequence analysis. *Applied and Environmental Microbiology* **62:** 668-675.

Hallam SJ, Girguis PR, Preston CM, Richardson PM, DeLong EF (2003). Identification of Methyl Coenzyme M Reductase A (mcrA) Genes Associated with Methane-Oxidizing Archaea. *Applied and Environmental Microbiology* **69:** 5483-5491.

Hallam SJ (2004). Reverse Methanogenesis: Testing the Hypothesis with Environmental Genomics. *Science* **305:** 1457-1462.

Handelsman J, Rondon M, Brady S, Clardy J, Goodman R (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry and biology* **5:** R245-249.

Handelsman J (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews : MMBR* **68:** 669-685.

Hanson R, Hanson T (1996). Methanotrophic bacteria. *Microbiological reviews* **60:** 439-471.

Hardenbol P, Banér J, Jain M, Nilsson M, Namsaraev E, Karlin-Neumann G *et al* (2003). Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnology* **21:** 673-678.

Haritash AK, Kaushik CP (2009). Biodegradation aspects of Polycyclic Aromatic Hydrocarbons (PAHs): A review. *Journal of Hazardous Materials* **169:** 1-15.

Harris JK, Kelley ST, Pace NR (2004). New Perspective on Uncultured Bacterial Phylogenetic Division OP11. *Applied and Environmental Microbiology* **70:** 845-849.

He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG *et al* (2010a). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods* **7:** 807-812.

He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC *et al* (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal* **1:** 67-77.

He Z, Deng Y, Van Nostrand JD, Tu Q, Xu M, Hemme CL *et al* (2010b). GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. *The ISME Journal* **4:** 1167-1179.

He Z, Piceno Y, Deng Y, Xu M, Lu Z, DeSantis T *et al* (2011a). The phylogenetic composition and structure of soil microbial communities shifts in response to elevated carbon dioxide. *The ISME Journal* **6:** 259-272.

He Z, Van Nostrand JD, Deng Y, Zhou J (2011b). Development and applications of functional gene microarrays in the analysis of the functional diversity, composition, and structure of microbial communities. *Frontiers of Environmental Science and Engineering in China* **5:** 1-20.

He ZL, VAN NOSTRAND JD, WU LY, ZHOU JZ (2008). Development and application of functional gene arrays for microbial community analysis. *The Transactions of Nonferrous Metals Society of China* **18:** 1319-1327.

Henne A, Schmitz R, Bömeke M, Gottschalk G, Daniel R (2000). Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on Escherichia coli. *Applied and environmental microbiology* **66:** 3113-3116.

Heyer J (2005). Methylohalobius crimeensis gen. nov., sp. nov., a moderately halophilic, methanotrophic bacterium isolated from hypersaline lakes of Crimea. *International Journal of Systematic and Evolutionary Microbiology* **55:** 1817-1826.

Hinrichs K, Hayes J, Sylva S, Brewer P, DeLong E (1999). Methane-consuming archaebacteria in marine sediments. *Nature* **398:** 802-805.

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW *et al* (2007). Genome-wide in situ exon capture for selective resequencing. *Nature Genetics* **39:** 1522-1527.

Hodges E, Smith AD, Kendall J, Xuan Z, Ravi K, Rooks M *et al* (2009). High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Research* **19:** 1593-1605.

Hoehler TM, Alperin MJ, Albert DB, Martens CS (1994). Field and laboratory studies of methane oxidation in an anoxic marine sediment - evidence for a methanogen-sulfate reducer consortium. *Global Biogeochemical Cycles* **8:** 451-463.

Hoff KJ (2009). The effect of sequencing errors on metagenomic gene prediction. *BMC Genomics* **10:** 520.

Hong X, Doddapaneni H, Comeron J, Rodesch M, Halvensleben H, Nien C *et al* (2012). Microarray-Based Capture of Novel Expressed Cell Type-Specific Transfrags (CoNECT) to Annotate Tissue-Specific Transcription in Drosophila melanogaster. *G3 (Bethesda, Md)* **2:** 873-882.

Hook SE, Northwood KS, Wright ADG, McBride BW (2008). Long-Term Monensin Supplementation Does Not Significantly Affect the Quantity or Diversity of Methanogens in the Rumen of the Lactating Dairy Cow. *Applied and Environmental Microbiology* **75:** 374-380.

Hori T, Haruta S, Ueno Y, Ishii M, Igarashi Y (2006). Dynamic Transition of a Methanogenic Population in Response to the Concentration of Volatile Fatty Acids in a Thermophilic Anaerobic Digester. *Applied and Environmental Microbiology* **72:** 1623-1630.

Horz HP, Rich V, Avrahami S, Bohannan BJM (2005). Methane-Oxidizing Bacteria in a California Upland Grassland Soil: Diversity and Response to Simulated Global Change. *Applied and Environmental Microbiology* **71:** 2642-2652.

Hou S, Makarova KS, Saw JHW, Senin P, Ly BV, Zhou Z *et al* (2008). Complete genome sequence of the extremely acidophilic methanotroph isolate V4, Methylacidiphilum infernorum, a representative of the bacterial phylum Verrucomicrobia. *Biology Direct* **3:** 26.

Huang X, Li Y, Niu Q, Zhang K (2007). Suppression Subtractive Hybridization (SSH) and its modifications in microbiological research. *Applied Microbiology and Biotechnology* **76:** 753-760.

Hugenholtz P, Goebel B, Pace N (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology* **180:** 4765-4774.

Hugenholtz P (2002). Exploring prokaryotic diversity in the genomic era. *Genome biology* **3:** REVIEWS0003.

Hughes T, Mao M, Jones A, Burchard J, Marton M, Shannon K *et al* (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* **19:** 342-347.

Hunkapiller T, Kaiser R, Koop B, Hood L (1991). Large-scale and automated DNA sequence determination. *Science* **254:** 59-67.

Huyghe A, Francois P, Charbonnier Y, Tangomo-Bento M, Bonetti EJ, Paster BJ *et al* (2008). Novel Microarray Design Strategy To Study Complex Bacterial Communities. *Applied and Environmental Microbiology* **74:** 1876-1885.

Hyman E (1988). A new method of sequencing DNA. *Analytical biochemistry* **174:** 423-436.

Hysom DA, Naraghi-Arani P, Elsheikh M, Carrillo AC, Williams PL, Gardner SN (2012). Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments. *PLoS One* **7:** e34560.

Im J, Lee S-W, Yoon S, DiSpirito AA, Semrau JD (2011). Characterization of a novel facultative Methylocystis species capable of growth on methane, acetate and ethanol. *Environmental Microbiology Reports* **3:** 174-181.

Imachi H, Sakai S, Sekiguchi Y, Hanada S, Kamagata Y, Ohashi A *et al* (2008). Methanolinea tarda gen. nov., sp. nov., a methane-producing archaeon isolated from a methanogenic digester sludge. *International Journal of Systematic and Evolutionary Microbiology* **58:** 294-301.

Inagaki F, Tsunogai U, Suzuki M, Kosaka A, Machiyama H, Takai K *et al* (2004). Characterization of C1-Metabolizing Prokaryotic Communities in Methane Seep Habitats at the Kuroshima Knoll, Southern Ryukyu Arc, by Analyzing pmoA, mmoX, mxaF, mcrA, and 16S rRNA Genes. *Applied and Environmental Microbiology* **70:** 7445-7455.

Islam T, Jensen S, Reigstad LJ, Larsen O, Birkeland NK (2008). Methane oxidation at 55 C and pH 2 by a thermoacidophilic bacterium belonging to the Verrucomicrobia phylum. *Proceedings of the National Academy of Sciences* **105:** 300-304.

Islas-Lima S, Thalasso F, Gómez-Hernandez J (2004). Evidence of anoxic methane oxidation coupled to denitrification. *Water Research* **38:** 13-16.

Iwai S, Kurisu F, Urakawa H, Yagi O, Furumai H (2007). Development of a 60-mer oligonucleotide microarray on the basis of benzene monooxygenase gene diversity. *Applied Microbiology and Biotechnology* **75:** 929-939.

Iwai S, Kurisu F, Urakawa H, Yagi O, Kasuga I, Furumai H (2008). Development of an oligonucleotide microarray to detect di- and monooxygenase genes for benzene degradation in soil. *FEMS Microbiology Letters* **285:** 111-121.

Jaing C, Gardner S, McLoughlin K, Mulakken N, Alegria-Hartman M, Banda P *et al* (2008). A functional gene array for detection of bacterial virulence elements. *PLoS One* **3:** e2163.

Jaiziri F, Hill D, Parisot N, Denonfoux J, Dugat-Bony E, Peyretaillade E *et al* (2012). MetaExploArrays : a large-scale oligonucleotide probe design software for explorative DNA microarrays *13th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT 2012), Beijing, China***:** Accepted.

Jetten M, Stams A, Zehnder A (1992). Methanogenesis from acetate: a comparison of the acetate metabolism in *Methanothrix soehngenii* and *Methanosarcina* spp. *FEMS Microbiology Letters* **88:** 181-198.

Jiang T, Yang L, Jiang H, Tian G, Zhang X (2011). High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. *Science China Life Sciences* **54:** 945-952.

Juottonen H, Galand PE, Tuittila E-S, Laine J, Fritze H, Yrjala K (2005). Methanogen communities and Bacteria along an ecohydrological gradient in a northern raised bog complex. *Environmental microbiology* **7:** 1547-1557.

Juottonen H, Galand PE, Yrjälä K (2006). Detection of methanogenic Archaea in peat: comparison of PCR primers targeting the mcrA gene. *Research in Microbiology* **157:** 914-921.

Kafatos F, Jones C, Efstratiadis A (1979). Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic acids research* **7:** 1541-1552.

Kane M, Jatkoe T, Stumpf C, Lu J, Thomas J, Madore S (2000). Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic acids research* **28:** 4552-4557.

Karthikeyan R, Bhandari A (2001). Anaerobic biotransformation of aromatic and polycyclic aromatic hydrocarbons in soil microcosms: a review. *Journal of Hazardous Substance Research* **3:** 1-19.

Kaserer H (1905). Ueber die oxydation des wasserstofes und des methane durch mikroorganismen. *Z landw Versuchsw in Osterreich* **8:** 789-792.

Kawasaki E (2006). The end of the microarray Tower of Babel: will universal standards lead the way? *Journal of biomolecular techniques : JBT* **17:** 200-206.

Keith L, Telliard W (1976). Priority Pollutants: I. A Perspective View. *Environmental Science and Technology* **13:** 416-423.

Kelly JJ, Siripong S, McCormack J, Janus LR, Urakawa H, El Fantroussi S *et al* (2005). DNA microarray detection of nitrifying bacterial 16S rRNA in wastewater treatment plant samples. *Water Research* **39:** 3229-3238.

Kelly LC, Cockell CS, Piceno YM, Andersen GL, Thorsteinsson T, Marteinsson V (2010). Bacterial Diversity of Weathered Terrestrial Icelandic Volcanic Glasses. *Microbial Ecology* **60:** 740-752.

Kelly LC, Cockell CS, Herrera-Belaroussi A, Piceno Y, Andersen G, DeSantis T *et al* (2011). Bacterial Diversity of Terrestrial Crystalline Volcanic Rocks, Iceland. *Microbial Ecology* **62:** 69-79.

Kemnitz D, Chin K-J, Bodelier P, Conrad R (2004). Community analysis of methanogenic archaea within a riparian flooding gradient. *Environmental microbiology* **6:** 449-461.

Keppler F, Hamilton JTG, Braß M, Röckmann T (2006). Methane emissions from terrestrial plants under aerobic conditions. *Nature* **439:** 187-191.

Khan AA, Wang RF, Cao WW, Doerge DR, Wennerstrom D, Cerniglia CE (2001). Molecular Cloning, Nucleotide Sequence, and Expression of Genes Encoding a Polycyclic Aromatic Ring Dioxygenase from Mycobacterium sp. Strain PYR-1. *Applied and Environmental Microbiology* **67:** 3577-3585.

Kim SJ, Kweon O, Freeman JP, Jones RC, Adjei MD, Jhoo JW *et al* (2006a). Molecular Cloning and Expression of Genes Encoding a Novel Dioxygenase Involved in Low- and High-Molecular-Weight Polycyclic Aromatic Hydrocarbon Degradation in Mycobacterium vanbaalenii PYR-1. *Applied and Environmental Microbiology* **72:** 1045-1054.

Kim SJ, Kweon O, Jones RC, Freeman JP, Edmondson RD, Cerniglia CE (2006b). Complete and Integrated Pyrene Degradation Pathway in Mycobacterium vanbaalenii PYR-1 Based on Systems Biology. *Journal of bacteriology* **189:** 464-472.

Kirk JL, Beaudette LA, Hart M, Moutoglis P, Klironomos JN, Lee H *et al* (2004). Methods of studying soil microbial diversity. *Journal of Microbiological Methods* **58:** 169-188.

Kitamura K, Fujita T, Akada S, Tonouchi A (2011). Methanobacterium kanagiense sp. nov., a hydrogenotrophic methanogen, isolated from rice-field soil. *International Journal of Systematic and Evolutionary Microbiology* **61:** 1246-1262.

Kleikemper J, Pombo SA, Schroth MH, Sigler WV, Pesaro M, Zeyer J (2005). Activity and Diversity of Methanogens in a Petroleum Hydrocarbon-Contaminated Aquifer. *Applied and Environmental Microbiology* **71:** 149-158.

Kleinow K, James M, Tong Z, Venugopalan C (1998). Bioavailability and biotransformation of benzo(a)pyrene in an isolated perfused In situ catfish intestinal preparation. *Environmental health perspectives* **106:** 155-166.

Knittel K, Losekann T, Boetius A, Kort R, Amann R (2005). Diversity and Distribution of Methanotrophic Archaea at Cold Seeps. *Applied and Environmental Microbiology* **71:** 467-479.

Knittel K, Boetius A (2009). Anaerobic Oxidation of Methane: Progress with an Unknown Process. *Annual Review of Microbiology* **63:** 311-334.

Konstantinidis KT, Ramette A, Tiedje JM (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361:** 1929-1940.

Kopecky J, Kyselkova M, Omelka M, Cermak L, Novotna J, Grundmann GL *et al* (2011). Actinobacterial community dominated by a distinct clade in acidic soil of a waterlogged deciduous forest. *FEMS Microbiology Ecology* **78:** 386-394.

Korbel JO, Jensen LJ, von Mering C, Bork P (2004). Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology* **22:** 911-917.

Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ *et al* (2010). Real-Time DNA Sequencing from Single Polymerase Molecules **472:** 431-455.

Korves T, Piceno Y, Tom L, Desantis T, Jones B, Andersen G *et al* (2012). Bacterial communities in commercial aircraft high-efficiency particulate air (HEPA) filters assessed by PhyloChip analysis. *Indoor air***:** doi: 10.1111/j.1600-0668.2012.00787.x.

Krüger M, Meyerdierks A, Glöckner F, Amann R, Widdel F, Kube M *et al* (2003). A conspicuous nickel protein in microbial mats that oxidize methane anaerobically. *Nature* **426:** 878-881.

Kump LR, Kasting JF, Crane RG (1999). The Earth System. *Prentice Hall, Upper Saddle River, NJ*.

Kurr M, Huber R, König H, Jannasch HW, Fricke H, Trincone A *et al* (1991). Methanopyrus kandleri, gen. and sp. nov. represents a novel group of hyperthermophilic methanogens, growing at 110°C. *Archives Of Microbiology* **156:** 239-247.

Kweon O, Kim SJ, Jones RC, Freeman JP, Adjei MD, Edmondson RD *et al* (2007). A Polyomic Approach To Elucidate the Fluoranthene-Degradative Pathway in Mycobacterium vanbaalenii PYR-1. *Journal of bacteriology* **189:** 4635-4647.

Kyselková M, Kopecký J, Felföldi T, Čermák L, Omelka M, Grundmann GL *et al* (2008). Development of a 16S rRNA gene-based prototype microarray for the detection of selected actinomycetes genera. *Antonie van Leeuwenhoek* **94:** 439-453.

Lee D, Zo Y, Kim S (1996). Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism. *Applied and environmental microbiology* **62:** 3112-3120.

Lee H, O'Connor BD, Merriman B, Funari VA, Homer N, Chen Z *et al* (2009). Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. *BMC Genomics* **10:** 646.

Lee HJ, Kim JM, Lee SH, Park M, Lee K, Madsen EL *et al* (2011). Gentisate 1,2-dioxygenase, in the third naphthalene catabolic gene cluster of Polaromonas naphthalenivorans CJ2, has a role in naphthalene degradation. *Microbiology* **157:** 2891-2903.

Lee N, Nielsen P, Andreasen K, Juretschko S, Nielsen J, Schleifer K *et al* (1999). Combination of fluorescent in situ hybridization and microautoradiography-a new tool for structure-function analyses in microbial ecology. *Applied and environmental microbiology* **65:** 1289-1297.

Lee PKH, Warnecke F, Brodie EL, Macbeth TW, Conrad ME, Andersen GL *et al* (2012). Phylogenetic Microarray Analysis of a Microbial Community Performing Reductive Dechlorination at a TCE-Contaminated Site. *Environmental science & technology* **46:** 1044-1054.

Lehmacher A, Klenk H (1994). Characterization and phylogeny of mcrII, a gene cluster encoding an isoenzyme of methyl coenzyme M reductase from hyperthermophilic *Methanothermus fervidus*. *Molecular & general genetics : MGG* **243:** 198-206.

Lehmann-Richter S, Grosskopf R, Liesack W, Frenzel P, Conrad R (1999). Methanogenic archaea and CO2-dependent methanogenesis on washed rice roots. *Environmental microbiology* **1:** 159-166.

Lehours AC, Batisson I, Guedon A, Mailhot G, Fonty G (2009). Diversity of Culturable Bacteria, from the Anaerobic Zone of the Meromictic Lake Pavin, Able to Perform Dissimilatory-Iron Reduction in Different in Vitro Conditions. *Geomicrobiology Journal* **26:** 212-223.

Leigh MB, Pellizari VH, Uhlík O, Sutka R, Rodrigues J, Ostrom NE *et al* (2007). Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs). *The ISME Journal* **1:** 134-148.

Lelieveld J, Crutzen PJ, Dentener FJ (1998). Changing concentration, lifetime and climate forcing of atmospheric methane *Tellus Series B-chemical and Physical Meteorology* **50:** 128-150.

Lemoine S, Combes F, Le Crom S (2009). An evaluation of custom microarray applications: the oligonucleotide design challenge. *Nucleic acids research* **37:** 1726-1739.

Levene M, Korlach J, Turner S, Foquet M, Craighead H, Webb W (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299:** 682-686.

Li T, Wu T-D, Mazéas L, Toffin L, Guerquin-Kern J-L, Leblon G *et al* (2008). Simultaneous analysis of microbial identity and function using NanoSIMS. *Environmental microbiology* **10:** 580-588.

Li W, Carpi G, Cagnacci F, Wittekindt NE, Zhao F, Qi J *et al* (2011). Metagenomic Profile of the Bacterial Communities Associated with Ixodes ricinus Ticks. *PLoS One* **6:** e25604.

Liang Y, Li G, Van Nostrand JD, He Z, Wu L, Deng Y *et al* (2009). Microarray-based analysis of microbial functional diversity along an oil contamination gradient in oil field. *FEMS Microbiology Ecology* **70:** 324-333.

Liang Y, Van Nostrand JD, Deng Y, He Z, Wu L, Zhang X *et al* (2010). Functional gene diversity of soil microbial communities from five oil-contaminated fields in China. *The ISME Journal* **5:** 403-413.

Liang Y, Van Nostrand JD, N'Guessan LA, Peacock AD, Deng Y, Long PE *et al* (2012). Microbial Functional Gene Diversity with a Shift of Subsurface Redox Conditions during In Situ Uranium Reduction. *Applied and Environmental Microbiology* **78:** 2966-2972.

Ligocki M, Pankow J (1989). Measurements of the gas/particule distributions of atmospheric organic compounds. *Environmental science & technology* **23:** 75-83.

Liles MR, Turkmen O, Manske BF, Zhang M, Rouillard J-M, George I *et al* (2010). A phylogenetic microarray targeting 16S rRNA genes from the bacterial division Acidobacteria reveals a lineage-specific distribution in a soil clay fraction. *Soil Biology and Biochemistry* **42:** 739-747.

Lin C, Huang X, Kolbanovskii A, Hingerty B, Amin S, Broyde S *et al* (2001). Molecular topology of polycyclic aromatic carcinogens determines DNA adduct conformation: a link to tumorigenic activity. *Journal of molecular biology* **306:** 1059-1080.

Lipscomb J (1994). Biochemistry of the soluble methane monooxygenase. *Annual review of microbiology* **48:** 371-399.

Liu W, Marsh T, Cheng H, Forney L (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and environmental microbiology* **63:** 4516-4522.

Liu Y, Whitman WB (2008). Metabolic, Phylogenetic, and Ecological Diversity of the Methanogenic Archaea. *Annals of the New York Academy of Sciences* **1125:** 171-189.

Lomans PB, Maas R, Luderer R, Op Den Camp HJM, Pol A, Van Der Drift C *et al* (1999). Isolation and Characterization of Methanomethylovorans hollandica gen. nov., sp. nov., Isolated from Freshwater Sediment, a Methylotrophic Methanogen Able To Grow on Dimethyl Sulfide and Methanethiol. *Applied and Environmental Microbiology* **65:** 3641-3650.

Lovett M, Kere J, Hinton L (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proceedings of the National Academy of Sciences of the United States of America* **88:** 9628-9632.

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J *et al* (2002a). Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* **68:** 5064-5081.

Loy A, Lehner A, Lee N, Adamczyk J, Meier H, Ernst J *et al* (2002b). Oligonucleotide Microarray for 16S rRNA Gene-Based Detection of All Recognized Lineages of Sulfate-Reducing Prokaryotes in the Environment. *Applied and Environmental Microbiology* **68:** 5064-5081.

Loy A, Schulz C, Lucker S, Schopfer-Wendels A, Stoecker K, Baranyi C *et al* (2005). 16S rRNA Gene-Based Oligonucleotide Microarray for Environmental Monitoring of the Betaproteobacterial Order "Rhodocyclales". *Applied and Environmental Microbiology* **71:** 1373-1386.

Loy A, Bodrossy L (2006). Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clinica Chimica Acta* **363:** 106-119.

Lu Z, He Z, Parisi VA, Kang S, Deng Y, Van Nostrand JD *et al* (2012). GeoChip-Based Analysis of Microbial Functional Gene Diversity in a Landfill Leachate-Contaminated Aquifer. *Environmental science & technology***:** 120523162719004.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar *et al* (2004). ARB: a software environment for sequence data. *Nucleic acids research* **32:** 1363-1371.

Lueders T, Chin K, Conrad R, Friedrich M (2001). Molecular analyses of methyl-coenzyme M reductase alpha-subunit (*mcrA*) genes in rice field soil and enrichment cultures reveal the methanogenic phenotype of a novel archaeal lineage. *Environmental microbiology* **3:** 194-204.

Lueders T, Pommerenke B, Friedrich MW (2004). Stable-Isotope Probing of Microorganisms Thriving at Thermodynamic Limits: Syntrophic Propionate Oxidation in Flooded Soil. *Applied and Environmental Microbiology* **70:** 5778-5786.

Luesken FA, Sanchez J, van Alen TA, Sanabria J, Op den Camp HJM, Jetten MSM *et al* (2011). Simultaneous Nitrite-Dependent Anaerobic Methane and Ammonium Oxidation Processes. *Applied and Environmental Microbiology* **77:** 6802-6807.

Luesken FA, Wu ML, Op den Camp HJM, Keltjens JT, Stunnenberg H, Francoijs K-J *et al* (2012). Effect of oxygen on the anaerobic methanotroph 'Candidatus Methylomirabilis oxyfera': kinetic and transcriptional analysis. *Environmental microbiology* **14:** 1024-1034.

Luton P, Wayne J, Sharp R, Riley P (2002). The *mcrA* gene as an alternative to 16S rRNA in the phylogenetic analysis of methanogen populations in landfill. *Microbiology* **148:** 3521-3530.

Machmüller A (2006). Medium-chain fatty acids and their potential to reduce methanogenesis in domestic ruminants. *Agriculture, Ecosystems & Environment* **112:** 107-114.

Mamanova L, Coffey A, Scott C, Kozarewa I, Turner E, Kumar A *et al* (2010). Target-enrichment strategies for next-generation sequencing. *Nature Methods* **7:** 111-118.

Mandalakis M, Gustafsson O, Alsberg T, Egebäck A, Reddy C, Xu L *et al* (2005). Contribution of biomass burning to atmospheric polycyclic aromatic hydrocarbons at three European background sites. *Environmental science & technology* **39:** 2976-2982.

Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N *et al* (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology* **30:** 349-353.

Marçais G, Kingsford C (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27:** 764-770.

Marchesi J, Weightman A, Cragg B, Parkes R, Fry J (2001). Methanogen and bacterial diversity and distribution in deep gas hydrate sediments from the Cascadia Margin as revealed by 16S rRNA molecular analysis. *FEMS Microbiology Ecology* **34:** 221-228.

Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L *et al* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-380.

Marsili L, Caruso A, Fossi M, Zanardelli M, Politi E, Focardi S (2001). Polycyclic aromatic hydrocarbons (PaHs) in subcutaneous biopsies of Mediterranean cetaceans. *Chemosphere* **44:** 147-154.

Martens C, Berner R (1974). Methane production in interstitial waters of sulfate-depleted marine sediments. *Science* **185:** 1167-1169.

Martin C, Morgavi DP, Doreau M, Jouany JP (2006). Comment réduire la production de méthane chez les ruminants ? *Fourrages* **187:** 283-300.

Mathrani IM, Boone DR, Mah RA, Fox GE, Lau PP (1988). Methanohalophilus zhilinae sp. nov. , an Alkaliphilic, Halophilic, Methylotrophic Methanogen. *International Journal of Systematic Bacteriology* **38:** 139-142.

McCarthy A (2010). Third Generation DNA Sequencing: Pacific Biosciences' Single Molecule Real Time Technology. *Chemistry & Biology* **17:** 675-676.

McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A *et al* (2011). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal* **6:** 610-618.

McDonald I, Bodrossy L, Chen Y, Murrell J (2008). Molecular ecology techniques for the study of aerobic methanotrophs. *Applied and environmental microbiology* **74:** 1305-1315.

McNally B, Singer A, Yu Z, Sun Y, Weng Z, Meller A (2010). Optical Recognition of Converted DNA Nucleotides for Single-Molecule DNA Sequencing Using Nanopore Arrays. *Nano Letters* **10:** 2237-2244.

Meckenstock RU, Safinowski M, Griebler C (2004). Anaerobic degradation of polycyclic aromatic hydrocarbons. *FEMS Microbiology Ecology* **49:** 27-36.

Meckenstock RU, Mouttaki H (2011). Anaerobic degradation of non-substituted aromatic hydrocarbons. *Current Opinion in Biotechnology* **22:** 406-414.

Mertes F, ElSharawy A, Sauer S, van Helvoort JMLM, van der Zaag PJ, Franke A *et al* (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings in Functional Genomics* **10:** 374-386.

Metzker ML (2005). Emerging technologies in DNA sequencing. *Genome Research* **15:** 1767-1776.

Metzker ML (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics* **11:** 31-46.

Meyer QC, Burton SG, Cowan DA (2007). Subtractive hybridization magnetic bead capture: A new technique for the recovery of full-length ORFs from the metagenome. *Biotechnology Journal* **2:** 36-40.

Meyerdierks A, Kube M, Kostadinov I, Teeling H, Glöckner FO, Reinhardt R *et al* (2010). Metagenome and mRNA expression analyses of anaerobic methanotrophic archaea of the ANME-1 group. *Environmental microbiology* **12:** 422-439.

Michard G, Viollier E, Jezequel D, Sarazin G (1994). Geochemical study of a crater lake: Pavin

Lake, France -- Identification, location and quantification of the chemical reactions in the lake. *Chemical Geology* **115:** 103-115.

Mihajlovski A, Alric M, Brugère J-F (2008). A putative new order of methanogenic Archaea inhabiting the human gut, as revealed by molecular analyses of the mcrA gene. *Research in Microbiology* **159:** 516-521.

Mihajlovski A, Doré J, Levenez F, Alric M, Brugère J-F (2010). Molecular evaluation of the human gut methanogenic archaeal microbiota reveals an age-associated increase of the diversity. *Environmental Microbiology Reports* **2:** 272-280.

Milferstedt K, Youngblut ND, Whitaker RJ (2010). Spatial structure and persistence of methanogen populations in humic bog lakes. *The ISME Journal* **4:** 764-776.

Militon C, Rimour S, Missaoui M, Biderre C, Barra V, Hill D *et al* (2007). PhylArray: phylogenetic probe design algorithm for microarray. *Bioinformatics* **23:** 2550-2557.

Miller M, Wasik S, Huang G, Shiu W, Mackay D (1985). Relationships between octanol-water partition coefficient and aqueous solubility. *Environmental science & technology* **19:** 522-529.

Miller SM, Tourlousse DM, Stedtfeld RD, Baushke SW, Herzog AB, Wick LM *et al* (2008). In Situ-Synthesized Virulence and Marker Gene Biochip for Detection of Bacterial Pathogens in Water. *Applied and Environmental Microbiology* **74:** 2200-2209.

Miller T, Wolin M (1985). *Methanosphaera stadtmaniae* gen. nov., sp. nov.: a species that forms methane by reducing methanol with hydrogen. *Archives Of Microbiology* **141:** 116-122.

Moletta R (2002). Procédés biologiques anaérobies, Dans Gestion des problèmes environnementaux dans les industries agroalimentaires. *Technique et documentation***:** Editions Lavoisier, Paris.

Morales SE, Holben WE (2011). Linking bacterial identities and ecosystem processes: can 'omic' analyses be more than the sum of their parts? *FEMS Microbiology Ecology* **75:** 2-16.

Moran JJ, Beal EJ, Vrentas JM, Orphan VJ, Freeman KH, House CH (2007). Methyl sulfides as intermediates in the anaerobic oxidation of methane. *Environmental microbiology* **0:** 071002213627002-???

Morasch B, Hunkeler D, Zopfi J, Temime B, Höhener P (2011). Intrinsic biodegradation potential of aromatic hydrocarbons in an alluvial aquifer – Potentials and limits of signature metabolite analysis and two stable isotope-based techniques. *Water Research* **45:** 4459-4469.

Munroe DJ, Harris TJR (2010). Third-generation sequencing fireworks at Marco Island. *Nature Biotechnology* **28:** 426-428.

Murrell JC, McDonald IR, Bourne DG (1998). Molecular methods for the study of methanotroph ecology. *FEMS Microbiology Ecology* **27:** 103-114.

Murrell JC, Gilbert B, McDonald IR (2000). Molecular biology and regulation of methane monooxygenase. *Archives Of Microbiology* **173:** 325-332.

Narihiro T, Terada T, Kikuchi K, Iguchi A, Ikeda M, Yamauchi T *et al* (2009a). Comparative Analysis of Bacterial and Archaeal Communities in Methanogenic Sludge Granules from Upflow Anaerobic Sludge Blanket Reactors Treating Various Food-Processing, High-Strength Organic Wastewaters. *Microbes and Environments* **24:** 88-96.

Narihiro T, Terada T, Ohashi A, Wu J-H, Liu W-T, Araki N *et al* (2009b). Quantitative detection of culturable methanogenic archaea abundance in anaerobic treatment systems using the sequence-specific rRNA cleavage method. *The ISME Journal* **3:** 522-535.

Narihiro T, Sekiguchi Y (2011). Oligonucleotide primers, probes and molecular methods for the environmental monitoring of methanogenic archaea. *Microbial biotechnology* **4:** 585-602.

Naumova A, Kim D-W, Nam S-H, Kim RN, Choi S-H, Park H-S (2010). Whole human exome capture for high-throughput sequencing. *Genome* **53:** 568-574.

Nemir A, David MM, Perrussel R, Sapkota A, Simonet P, Monier J-M *et al* (2010). Comparative phylogenetic microarray analysis of microbial communities in TCE-contaminated soils. *Chemosphere* **80:** 600-607.

Nercessian O, Prokofeva M, Lebedinski A, L'Haridon S, Cary C, Prieur D *et al* (2004). Design of 16S rRNA-targeted oligonucleotide probes for detecting cultured and uncultured archaeal lineages in high-temperature environments. *Environmental microbiology* **6:** 170-182.

Nettmann E, Bergmann I, Mundt K, Linke B, Klocke M (2008). Archaea diversity within a commercial biogas plant utilizing herbal biomass determined by 16S rDNA andmcrA analysis. *Journal of Applied Microbiology* **105:** 1835-1850.

Neufeld JD, Mohn WW, de Lorenzo V (2006). Composition of microbial communities in hexachlorocyclohexane (HCH) contaminated soils from Spain revealed with a habitat-specific microarray. *Environmental microbiology* **8:** 126-140.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C *et al* (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461:** 272-276.

Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A *et al* (2010). Use of Ichip for High-Throughput In Situ Cultivation of "Uncultivable" Microbial Species. *Applied and Environmental Microbiology* **76:** 2445-2450.

Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary B, Landegren U (1994). Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* **265:** 2085-2088.

Nölling J, Elfner A, Palmer J, Steigerwald V, Pihl T, Lake J *et al* (1996). Phylogeny of Methanopyrus kandleri based on methyl coenzyme M reductase operons. *International journal of systematic bacteriology* **46:** 1170-1173.

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

Nunoura T, Oida H, Toki T, Ashi J, Takai K, Horikoshi K (2006). Quantification of mcrA by quantitative fluorescent PCR in sediments from methane seep of the Nankai Trough. *FEMS Microbiology Ecology* **57:** 149-157.

Nuwaysir EF (2002). Gene Expression Analysis Using Oligonucleotide Arrays Produced by Maskless Photolithography. *Genome Research* **12:** 1749-1755.

Nyrén P, Lundin A (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical biochemistry* **151:** 504-509.

O'Brien J, Wolkin R, Moench T, Morgan J, Zeikus J (1984). Association of hydrogen metabolism with unitrophic or mixotrophic growth of Methanosarcina barkeri on carbon monoxide. *Journal of bacteriology* **158:** 373-375.

Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007). Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4:** 907-909.

Op den Camp HJM, Islam T, Stott MB, Harhangi HR, Hynes A, Schouten S *et al* (2009). Environmental, genomic and taxonomic perspectives on methanotrophicVerrucomicrobia. *Environmental Microbiology Reports* **1:** 293-306.

Oremland R, Marsh L, Polcin S (1982). Methane production and simultaneous sulphate reduction in anoxic, salt marsh sediments. *Nature* **296:** 143-145.

Oremland R, Polcin S (1982). Methanogenesis and sulfate reduction: competitive and noncompetitive substrates in estuarine sediments. *Applied and Environmental Microbiology* **44:** 1270-1276.

Orphan VJ (2001). Methane-Consuming Archaea Revealed by Directly Coupled Isotopic and Phylogenetic Analysis. *Science* **293:** 484-487.

Orphan VJ, Turk KA, Green AM, House CH (2009). Patterns of15N assimilation and growth of methanotrophic ANME-2 archaea and sulfate-reducing bacteria within structured syntrophic consortia revealed by FISH-SIMS. *Environmental microbiology* **11:** 1777-1791.

Overbeek R, Fonstein M, D'Souza M, Pusch G, Maltsev N (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* **96:** 2896-2901.

Pace NR (1997). A Molecular View of Microbial Diversity and the Biosphere. *Science* **276:** 734-740.

Pareek CS, Smoczynski R, Tretyn A (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics* **52:** 413-435.

Patel GB, Sprott GD, Fein JE (1990). Isolation and Characterization of Methanobacterium espanolae sp. nov., a Mesophilic, Moderately Acidiphilic Methanogen. *International Journal Of Systematic And Evolutionary Microbiology* **40:** 12-18.

Pathak A, Shanker R, Garg SK, Manickam N (2011). Profiling of biodegradation and bacterial 16S rRNA genes in diverse contaminated ecosystems using 60-mer oligonucleotide microarray. *Applied Microbiology and Biotechnology* **90:** 1739-1754.

Peng R-H, Xiong A-S, Xue Y, Fu X-Y, Gao F, Zhao W *et al* (2008). Microbial biodegradation of polyaromatic hydrocarbons. *FEMS microbiology reviews* **32:** 927-955.

Pennings JLA, de Wijs JLJ, Keltjens JT, van der Drift C (1997). Medium-reductant directed expression of methyl coenzymeM reductase isoenzymes in Methanobacterium thermoautotrophicum (strain ΔH). *FEBS Letters* **410:** 235-237.

Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009). Metagenomic Pyrosequencing and Microbial Identification. *Clinical Chemistry* **55:** 856-866.

Pihl T, Sharma S, Reeve J (1994). Growth phase-dependent transcription of the genes that encode the two methyl coenzyme M reductase isoenzymes and N5-methyltetrahydromethanopterin:coenzyme M methyltransferase in *Methanobacterium thermoautotrophicum* delta H. *Journal of bacteriology* **176:** 6384-6391.

Plasencia A, Bañeras L, Llirós M, Casamayor EO, Borrego C (2010). Maintenance of previously uncultured freshwater archaea from anoxic waters under laboratory conditions. *Antonie van Leeuwenhoek* **99:** 403-408.

Pol A, Heijmans K, Harhangi HR, Tedesco D, Jetten MSM, Op den Camp HJM (2007). Methanotrophy below pH 1 by a new Verrucomicrobia species. *Nature* **450:** 874-878.

Polz M, Cavanaugh C (1998). Bias in template-to-product ratios in multitemplate PCR. *Applied and environmental microbiology* **64:** 3724-3730.

Porat I, Vishnivetskaya TA, Mosher JJ, Brandt CC, Yang ZK, Brooks SC *et al* (2010). Characterization of Archaeal Community in Contaminated and Uncontaminated Surface Stream Sediments. *Microbial Ecology* **60:** 784-795.

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL *et al* (2007). Multiplex amplification of large sets of human exons. *Nature Methods* **4:** 931-936.

Pozhitkov AE, Tautz D, Noble PA (2007). Oligonucleotide microarrays: widely applied poorly understood. *Briefings in Functional Genomics and Proteomics* **6:** 141-148.

Preza D, Olsen I, Willumsen T, Boches SK, Cotton SL, Grinde B *et al* (2008). Microarray analysis of the microflora of root caries in elderly. *European Journal of Clinical Microbiology & Infectious Diseases* **28:** 509-517.

Proctor LM, Lai R, Gunsalus RP (1997). The methanogenic archaeon Methanosarcina thermophila TM-1 possesses a high-affinity glycine betaine transporter involved in osmotic adaptation. *Applied and Environmental Microbiology* **63:** 2252-2257.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J *et al* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research* **35:** 7188-7196.

Prufer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J (2008). PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics* **24:** 1530-1531.

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.

Qu W, Hashimoto Si, Morishita S (2009). Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. *Genome Research* **19:** 1309-1315.

Quince C, Curtis TP, Sloan WT (2008). The rational exploration of microbial diversity. *The ISME Journal* **2:** 997-1006.

Rademacher A, Zakrzewski M, Schlüter A, Schönberg M, Szczepanowski R, Goesmann A *et al* (2012). Characterization of microbial biofilms in a thermophilic biogas system by high-throughput metagenome sequencing. *FEMS Microbiology Ecology* **79:** 785-799.

Raghoebarsing AA, Pol A, van de Pas-Schoonen KT, Smolders AJP, Ettwig KF, Rijpstra WIC *et al* (2006). A microbial consortium couples anaerobic methane oxidation to denitrification. *Nature* **440:** 918-921.

Rajendhran J, Gunasekaran P (2008). Strategies for accessing soil metagenome for desired applications. *Biotechnology Advances* **26:** 576-590.

Rajilić-Stojanović M, Heilig HGHJ, Molenaar D, Kajander K, Surakka A, Smidt H *et al* (2009). Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environmental microbiology* **11:** 1736-1751.

Ramakrishnan B, Lueders T, Dun¢eld PF, Conrad R, Friedrich MW (2001). Archaeal community structures in rice soils from di¡erent geographical regions before and after initiation of methane production. *FEMS Microbiology Ecology* **37:** 175-186.

Ramaswamy V, Boucher O, Haigh J, Hauglustaine D, Haywood J, Myhre G *et al* (2001). Radiative forcing of climate change. In: Climate Change 2001: The Scientifi c Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA***:** 349-416.

Ramette A (2009). Quantitative Community Fingerprinting Methods for Estimating the Abundance of Operational Taxonomic Units in Natural Microbial Communities. *Applied and Environmental Microbiology* **75:** 2495-2505.

Rappé MS, Giovannoni SJ (2003). The uncultured microbial majority. *Annual Review of Microbiology* **57:** 369-394.

Rastogi G, Osman S, Vaishampayan PA, Andersen GL, Stetler LD, Sani RK (2009). Microbial Diversity in Uranium Mining-Impacted Soils as Revealed by High-Density 16S Microarray and Clone Library. *Microbial Ecology* **59:** 94-108.

Rastogi G, Osman S, Kukkadapu R, Engelhard M, Vaishampayan PA, Andersen GL *et al* (2010). Microbial and Mineralogical Characterizations of Soils Collected from the Deep Biosphere of the Former Homestake Gold Mine, South Dakota. *Microbial Ecology* **60:** 539-550.

Rastogi G, Barua S, Sani RK, Peyton BM (2011). Investigation of Microbial Populations in the Extremely Metal-Contaminated Coeur d'Alene River Sediments. *Microbial Ecology* **62:** 1-13.

Ratner AJ, Armougom F, Henry M, Vialettes B, Raccah D, Raoult D (2009). Monitoring Bacterial Community of Human Gut Microbiota Reveals an Increase in Lactobacillus in Obese Patients and Methanogens in Anorexic Patients. *PLoS One* **4:** e7125.

Ravindra, Mittal A, Van Grieken R (2001). Health risk assessment of urban suspended particulate matter with special reference to polycyclic aromatic hydrocarbons: a review. *Reviews on environmental health* **16:** 169-189.

Reeburgh WS (1976). Methane consumption in Cariaco trench waters and sediments. *Earth and Planetary Science Letters* **28:** 337-344.

Reeve J, Nölling J, Morgan R, Smith D (1997). Methanogenesis: Genes, Genomes, and Who's on First? *Journal of bacteriology* **179:** 5975-5986.

Relógio A, Schwager C, Richter A, Ansorge W, Valcárcel J (2002). Optimization of oligonucleotide-based DNA microarrays. *Nucleic acids research* **30:** e51.

Rhee SK, Liu X, Wu L, Chong SC, Wan X, Zhou J (2004). Detection of Genes Involved in Biodegradation and Biotransformation in Microbial Communities by Using 50-Mer Oligonucleotide Microarrays. *Applied and Environmental Microbiology* **70:** 4303-4317.

Riesenfeld CS, Schloss PD, Handelsman J (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. *Annual Review of Genetics* **38:** 525-552.

Rimour S, Hill D, Militon C, Peyret P (2005). GoArrays: highly dynamic and efficient microarray probe design. *Bioinformatics* **21:** 1094-1103.

Rinta-Kanto JM, Bürgmann H, Gifford SM, Sun S, Sharma S, del Valle DA *et al* (2011). Analysis of sulfur-related transcription by Roseobacter communities using a taxon-specific functional gene microarray. *Environmental microbiology* **13:** 453-467.

Rocheleau S, Greer C, Lawrence J, Cantin C, Laramee L, Guiot S (1999). Differentiation of methanosaeta concilii and methanosarcina barkeri in anaerobic mesophilic granular sludge by fluorescent In situ hybridization and confocal scanning laser microscopy. *Applied and Environmental Microbiology* **65:** 2222-2229.

Roh SW, Abell GCJ, Kim K-H, Nam Y-D, Bae J-W (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in biotechnology* **28:** 291-299.

Rohland N, Reich D (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* **22:** 939-946.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242:** 84-89.

Rondon M, August P, Bettermann A, Brady S, Grossman T, Liles M *et al* (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and environmental microbiology* **66:** 2541-2547.

Rospert S, Linder D, Ellermann J, Thauer R (1990). Two genetically distinct methyl-coenzyme M reductases in *Methanobacterium thermoautotrophicum* strain Marburg and delta H. *European journal of biochemistry / FEBS* **194:** 871-877.

Rothberg JM, Leamon JH (2008). The development and impact of 454 sequencing. *Nature Biotechnology* **26:** 1117-1124.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M *et al* (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475:** 348-352.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al* (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* **5:** e77.

Rychlik W, Spencer W, Rhoads R (1990). Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic acids research* **18:** 6409-6412.

Safinowski M, Meckenstock RU (2006). Methylation is the initial reaction in anaerobic naphthalene degradation by a sulfate-reducing enrichment culture. *Environmental microbiology* **8:** 347-352.

Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H *et al* (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230:** 1350-1354.

Sait M, Hugenholtz P, Janssen P (2002). Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. *Environmental microbiology* **4:** 654-666.

Saito A, Iwabuchi T, Harayama S (2000). A novel phenanthrene dioxygenase from Nocardioides sp. Strain KP7: expression in Escherichia coli. *Journal of bacteriology* **182:** 2134-2141.

Sakai S, Imachi H, Sekiguchi Y, Ohashi A, Harada H, Kamagata Y (2007). Isolation of Key Methanogens for Global Methane Emission from Rice Paddy Fields: a Novel Isolate

Affiliated with the Clone Cluster Rice Cluster I. *Applied and Environmental Microbiology* **73:** 4326-4331.

Sakai S, Imachi H, Hanada S, Ohashi A, Harada H, Kamagata Y (2008). Methanocella paludicola gen. nov., sp. nov., a methane-producing archaeon, the first isolate of the lineage 'Rice Cluster I', and proposal of the new archaeal order Methanocellales ord. nov. *International Journal of Systematic and Evolutionary Microbiology* **58:** 929-936.

Sakai S, Conrad R, Liesack W, Imachi H (2010). Methanocella arvoryzae sp. nov., a hydrogenotrophic methanogen isolated from rice field soil. *International Journal of Systematic and Evolutionary Microbiology* **60:** 2918-2923.

Sakai S, Ehara M, Tseng IC, Yamaguchi T, Brauer SL, Cadillo-Quiroz H *et al* (2011). Methanolinea mesophila sp. nov., a hydrogenotrophic methanogen isolated from rice field soil, and proposal of the archaeal family Methanoregulaceae fam. nov. within the order Methanomicrobiales. *International Journal of Systematic and Evolutionary Microbiology* **62:** 1389-1395.

Samanta S, Singh O, Jain R (2002). Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation. *Trends in biotechnology* **20:** 243-248.

Sanger F, Nicklen S, Coulson A (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74:** 5463-5467.

Sanguin H, Herrera A, Oger-Desfeux C, Dechesne A, Simonet P, Navarro E *et al* (2006a). Development and validation of a prototype 16S rRNA-based taxonomic microarray for Alphaproteobacteria. *Environmental microbiology* **8:** 289-307.

Sanguin H, Remenant B, Dechesne A, Thioulouse J, Vogel TM, Nesme X *et al* (2006b). Potential of a 16S rRNA-Based Taxonomic Microarray for Analyzing the Rhizosphere Effects of Maize on Agrobacterium spp. and Bacterial Communities. *Applied and Environmental Microbiology* **72:** 4302-4312.

Sanguin H, Sarniguet A, Gazengel K, Moënne-Loccoz Y, Grundmann GL (2009). Rhizosphere bacterial communities associated with disease suppressiveness stages of take-all decline in wheat monoculture. *New Phytologist* **184:** 694-707.

Sapkota AR, Berger S, Vogel TM (2009). Human Pathogens Abundant in the Bacterial Metagenome of Cigarettes. *Environmental Health Perspectives* **118:** 351-356.

Schena M, Shalon D, Davis R, Brown P (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270:** 467-470.

Schink B (1997). Energetics of syntrophic cooperation in methanogenic degradation. *Microbiology and molecular biology reviews : MMBR* **61:** 262-280.

Schloss PD, Handelsman J (2004). Status of the Microbial Census. *Microbiology and Molecular Biology Reviews* **68:** 686-691.

Schlüter A, Bekel T, Diaz NN, Dondrup M, Eichenlaub R, Gartemann K-H *et al* (2008). The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *Journal of Biotechnology* **136:** 77-90.

Schneider GF, Dekker C (2012). DNA sequencing with nanopores. *Nature Biotechnology* **30:** 326-328.

Schönhuber W, Fuchs B, Juretschko S, Amann R (1997). Improved sensitivity of whole-cell hybridization by the combination of horseradish peroxidase-labeled oligonucleotides and tyramide signal amplification. *Applied and environmental microbiology* **63:** 3268-3273.

Schönmann S, Loy A, Wimmersberger C, Sobek J, Aquino C, Vandamme P *et al* (2009). 16S rRNA gene-based phylogenetic microarray for simultaneous identification of members of the genusBurkholderia. *Environmental microbiology* **11:** 779-800.

Seo J-S, Keum Y-S, Li QX (2009). Bacterial Degradation of Aromatic Compounds. *International Journal of Environmental Research and Public Health* **6:** 278-309.

Sessitsch A, Hackl E, Wenzl P, Kilian A, Kostic T, Stralis-Pavese N *et al* (2006). Diagnostic microbial microarrays in soil ecology. *The New phytologist* **171:** 719-735.

Severgnini M, Cremonesi P, Consolandi C, Caredda G, De Bellis G, Castiglioni B (2009). ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design. *Nucleic acids research* **37:** e109.

Shendure J, Ji H (2008). Next-generation DNA sequencing. *Nature Biotechnology* **26:** 1135-1145.

Siegert M, Cichocka D, Herrmann S, Gründger F, Feisthauer S, Richnow H-H *et al* (2011). Accelerated methanogenesis from aliphatic and aromatic hydrocarbons under iron- and sulfate-reducing conditions. *FEMS Microbiology Letters* **315:** 6-16.

Singer A, McNally B, Torre RD, Meller A (2012). DNA Sequencing by Nanopore-Induced Photon Emission **870:** 99-114.

Skillman LC, Evans PN, Naylor GE, Morvan B, Jarvis GN, Joblin KN (2004). 16S ribosomal DNA-directed PCR primers for ruminal methanogens and identification of methanogens colonising young lambs. *Anaerobe* **10:** 277-285.

Slater T (1984). Free-radical mechanisms in tissue injury. *The Biochemical journal* **222:** 1-15.

Small J, Call DR, Brockman FJ, Straub TM, Chandler DP (2001). Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Applied and environmental microbiology* **67:** 4708-4716.

Smith KS, Ingram-Smith C (2007). Methanosaeta, the forgotten methanogen? *Trends in Microbiology* **15:** 150-155.

Smith R, Howes B, Garabedian S (1991). In situ measurement of methane oxidation in groundwater by using natural-gradient tracer tests. *Applied and Environmental Microbiology* **57:** 1997-2004.

Soclo H, Garrigues P, Ewald M (2000). Origin of Polycyclic Aromatic Hydrocarbons (PAHs) in Coastal Marine Sediments: Case Studies in Cotonou (Benin) and Aquitaine (France) Areas. *Marine Pollution Bulletin* **40:** 387-396.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al* (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* **103:** 12115-12120.

Sohngen N (1906). Uber bakterien welche methan ab kohlenstoffnahrung und energiequelle gerbrauchen. *Z Bakteriol Parazitenk (Infektionster)* **15:** 513-517.

Southern E (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of molecular biology* **98:** 503-517.

Sprenger W, Van Belzen M, Rosenberg J, Hackstein J, Keltjens J (2000). *Methanomicrococcus blatticola* gen. nov., sp. nov., a methanol- and methylamine-reducing methanogen from the hindgut of the cockroach Periplaneta americana. *International journal of systematic and evolutionary microbiology* **50:** 1989-1999.

Springer E, Sachs M, Woese C, Boone D (1995). Partial gene sequences for the A subunit of methyl-coenzyme M reductase (mcrI) as a phylogenetic tool for the family Methanosarcinaceae. *International journal of systematic bacteriology* **45:** 554-559.

Stams AJM, Plugge CM (2009). Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nature Reviews Microbiology* **7:** 568-577.

Steward GF, Jenkins BD, Ward BB, Zehr JP (2004). Development and Testing of a DNA Macroarray To Assess Nitrogenase (nifH) Gene Diversity. *Applied and Environmental Microbiology* **70:** 1455-1465.

Stewart FJ, Ottesen EA, DeLong EF (2010). Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. *The ISME Journal* **4:** 896-907.

Stoecker K (2006). From the Cover: Cohn's Crenothrix is a filamentous methane oxidizer with an unusual methane monooxygenase. *Proceedings of the National Academy of Sciences* **103:** 2363-2367.

Stralis-Pavese N, Abell GCJ, Sessitsch A, Bodrossy L (2011). Analysis of methanotroph community composition using a pmoA-based microbial diagnostic microarray. *Nature Protocols* **6:** 609-624.

Strous M, Jetten MSM (2004). Anaerobic Oxidation of Methane and Ammonium. *Annual Review of Microbiology* **58:** 99-117.

Summerer D (2009). Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing. *Genomics* **94:** 363-368.

Summerer D, Wu H, Haase B, Cheng Y, Schracke N, Stahler CF *et al* (2009). Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Research* **19:** 1616-1621.

Summerer D, Schracke N, Wu H, Cheng Y, Bau S, Stähler CF *et al* (2010). Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform. *Genomics* **95:** 241-246.

Suzuki M, Giovannoni S (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmental microbiology* **62:** 625-630.

Swerdlow H, Wu S, Harke H, Dovichi N (1990). Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *Journal of chromatography* **516:** 61-67.

Szeliga J, Dipple A (1998). DNA adduct formation by polycyclic aromatic hydrocarbon dihydrodiol epoxides. *Chemical research in toxicology* **11:** 1-11.

Taroncher-Oldenburg G, Griner EM, Francis CA, Ward BB (2003). Oligonucleotide Microarray for the Study of Functional Gene Diversity in the Nitrogen Cycle in the Environment. *Applied and Environmental Microbiology* **69:** 1159-1171.

Tas N, van Eekert MHA, Schraa G, Zhou J, de Vos WM, Smidt H (2009). Tracking Functional Guilds: "Dehalococcoides" spp. in European River Basins Contaminated with Hexachlorobenzene. *Applied and Environmental Microbiology* **75:** 4696-4704.

Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ *et al* (2010). Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Research* **20:** 1420-1431.

Terrat S, Peyretaillade E, Gonçalves O, Dugat-Bony E, Gravelat F, Moné A *et al* (2010). Detecting variants with Metabolic Design, a new software tool to design probes for explorative functional DNA microarray development. *BMC Bioinformatics* **11:** 478.

Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH *et al* (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* **27:** 1025-1031.

Thauer R (1998). Biochemistry of methanogenesis: a tribute to Marjory Stephenson. 1998 Marjory Stephenson Prize Lecture. *Microbiology* **144:** 2377-2406.

Thauer RK, Kaster A-K, Seedorf H, Buckel W, Hedderich R (2008). Methanogenic archaea: ecologically relevant differences in energy conservation. *Nature Reviews Microbiology* **6:** 579-591.

Thompson J, Marcelino L, Polz M (2002). Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by 'reconditioning PCR'. *Nucleic acids research* **30:** 2083-2088.

Thorsen O, Smith B, Sosa CP, Jiang K, Lin H, Peters A *et al*: Parallel genomic sequence-search on a massively parallel system. *Proceedings of the 4th international conference on Computing frontiers*; Ischia, Italy. ACM: 2007.

Tian J, Wang Y, Dong X (2009). Methanoculleus hydrogenitrophicus sp. nov., a methanogenic archaeon isolated from wetland soil. *International Journal of Systematic and Evolutionary Microbiology* **60:** 2165-2169.

Timp W, Mirsaidov UM, Deqiang W, Comer J, Aksimentiev A, Timp G (2010). Nanopore Sequencing: Electrical Measurements of the Code of Life. *IEEE Transactions on Nanotechnology* **9:** 281-294.

Tiquia S, Gurczynski S, Zholi A, Devol A (2006). Diversity of biogeochemical cycling genes from Puget Sound sediments using DNA microarrays. *Environmental technology* **27:** 1377-1389.

Tomiuk S, Hofmann K (2001). Microarray probe selection strategies. *Briefings in bioinformatics* **2:** 329-340.

Torsvik V, Goksøyr J, Daae FL (1990). High diversity in DNA of soil bacteria. *Applied and Environmental Microbiology* **53:** 782-787.

Trably E, Patureau D, Delgenes J (2003). Enhancement of polycyclic aromatic hydrocarbons removal during anaerobic treatment of urban sludge. *Water science and technology : a journal of the International Association on Water Pollution Research* **48:** 53-60.

Treccani V, Walker N, Wiltshire G (1954). The metabolism of naphthalene by soil bacteria. *Journal of general microbiology* **11:** 341-348.

Trotsenko I, Medvedkova K, Khmelenina V, Eshinimaev B (2009). Thermophilic and thermotolerant aerobic methanotrophs. *Microbiology* **78:** 435-450.

Trotsenko YA, Khmelenina VN (2005). Aerobic methanotrophic bacteria of cold ecosystems. *FEMS Microbiology Ecology* **53:** 15-26.

Turcatti G, Romieu A, Fedurco M, Tairi AP (2007). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research* **36:** e25-e25.

Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J (2009a). Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6:** 315-316.

Turner EH, Ng SB, Nickerson DA, Shendure J (2009b). Methods for Genomic Partitioning. *Annual Review of Genomics and Human Genetics* **10:** 263-284.

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

Uchiyama T, Ito K, Mori K, Tsurumaru H, Harayama S (2010). Iron-Corroding Methanogen Isolated from a Crude-Oil Storage Tank. *Applied and Environmental Microbiology* **76:** 1783-1788.

Ustek D, Sirma S, Gumus E, Arikan M, Cakiris A, Abaci N *et al* (2012). A genome-wide analysis of lentivector integration sites using targeted sequence capture and next generation sequencing technology. *Infection, Genetics and Evolution* **12:** 1349-1354.

Valencia A, Schnoes AM, Brown SD, Dodevski I, Babbitt PC (2009). Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Computational Biology* **5:** e1000605.

Valentine D (2002). Biogeochemistry and microbial ecology of methane oxidation in anoxic environments: a review. *Antonie van Leeuwenhoek* **81:** 271-282.

Valm AM, Mark WJL, Rieken CW, Hasegawa Y, Sogin M, Oldenbourg R *et al* (2011). Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. *Proceedings of the National Academy of Sciences of the United States of America* **108:** 4152-4157.

Valm AM, Mark Welch JL, Borisy GG (2012). CLASI-FISH: Principles of combinatorial labeling and spectral imaging. *Systematic and Applied Microbiology*.

Van Jaarsveld J, Van Pul W, De Leeuw F (1997). <Van-Jaarsveld_1997.pdf>. *Atmospheric Environment* **31:** 1011-1024.

Van Nostrand JD, Wu W-M, Wu L, Deng Y, Carley J, Carroll S *et al* (2009). GeoChip-based analysis of functional microbial communities during the reoxidation of a bioreduced uranium-contaminated aquifer. *Environmental microbiology* **11:** 2611-2626.

Van Nostrand JD, Wu L, Wu WM, Huang Z, Gentry TJ, Deng Y *et al* (2011). Dynamics of Microbial Community Composition and Function during In Situ Bioremediation of a Uranium-Contaminated Aquifer. *Applied and Environmental Microbiology* **77:** 3860-3869.

Varley KE, Mitra RD (2008). Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Research* **18:** 1844-1850.

Venkataraman C, Negi G, Sardar S, Rastogi R (2002). Size distributions ofpolycyclic aromatic hydrocarbons in aerosol emissions from biofuel combustion **33:** 503-518.

Venter JC (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* **304:** 66-74.

Vianna ME, Holtgraewe S, Seyfarth I, Conrads G, Horz HP (2008). Quantitative Analysis of Three Hydrogenotrophic Microbial Groups, Methanogenic Archaea, Sulfate-Reducing Bacteria, and Acetogenic Bacteria, within Plaque Biofilms Associated with Human Periodontal Disease. *Journal of bacteriology* **190:** 3779-3785.

Vieites JM, Guazzaroni Ma-E, Beloqui A, Golyshin PN, Ferrer M (2009). Metagenomics approaches in systems microbiology. *FEMS microbiology reviews* **33:** 236-255.

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. *Nature Reviews Microbiology* **7:** doi:10.1038/nrmicro2119.

von Wintzingerode F, Göbel U, Stackebrandt E (1997). Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS microbiology reviews* **21:** 213-229.

Vorobev AV, Baani M, Doronina NV, Brady AL, Liesack W, Dunfield PF *et al* (2010). Methyloferula stellata gen. nov., sp. nov., an acidophilic, obligately methanotrophic bacterium that possesses only a soluble methane monooxygenase. *International Journal of Systematic and Evolutionary Microbiology* **61:** 2456-2463.

Wagner M, Smidt H, Loy A, Zhou J (2007). Unravelling Microbial Communities with DNA-Microarrays: Challenges and Future Directions. *Microbial Ecology* **53:** 498-506.

Walter A, Knapp BA, Farbmacher T, Ebner C, Insam H, Franke-Whittle IH (2012). Searching for links in the biotic characteristics and abiotic parameters of nine different biogas plants. *Microbial biotechnology***:** doi: 10.1111/j.1751-7915.2012.00361.x.

Wang Z, Keppler F, Greule M, Hamilton JTG (2011). Non-microbial methane emissions from fresh leaves: Effects of physical wounding and anoxia. *Atmospheric Environment* **45:** 4915-4921.

Ward B, Bouskill N (2011). The utility of functional gene arrays for assessing community composition, relative abundance, and distribution of ammonia-oxidizing bacteria and archaea. *Methods in enzymology* **496:** 373-396.

Ward BB, Eveillard D, Kirshtein JD, Nelson JD, Voytek MA, Jackson GA (2007). Ammonia-oxidizing bacterial community composition in estuarine and oceanic environments assessed using a functional gene microarray. *Environmental microbiology* **9:** 2522-2538.

Wartiainen I (2006). Methylobacter tundripaludum sp. nov., a methane-oxidizing bacterium from Arctic wetland soil on the Svalbard islands, Norway (78 N). *International Journal of Systematic and Evolutionary Microbiology* **56:** 109-113.

Wei X, Ju X, Yi X, Zhu Q, Qu N, Liu T *et al* (2011). Identification of sequence variants in genetic disease-causing genes using targeted next-generation sequencing. *PLoS One* **6:** e29500.

White H (1988). Manual Oligonucleotide Synthesis Using the Phosphoramidite Method. *Methods in molecular biology* **4:** 193-213.

Whitman W, Coleman D, Wiebe W (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences of the United States of America* **95:** 6578-6583.

Widdel F (1986). Growth of Methanogenic Bacteria in Pure Culture with 2-Propanol and Other Alcohols as Hydrogen Donors. *Applied and Environmental Microbiology* **51:** 1056-1062.

Widdel F, Wolfe RS (1989). Expression of secondary alcohol dehydrogenase in methanogenic bacteria and purification of the F420-specific enzyme from *Methanogenium thermophilum* strain TCI *Archives Of Microbiology* **152:** 322-328.

Wild S, Jones K (1995). Polynuclear aromatic hydrocarbons in the United Kingdom environment: a preliminary source inventory and budget. *Environmental pollution* **88:** 91-108.

Winfield MO, Wilkinson PA, Allen AM, Barker GLA, Coghill JA, Burridge A *et al* (2012). Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnology Journal* **10:** 733-742.

Wirth R, Kovács E, Maróti G, Bagi Z, Rákhely G, Kovács KL (2012). Characterization of a biogas-producing microbial community by short-read next generation DNA sequencing. *Biotechnology for Biofuels* **5:** 41.

Woese C (1987). Bacterial evolution. *Microbiological reviews* **51:** 221-271.

Woese C, Kandler O, Wheelis M (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America* **87:** 4576-4579.

Wolber P, Collins P, Lucas A, De Witte A, Shannon K (2006). The Agilent in situ-synthesized microarray platform. *Methods in enzymology* **410:** 28-57.

Wommack KE, Bhavsar J, Ravel J (2008). Metagenomics: Read Length Matters. *Applied and Environmental Microbiology* **74:** 1453-1463.

Wright A-DG, Pimm C (2003). Improved strategy for presumptive identification of methanogens using 16S riboprinting. *Journal of Microbiological Methods* **55:** 337-349.

Wu L, Thompson DK, Li G, Hurt RA, Tiedje JM, Zhou J (2001). Development and Evaluation of Functional Gene Arrays for Detection of Selected Genes in the Environment. *Applied and Environmental Microbiology* **67:** 5780-5790.

Wu L, Thompson D, Liu X, Fields M, Bagwell C, Tiedje J *et al* (2004). Development and evaluation of microarray-based whole-genome hybridization for detection of microorganisms within the context of environmental applications. *Environmental science & technology* **38:** 6775-6782.

Wu ML, de Vries S, van Alen TA, Butler MK, Op den Camp HJM, Keltjens JT *et al* (2010). Physiological role of the respiratory quinol oxidase in the anaerobic nitrite-reducing methanotroph 'Candidatus Methylomirabilis oxyfera'. *Microbiology* **157:** 890-898.

Wu Ming L, Ettwig Katharina F, Jetten Mike SM, Strous M, Keltjens Jan T, Niftrik Laura v (2011). A new intra-aerobic metabolism in the nitrite-dependent anaerobic methane-oxidizing bacteriumCandidatus‘Methylomirabilis oxyfera’. *Biochemical Society Transactions* **39:** 243-248.

Xie J, He Z, Liu X, Van Nostrand JD, Deng Y, Wu L *et al* (2010). GeoChip-Based Analysis of the Functional Gene Diversity and Metabolic Potential of Microbial Communities in Acid Mine Drainage. *Applied and Environmental Microbiology* **77:** 991-999.

Xiong J, Wu L, Tu S, Van Nostrand JD, He Z, Zhou J *et al* (2010). Microbial Communities and Functional Genes Associated with Soil Arsenic Contamination and the Rhizosphere of the Arsenic-Hyperaccumulating Plant Pteris vittata L. *Applied and Environmental Microbiology* **76:** 7277-7284.

Xu M, Wu W-M, Wu L, He Z, Van Nostrand JD, Deng Y *et al* (2010). Responses of microbial community functional structures to pilot-scale uranium in situ bioremediation. *The ISME Journal* **4:** 1060-1070.

Yanagita K, Kamagata Y, Kawaharasaki M, Suzuki T, Nakamura Y, Minato H (2000). Phylogenetic analysis of methanogens in sheep rumen ecosystem and detection of Methanomicrobium mobile By fluorescence in situ hybridization. *Bioscience, biotechnology, and biochemistry* **64:** 1737-1742.

Yashiro Y, Sakai S, Ehara M, Miyazaki M, Yamaguchi T, Imachi H (2009). Methanoregula formicica sp. nov., a methane-producing archaeon isolated from methanogenic sludge. *International Journal of Systematic and Evolutionary Microbiology* **61:** 53-59.

Yergeau E, Schoondermark-Stolk SA, Brodie EL, Déjean S, DeSantis TZ, Gonçalves O *et al* (2008). Environmental microarray analyses of Antarctic soil microbial communities. *The ISME Journal* **3:** 340-351.

Yergeau E, Bokhorst S, Kang S, Zhou J, Greer CW, Aerts R *et al* (2011). Shifts in soil microorganisms in response to warming are consistent across a range of Antarctic environments. *The ISME Journal* **6:** 692-702.

Yu K, Zhang T (2012). Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS One* **7:** e38183.

Yu Y, Lee C, Kim J, Hwang S (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnology and bioengineering* **89:** 670-679.

Yu Y, Kim J, Hwang S (2006). Use of real-time PCR for group-specific quantification of aceticlastic methanogens in anaerobic processes: population dynamics and community structures. *biotechnology and bioengineering* **93:** 424-433.

Yu Z, Garcia-Gonzalez R, Schanbacher FL, Morrison M (2007). Evaluations of Different Hypervariable Regions of Archaeal 16S rRNA Genes in Profiling of Methanogens by Archaea-Specific PCR and Denaturing Gradient Gel Electrophoresis. *Applied and Environmental Microbiology* **74:** 889-893.

Yuan S, Chang B (2007). Anaerobic degradation of five polycyclic aromatic hydrocarbons from river sediment in Taiwan. *Journal of environmental science and health Part B, Pesticides, food contaminants, and agricultural wastes* **42:** 63-69.

Zehnder A, Brock T (1979). Methane formation and methane oxidation by methanogenic bacteria. *Journal of bacteriology* **137:** 420-432.

Zengler K (2002). Cultivating the uncultured. *Proceedings of the National Academy of Sciences* **99:** 15681-15686.

Zengler K, Walcher M, Clark G, Haller I, Toledo G, Holland T *et al* (2005). High-throughput cultivation of microorganisms using microcapsules. *Methods in enzymology* **397:** 124-130.

Zepp FK, Holliger C, Grosskopf R, Liesack W, Nozhevnikova AN, Müller B *et al* (1999). Vertical distribution of methanogens in the anoxic sediment of Rotsee (Switzerland). *Applied and Environmental Microbiology* **65:** 2402-2408.

Zhang G, Jiang N, Liu X, Dong X (2008a). Methanogenesis from Methanol at Low Temperatures by a Novel Psychrophilic Methanogen, "Methanolobus psychrophilus" sp. nov., Prevalent in Zoige Wetland of the Tibetan Plateau. *Applied and Environmental Microbiology* **74:** 6114-6120.

Zhang G, Tian J, Jiang N, Guo X, Wang Y, Dong X (2008b). Methanogen community in Zoige wetland of Tibetan plateau and phenotypic characterization of a dominant uncultured methanogen cluster ZC-I. *Environmental microbiology* **10:** 1850-1860.

Zhang S, Wang Q, Xie S (2011). Stable isotope probing identifies anthracene degraders under methanogenic conditions. *Biodegradation* **23:** 221-230.

Zhang T, Fang HHP (2006). Applications of real-time polymerase chain reaction for quantification of microorganisms in environmental samples. *Applied Microbiology and Biotechnology* **70:** 281-289.

Zhang X, LY. Y (1997). Carboxylation as an initial reaction in the anaerobic metabolism of naphthalene and phenanthrene by sulfidogenic consortia. *Applied and Environmental Microbiology* **63:** 4759-4764.

Zheng D, Raskin L (2000). Quantification of Methanosaeta Species in Anaerobic Bioreactors Using Genus- and Species-Specific Hybridization Probes. *Microbial Ecology* **39:** 246-262.

Zhou J, Thompson D (2002). Challenges in applying microarrays to environmental studies. *Current Opinion in Biotechnology* **13:** 204-207.

Zhou J (2003). Microarrays for bacterial detection and microbial community analysis. *Current opinion in microbiology* **6:** 288-294.

Zhou NY, Fuenmayor SL, Williams PA (2001). nag Genes of Ralstonia (Formerly Pseudomonas) sp. Strain U2 Encoding Enzymes for Gentisate Catabolism. *Journal of bacteriology* **183:** 700-708.

Ziv J, Lempel A (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory* **23:** 337-343.

Annexe

MetaExploArrays : a large-scale oligonucleotide probe design software for explorative DNA microarrays

Faouzi Jaziri [1,2,3], David R.C. Hill[1,2]
1 Clermont Université, Université Blaise Pascal
2 CNRS, UMR 6158, ISIMA / LIMOS, F-63173
AUBIERE, FRANCE
e-mail: David.Hill@univ-bpclermont.fr

Nicolas Parisot[3,4], Jérémie Denonfoux[3,4], Eric Dugat-Bony[3,4], Eric Peyretaillade[3,4], Pierre Peyret[3,4]
3 EA CIDAM, EA 4678, Université d'Auvergne
CRNH AUVERGNE IFR SANTE
4 Clermont Université, Université d'Auvergne
AUBIERE, FRANCE
e-mail: Pierre.Peyret@u-clermont1.fr

***Abstract*—Oligonucleotide arrays are miniaturized DNA microarrays that consist of a small solid surface onto which are spotted thousands of short single-stranded DNA fragments called oligonucleotide probes. Each probe must be specific and sensitive to hybridize only to its complementary target allowing the rapid identification and quantification of targets in complex samples. Probe design is a crucial step in successful oligonuclrotide array experiment. However, with the rapid growth of environmental databases (metagenomics programs coupled to next generation sequencing) the selection of oligonucleotide probes becomes a very difficult task. The computational capacity requirements of probe design algorithms have thus hugely increased. Consequently, the use of parallel and distributed architectures can considerably reduce the complexity and the computational time of these algorithms. In this paper we present a new efficient algorithm of oligonucleotide probe selection for an individual specific nucleic acid sequence or a group of sequences. We used a model driven engineering approach to simultaneous design of thousands of sensitive, specific, isothermal and explorative probes, on both PC, multiprocessor, cluster and grid computing.**

Keywords-meta-programming; model driven engineering; parallelization; oligonucleotide probe design; explorative probes, thermodynamic, complex environments.

I. INTRODUCTION

The latest biodiversity estimate predicts about 9 million species on our planet [11]. However, only less than 2 million species were already described scientifically. Most non-described species are microorganisms which are widely found in almost all environmental habitats present in nature, even the most extreme. They play a critical role in the ecosystems functioning. However, the study of this huge microbial biocomplexity and the great number of existing microorganisms requires the use of high-throughput molecular tools allowing simultaneous analysis.

DNA microarrays are currently one of the most promising approaches to address this task. These approaches are based on the ability of complementary strands of DNA to hybridize to one another in solution with high specificity. Oligonucleotide arrays can study the presence, or the expression levels of several thousands of genes, combining qualitative and quantitative aspects in only one experiment. Oligonucleotide arrays have been widely used for gene detection, gene expression quantification and profiling environmental communities in a flexible and easy-to-use way.

The probe selection is a difficult task that must exploit newly available high-density microarray formats. Oligonucleotide arrays can currently contain up to 4 million probes. These probes are usually between 20 and 70 bases long. Good oligonucleotide probes must have the following two properties: sensitive and specific.

The sensitivity of a probe represents its capacity to produce a strong signal if its complementary target is present in the sample hybridized. The sensitivity generally increases with probe length. However, the specificity generally decreases with probe length. The sensitivity decreases on the possibility of the formation of stable secondary self-structures by the probe and also by the target limiting the interaction for the duplex formation.

The specificity of oligonucleotide probes remains the main factor of the probes quality measure for the majority of users and probe design algorithms [6]. A specific probe mustn't produce significant signal if its complementary target doesn't exist in the sample hybridized. It mustn't cross-hybridize to other targets. To check probe specificity, algorithms such as BLAST [1] or suffix array method are usually used. Low complexity regions such as those containing long homopolymers[1] or a large number of bases G and C, may also affect probe specificity and must therefore be avoided [18].

All probes in a microarray experiment must have similar thermodynamic behaviors to obtain the best hybridization efficiency. They must be as uniform as possible [17] because they will hybridize under the same conditions (salt concentration, temperature, etc.). Probes must have homogeneous structural properties such as probe length, G + C content, melting temperature or binding capacities.

The use of specific probes in oligonucleotide arrays allows us to simultaneously study several thousand known organisms. However, it is also important to design

[1] A homopolymer is a sequence of identical bases, like "AAAAAA"

explorative probes that can detect unknown sequences in environmental samples and anticipate genetic variations [3].

Here, we present a large-scale oligonucleotide probe design method. Our approach allows the design of both known and explorative high quality oligonucleotide probes, for one or a group of nucleic acid sequences. Selected probes are specific, sensitive and isothermal. We introduced an efficient parallelization method to design probes for large groups of nucleic acid sequences. We used a model driven engineering approach to automatically generate source codes, to select oligonucleotide probes on one of the architectures: PC, multiprocessor, cluster or grid computing.

II. Related Works and Limitations

Functional Gene Arrays (FGAs), targeting key genes encoding enzymes involved in metabolic processes, as well as Phylogenetic Oligonucleotide Arrays (POAs), targeting the SSU rRNA genes, are known as the main approaches to study the microbial diversity in complex environments [17]. Several oligonucleotide probe design algorithms have been proposed to select probes for phylogenetic or functional arrays.

The PROBE_DESIGN tool of the ARB software package [7] was proposed to design probes with a length of 10 – 100 nucleotides. First, a target group of organisms must be specified. All possible signature sequences are then searched. Finally, selected probes can be matched against a database using the ARB Probe Match software. The Tm and the GC-content of probes are checked. ARB also provides a set of already published probes, each targeting distinct phylogenetic groups. However, this tool doesn't allow selecting explorative probes and it is not well suited for large scale oligonucleotide probe design.

The PRIMROSE [2] program is a Perl application proposed to identify both 16S rRNA probes and PCR primers. PRIMROSE is especially useful for the design of degenerate probes. First, a multiple alignment is produced for a given group of sequences. Every probe is subsequently tested against an input database, to detect potential cross-hybridizations and to verify the coverage of the targeted group of sequences. PRIMROSE doesn't check possible secondary structures, the low complexity or the Tm of the selected probes.

OligoArray (v 2.1) [13] is a Java program that designs specific oligonucleotide probes at the genomic scale. It selects probes for an individual sequence. This program takes a FASTA file of non-redundant sequences as input. This file is used for the design and to check the specificity of probes. OligoArray uses the Blast program and a thermodynamic approach to check the specificity of probes. It uses OligoArrayAux [2] to predict stable secondary structures. OligoArray only allows the design of oligonucleotides for use on microarrays for gene expression profiling.

ROSO [12] is a C program used to design oligonucleotide probes, at the genomic scale, using Blast program to check the specificity of each probe. It selects probes for an input sequence according to their Tm, GC content, stable secondary structures and homopolymers.

ORMA (Oligonucleotide Retrieving for Molecular Applications) [14] is one of the most recent tools proposed for probes design. This program is composed of a set of scripts developed under Matlab. ORMA was first applied to the design of probes targeting 16S rRNA genes. It can also be used on any set of highly correlated sequences. ORMA uses Blast program to check the oligonucleotides specificity, but it doesn't check homopolymers or secondary structures of the oligonucleotides selected.

All of these programs allow designing only known probes. A few tools such as PhylArray [5] [10] were designed with the possibility of designing explorative probes. This algorithm selects probes for a group of SSU rRNA sequences to globally monitor known and unknown bacterial communities. PhylArray uses a multiple sequence alignment to construct a degenerate consensus sequence of a group of sequences. All possible oligonucleotides are then generated from this consensus sequence. The oligonucleotides obtained are finally checked for cross-hybridization using Blast program. For large groups of sequences, the probes selection can take several days. The low complexity, the GC content and the secondary structures are not checked by PhylArray.

Among these probes design algorithms, we can distinguish approaches used to select probes for a single input sequence and approaches used to select probes for a group of sequences. The probes selection for a single sequence can often take only few minutes for a single design. However the probes design for a large group of sequences often requires a considerable computation time and can take up to few days for only one design. These approaches make an intensive use of time consuming bioinformatics algorithms such as a BLAST [1] used against a large database of sequences. These approaches require important computing resources especially when dealing with complex environments [19]. For this kind of application, distributed architecture like clusters or Computing Grids [15] provide an efficient approach to meet the continuously evolving computational needs of bioinformatics [4] [9].

Oligonucleotide microarrays, with the opportunity to survey both known and unknown microorganisms through explorative probe design, are one of the most powerful approaches for a better understanding of microbial community functioning. However, the most proposed approaches for probe design don't select explorative probes. These approaches allow studying only known microorganisms with available sequences in public databases. However, the vast majority of microbial species is still non-described and is not represented by sequences in public databases.

Even in PhylArray which was designed to select exploratory probes, the quality of these probes is not tested and several important criteria for the selection of efficient probes are not taken into account (secondary structures, homopolymers, GC content, etc.). Consequently, it is very important to improve this innovative concept of explorative probes and systematically integrate it into probe design methods.

[2] http://mfold.rna.albany.edu/?q=DINAMelt/OligoArrayAux

In this work, we present a new approach that allows selecting oligonucleotide probes for a specific nucleic acid sequence or a group of nucleic acid sequences. All probes selected are sensitive, specific and isothermal. Our method selects and improves the quality of explorative probes. We proposed an efficient parallelization method and a model driven engineering approach to reduce the complexity and the computational time of this probe selection software. We automatically generate source code to select probes on PC, multiprocessor, cluster or grid computing.

III. MATERIAL AND METHODS

A. *Implementation*

Our method was implemented in a program called «MetaExploArrays». MetaExploArrays is written in C++ and was developed under Linux CentOS 5.4. It uses two other programs: Blast [1] and UNAFold [8]. The user must first specify the desired architecture (PC, multiprocessor, cluster or grid computing), the number of processors to use and the size of oligonucleotide probes. MetaExploArrays generates the source codes needed depending on the chosen architecture, the number of processors and other input parameters such as the probe size. Source codes are then compiled and probes selection is running. This meta-programming was achieved with a Model Driven Engineering approach (Figure 1).

Two kinds of input file are accepted by the program: a FASTA file containing a specific nucleic acid sequence or an alignment file, in Clustal format, containing the result of multiple alignment of a group of nucleic acid sequences. For similarity search, the user can provide, as input, a file containing a database of nucleic acid sequences in FASTA format. It can also use one of the predefined databases, available in MetaExploArrays. These predefined databases are:

- A high-quality formatted database composed of about 70,000 prokaryotic SSU rRNA sequences representing 2072 prokaryotic genera. This database was created as described in [5]. It can be used to design probes for Phylogenetic Oligonucleotide Arrays.
- A wide formatted database, in FASTA format, dedicated to microbial communities (all high quality coding DNA sequences (CDSs) from Prokaryotes, Fungi and Environmental taxonomic divisions of the EMBL databank). This database can be used to design probes for Functional Gene Arrays.

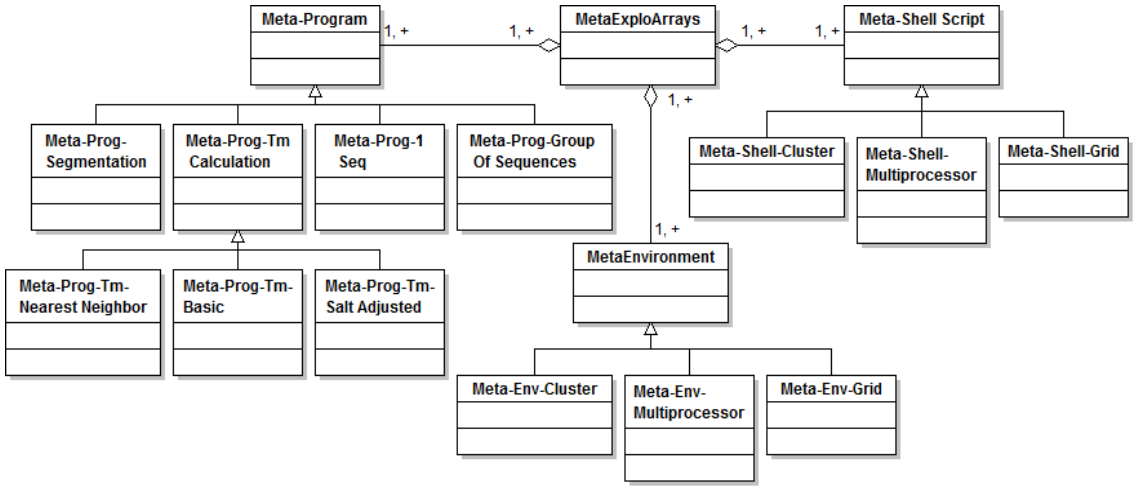

Figure 1. The UML Meta-model of the proposed algorithm

B. *Algorithm*

Before selecting oligonucleotide probes, the user must set the value of the following parameters:

- The length of the desired oligonucleotides (l),
- The maximum number of cross-hybridization authorized to keep an oligonucleotide (C),
- The specificity threshold (S) to consider a cross-hybridization,
- The maximum Tm (T) and the minimum Tm (t) of oligonucleotides,
- The method used to calculate the melting temperature Tm of the oligonucleotides. Our algorithm allows using one of the following methods: Basic (1), Salt Adjusted (2) or Nearest Neighbor (3). MetaExploArrays automatically generates the source code of the chosen method.

$$Tm = 64.9 + 41 * (yG + zC – 16.4) / (wA + xT + yG + zC) \quad (1)$$

$$Tm = 79.8 + 18.5 * \log_{10}([Na^+]) + 58.4 * (yG + zC) / (wA + xT + yG + zC) + 11.8 * (yG + zC)^2 / (wA + xT + yG + zC)^2 – 820 / (wA + xT + yG + zC) \quad (2)$$

$$Tm = (\Delta H - 3.4\ kcal) / (\Delta S + R * \ln(C_T) + 16.6 * \log_{10}([Na^+]) – 273.15 \quad (3)$$

Where w,x,y,z are respectively the number of the bases A,T,G,C in the oligonucleotide; ΔH: the sum of nearest neighbor enthalpy changes; ΔS: the sum of nearest neighbor entropy changes; R: the gas constant 1.987 cal. K-1.mol-1; C_T: the molar concentration of oligonucleotide probe; [Na+]: Salt concentration.

1) Oligonucleotide probe design for one nucleic acid sequence:

Several steps are needed to select probes for a specific nucleic acid sequence.

First, the input sequence is read, to extract oligonucleotides, by using a moving window length equal to the length of the oligonucleotide. The sequence window is moved by 1 nucleotide, along the input sequence, to get all potential oligonucleotides.

Then, the obtained oligonucleotides are examined to eliminate the prohibited ones. We keep only the oligonucleotides that meet the following criteria:

- the percentage of G+C is between 40 and 60 percent of the oligonucleotide length.
- the oligonucleotide doesn't contain a homopolymer longer than 5 nucleotides.
- t <= Tm <= T.

Once an oligonucleotide meets these criteria, it is tested for the absence of secondary structures. The UNAFold program [8] is used to compute the minimum free energies of all possible secondary structures. We use a Na+ concentration of 0.5 M and a temperature equal to the mean of the Tm range set by the user (temperature = (t +T) / 2). If, at this temperature, an oligonucleotide contains a secondary structure with a negative ΔG, it will be rejected.

The last step is the test of specificity. The Blast program [1] is used to check the specificity of the potential probe, against the input database chosen by the user. The Blast program is running with the following parameters: word size W=7, low-complexity Filtering F=false, Expectation value E=1000, one-line descriptions v=5, number of reported alignments b=3000.

The output result file of Blast is then parsed to calculate, for each oligonucleotide, the number of cross-hybridizations with a specificity threshold > S (S is the specificity threshold to consider a cross-hybridization). If this number is less than or equal to the maximum number of cross-hybridization set by the user, the oligonucleotide will be saved in the result file.

2) Oligonucleotide probe design for a group of nucleic acid sequences:

To design probes for a group of nucleic acid sequences that represents a given taxon, MetaExploArrays takes as input, an alignment file in Clustal format. This file is the result of the multiple alignment of the group of sequences for which we will select probes. Several steps are needed (fig. 2).

First, a consensus sequence is created from the alignment file, using the IUPAC code. The aim is not only to obtain a common sequence that entirely represents the taxon targeted, but also to improve alignment and correct possible sequencing errors. Indeed, in each column of the alignment, the number of unknown nucleotides ("N" or "-") is counted. If this number is greater than N_S / 2 (N_S is the total number of sequences aligned), a gap "-" is inserted in the consensus sequence at this position. Else if this number is greater than 0 but less than N_S / 2, all the unknown bases at this position are replaced by the degenerate base calculated from the specific bases of this position.

Next, the algorithm reads the consensus sequence to find all possible subsequences that don't contain gaps ("-"), incrementing a window of length l (l is the length of oligonucleotides) along the consensus sequence. For each subsequence i found at the position pi, MetaExploArrays extracts all possible known oligonucleotides from the sequences used to do alignment, at the position pi. Each obtained oligonucleotide is examined to keep only the sensitive and isothermal probes (Tm, GC-content, homopolymers and secondary structures are checked). If only one of the oligonucleotides extracted at the position pi, doesn't meet these tests, the current subsequence is deleted and MetaExploArrays checks the next subsequence. Otherwise, MetaExploArrays moves to the next step.

The third step is the specificity test of the probes using the Blast program. First, the FASTA database chosen by the user is formatted. All the sequences of the targeted taxon are removed. The sequences of each existing taxon are remote and distributed throughout the FASTA database, to improve the detection of cross-hybridizations. After running Blast (with the following parameters: W=7, F=false, e=1000, v=5, b=6000), the result file is parsed to get cross-hybridizations, based on the specificity threshold set by the user. The set and the number of non-redundant cross-hybridizations, of the current degenerated subsequence, are determined by adding the cross-hybridizations of all non-explorative oligonucleotides generated from this subsequence. If the calculated number is greater than the maximum number of cross-hybridization authorized, MetaExploArrays removes this subsequence and checks the next subsequence. Otherwise, it moves to the next step.

The fourth step is the extraction of the potential explorative probes. The degeneracy of the processed subsequence is calculated. If this degeneracy is greater

than “MaxDeg”, the selection of explorative probes is not permitted (MaxDeg is calculated based on the number of sequences of the targeted group. Its value varies from one group to another; it can reach up to 30 million oligonucleotides.). Otherwise, all possible probes are generated from the degenerated subsequence, using IUPAC code. These oligonucleotides are checked for homopolymers, GC-content and Tm. Only the good oligonucleotides are kept, and the prohibited ones are eliminated. To better monitor the quality of the explorative probes, the Blast program is used to test the specificity of the maintained oligonucleotides, against a database composed of only the nucleic acid sequences of the targeted taxon. An explorative probe is kept if:

- It has a cross-hybridization with only one mismatch, to at least one sequence of the targeted taxon
- And it isn’t a known probe: doesn't perfectly cross-hybridize (0 mismatches) with one of the sequences of the targeted taxon.

The explorative probes kept, are examined to check if they contain a secondary structure with a negative free energy. Then the specificity of these probes will be also checked: an explorative probe must be highly specific and mustn’t cross-hybridize with other taxa.

Finally, the specificity of the maintained explorative oligonucleotides, is checked against the formatted database created in the step 2. If an explorative probe cross-hybridize with at least one other taxon, it’s removed, else it’s kept. The final set of explorative probes that meet all criteria, is saved with the non explorative probes already obtained in the previous step (step 3).

The user result consists of a file containing all the valid subsequences of the targeted taxon, with explorative and non-explorative probes.

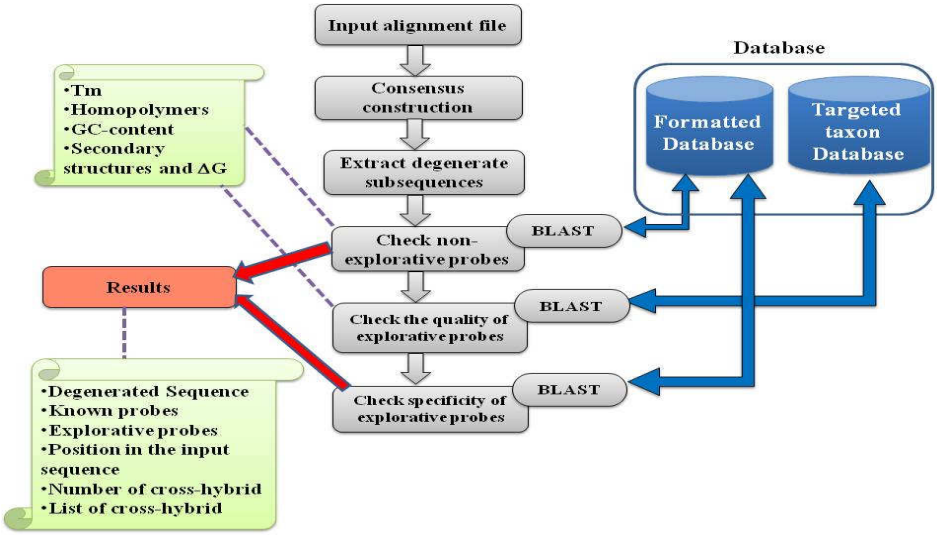

Figure 2. Summary of algorithm steps of selecting probes for a group of nucleic acid sequences

C. *Parallelization method*

The experimentation shows that the design of oligonucleotide probes for an individual nucleic acid sequence, with MetaExploArrays, doesn't require a long computation time. So, MetaExploArrays runs this kind of design only on a PC. However, selecting probes for a group of nucleic acid sequences requires a more important computation time. MetaExploArrays allows running this kind of design on a multiprocessor, a cluster or a grid computing. It contains a program generator that automatically writes the source code needed to parallelize and launch a probe design on the desired architecture. MetaExploArrays can also run a probe design for several groups of sequences simultaneously. So, the parallelization is done on two levels: intra- and inter- design.

First, the user must choose the architecture that he wants to use (a Multiprocessor, a Cluster or a Grid computing) and the number of processors to use on this architecture. Depending on this choice, MetaExploArrays generates 2 kinds of source codes:

- C++ programs that will be used to parallelize and achieve the selection of oligonucleotide probes for the input groups of sequences.
- Shell scripts to run the C++ programs on the desired architecture, using the specified processors number.

The choice of the architecture and the number of processor is always done by the user. However, MetaExploArrays helps the user to choose the best architecture available depending on the number of designs, on their sizes and complexity and also on the number of processors available. For example, if the user decided to use a multiprocessor with only two processors, for a probes design that requires several days of computation (the computation time required is estimated

by MetaExploArrays), MetaExploArrays recommends using more processors on a multiprocessor or a cluster, or using a grid computing if the user doesn't have enough processors available. In this case, MetaExploArrays indicates also the number of processors to use, to run this design on a grid with jobs of about 12 hours.

In MetaExploArrays, each group of sequences is first treated separately. The consensus sequence, constructed from the alignment file of each group of sequences, is read to extract all possible subsequences that don't contain gaps ("-"). The degeneracy of each subsequence is calculated. If this degeneracy is less than "MaxDeg", the subsequence is saved. A weight value is calculated for each saved subsequence based on its degeneracy, the estimated computation time required to process it and the value of "MaxDeg".

Once this step is performed for all targeted groups, all valid subsequences saved are collected and put in the same file. This file is then cut into "N_{Proc}" subfiles (N_{Proc} is the number of processors set by the user) depending on the weight value of each subsequence and the sum of all the weight values. The subfiles created will have almost the same weight (fig. 3). The mixture of all subsequences regardless of the targeted group to which they belong, allows a better load balancing between processors. It increases the probability to create sub-files with more balanced computation times.

The next step consists in selecting explorative and non-explorative probes from each subfile, using the procedure described in III.B.2). Finally, the result files obtained are parsed to regroup and save the probes of each input group of sequences.

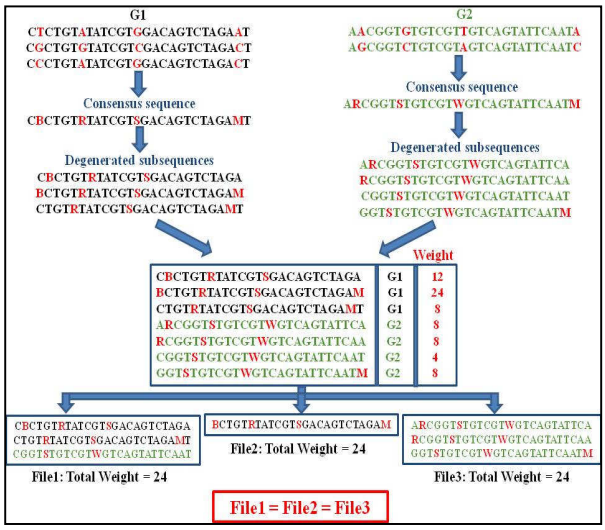

Figure 3. Example of parallelization to design probes for 2 groups of nucleic acid sequences, using 3 processors.

IV. RESULTS

MetaExploArrays allows selecting probes on different architectures: PC, Multiprocessor, Cluster or Grid computing. It helps users to choose the right architecture depending on the number of targeted groups on their sizes and complexity. MetaExploArrays contains a program generator, used to automatically write the source codes needed to select probes on the selected architecture. This meta-programming method is based on a Model Driving Engineering approach that reduces the complexity of our program. This approach was also used to improve the step of generating all possible oligonucleotides from degenerated consensus sequence, using IUPAC code. This step is based on a stack to absorb as much as possible the exponential nature of the problem. The stack depth and the number of nested loops depend on the length of the degenerated sequence to process. The use of meta-programming has considerably minimized the number of writing in memory, and reduced the amount of RAM used at this step. This amount is precisely limited to the probe size, plus a number of loop indexes. To check the gain obtained using this technique, we make a comparison between MetaExploArrays and our previous software PhylArray [10] (Table 1). This comparison is made to see the difference between our method and that used by PhylArray to generate all oligonucleotides from a consensus sequence. The second example illustrated in the table 1 processes a consensus sequence with a length equal to 64 bases. The results of this example show that our method is much more efficient than PhylArray. It significantly reduces the computation time and the amount of RAM consumed. Indeed, MetaExploArrays takes less than 2 minutes and only 11 Ko of RAM to process this example, while PhylArray takes more than 21 minutes and consumes a very large amount of RAM that exceeds 13 Go.

Our algorithm takes into account an important number of criteria to select efficient oligonucleotide probes, on different architectures. Table 2 illustrates the difference between MetaExploArrays and 3 popular probe design software.

MetaExploArrays can be applied to both, individual nucleic acid sequence and a group of nucleic acid sequences. To test our program, we first used it to select oligonucleotides for one sequence of "*Eubacterium eligens*" organism. Results are shown in table3. The experimentations show that selecting probes for 1 specific sequence doesn't require a long computation time. However, the processing of large groups of sequences can take a considerable computation time. To remedy this problem, we proposed an efficient parallelization method based on a parallelization intra-and inter design. The tests illustrated in table 4 and table 5, show that we have considerably reduced the computation time required to select probes, when running MetaExploArrays on a multiprocessor or a cluster. However, some jobs can be very long. These jobs consist in simultaneously selecting probes for several large groups of sequences. Dealing with this kind of design, MetaExploArrays also allows selecting probes on a grid computing. Source codes are automatically generated to submit and monitor jobs over the grid.

TABLE I. COMPARAISON OF THE METHOD OF OLIGONUCLEOTIDES GENERATION FROM A CONSENSUS SEQUENCE, OF METAEXPLOARRAYS AND PHYLARRAY

Consensus Sequence	Probe length	Nb Sub-sequences	Nb probes	Time		RAM	
				PhylArray	*MetaExploArrays*	*PhylArray*	*MetaExploArrays*
ANTCRSGNBBCNANKTANNCBGATKBCNGC	25	6	21676032	40s	3s (12x)	0.9 Go	11 Ko
NTKCNRARANKSTNCNANTCNNTBANGSNBRANTKCNRARNNCSTNGNANACNNTBANGSTBNC	25	40	838860800	21m28s	1m43s (12x)	13.2 Go	11 Ko

TABLE II. COMPARAISON OF METAEXPLOARRAYS AND THREE POPULAR SOFTWARE

Software	Criteria							Parallelization
	Specificity	*Tm*	*GC content*	*Low-complexity*	*Secondary structure*	*Degenerate probes*	*Explorative probes*	
PhylArray[10]	Blast	Basic	NO	NO	NO	YES	YES	Cluster
OligoArray[13]	Blast and thermodynamic	Nearest Neighbor	YES	YES	YES	NO	NO	Multiprocessor
ORMA[14]	Blast	Salt Adjusted	NO	YES	NO	YES	NO	NO
MetaExploArrays	Blast	-Basic -Salt Adjusted -Nearest neighbor	YES	YES	YES	YES	YES	-Multiprecessor, -Cluster -Grid – IDM approach

TABLE III. SELECT PROBES FOR SOME SEQUENCES OF "*EUBACTERIUM ELIGENS*" ORGANISM

Length of sequence	Parameters					Time
	probe length	*Tm Calculation*	*Tm*	*Specificity*	*Max Cross-hybridization*	
657	25	Salt Adjusted	35 - 70	0.92	10	4m22s
447	25	Salt Adjusted	35 - 70	0.92	10	3m01
366	25	Salt Adjusted	35 - 70	0.92	10	1m49

TABLE IV. SELECT PROBES FOR THE GENUS *RHODOVIBRIO*, USING A CLUSTER

Numbre of Processors	Parameters					Time
	probe length	*Tm Calculation*	*Tm*	*Specificity*	*Max Cross-hybridization*	
2	25	Nearest Neighbor	35 - 70	0.92	20	42m21s
5	25	Nearest Neighbor	35 - 70	0.92	20	17m54
10	25	Nearest Neighbor	35 - 70	0.92	20	9m21
20	25	Nearest Neighbor	35 - 70	0.92	20	4m55s

TABLE V. SIMULTANEOUSLY SELECT PROBES FOR 20 PROKARYOTIC GROUPS OF SEQUENCES, USING A CLUSTER

Targeted groups	Parameters					Time (hours)	
	Probe length	*Tm Calculation*	*Tm*	*Specificity*	*Max Cross-hybridization*	*1 core*	*100 cores*
CRONOBACTER; FLAVOBACTERIUM; MICROCOCCUS; STREPTOCOCCUS; TREPONEMA; CORYNEBACTERIUM; PSEUDOALTEROMONAS; PANTOEA; NOCARDIA; BRADYRHIZOBIUM; SPHINGOMONAS; HALOMONAS; SHEWANELLA; BURKHOLDERIA; RHODOVIBRIO; HELICOBACTER; NOCARDIOIDES; MARINOBACTER; VIBRIO; GEOTHERMOBACTERIUM	25	Salt Adjusted	35 - 70	0.92	150	232,5 hours (9,5 days)	6,5 hours

V. CONCLUSION

In summary, we present a novel oligonucleotide probe design software called MetaExploArrays that allows probe design for an individual nucleic acid sequence or a group of sequences. MetaExploArrays takes into account all significant known criteria to select sensitive, specific and isothermal probes. For each potential probe, MetaExploArrays checks the specificity, the number of cross-hybridization, the Tm, the GC content, the possible secondary structures formed and the existence of homopolymers.

MetaExploArrays allows selecting both known and explorative probes to identify new organisms or variant genes that are not yet discovered. Compared to our previous software PhylArray [10], a greater number of explorative probes are examined and only those of very high-quality are selected. The selected explorative probes of a given organism are highly specific and must not belong to any other organism.

MetaExploArrays is well adapted to multiple microarray applications. It can be used to design POA (Phylogenetic Oligonucleotide Arrays), FGAs (Functional Gene Arrays) or WGAs (whole genome arrays).

To deal with the considerable computation time required when selecting probes for a large group of nucleic acid sequences, we proposed an efficient parallelization method. The consensus sequence of an input group of sequences is fragmented into short subsequences of length equal to the length of the oligonucleotide probes desired. The processing of the obtained subsequences is equally shared across the available processors. When several groups of sequences are simultaneously processed, the parallelization is done on two levels: intra- and inter-design. All input groups are first combined and mixed, the design of probes is then done simultaneously, and the results are finally parsed and separated.

MetaExploArrays can be used to select probes on one of the following architectures: a PC, a multiprocessor, a cluster or a computing grid. It helps the users to choose the best architecture available. When using a grid computing, MetaExploArrays estimates the number of jobs required and creates, as far as possible, jobs of 12 hours.

Our software contains a program generator that automatically writes source codes. This meta-programming is based on a model driven engineering approach that considerably reduces the complexity of our program and significantly increases performance up to 12x when compared to our previous software [10].

ACKNOWLEDGMENT

We thank the Auvergne Regional Council for the funding of F. Jaziri scholarships.

REFERENCES

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J., "Basic local alignment search tool.", J. Mol. Biol., 215, 1990, 403-410.

[2] Ashelford, K.E., Weightman, A.J. and Fry, J.C., "PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database", Nucleic Acids Research, 30, 2002, 3481–3489.

[3] Dugat-Bony, E., Missaoui, M., Peyretaillade, E., Biderre-Petit, C., Bouzid, O., et al., "HiSpOD: probe design for functional DNA microarrays", Bioinformatics, 2011, 27: 641–648.

[4] Foster I., Kesselman C., "The Grid 2: Blueprint for a New Computing Infrastructure", Morgan Kaufmann, 2004, 748 pp.

[5] Jaziri, F., Missaoui, M., Cipiere, S., Peyret, P., Hill, D.R.C., "Large Scale Parallelization Method of 16S rRNA Probe Design Algorithm on Distributed Architecture: Application to Grid Computing", International Conference on Informatics and Computational Intelligence, 2011, 35-40.

[6] Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D., and Madore, S.J., "Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays.", Nucleic Acids Res, 2000, 28: 4552–4557.

[7] Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, •• et al., "ARB: a software environment for sequence data.", Nucleic Acids Res, 32, 2004, 1363–1371.

[8] Markham, N. R., and Zuker, M., "UNAFold: software for nucleic acid folding and hybridization", In Keith, J. M., editor, Bioinformatics, Volume II. Structure, Function and Applications, number 453 in Methods in Molecular Biology, 2008, chp1 3–31.

[9] Melab, N., Cahon, S., Talbi, E.G., "Grid Computing for parallel bioinspired algorithms", Journal of parallel and Distributed Computing, 66, 2005, 1052-1061.

[10] Militon, C., Rimour, S., Missaoui, M., Biderre, C., Barra, V., Hill, D., Moné, A., Gagne, G., Meier, H., Peyretaillade, E. and Peyret, P., "PhylArray: phylogenetic probe design algorithm for microarray", Bioinformatics, 23, 2007, 2550-2557.

[11] Mora, C., Tittensor, DP., Adl, S., Simpson, AGB., Worm, B., "How Many Species Are There on Earth and in the Ocean?", PLoS Biol 9(8), 2011, e1001127. doi:10.1371/journal.pbio.1001127.

[12] Reymond, N., Charles, H., Duret, L., Calevro, F., Beslon, G., and Fayard, J.M., "ROSO: optimizing oligonucleotide probes for microarrays", Bioinformatics 20, 2004, 271–273.

[13] Rouillard, JM., Herbert, CJ., Zuker, M., "OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach", Nucleic Acids Research, 31, 2003, 3057-3062.

[14] Severgnini, M., Cremonesi, P., Consolandi, C., Caredda, G., De Bellis, G., and Castiglioni, B., "ORMA: a tool for identification of species-specific variations in 16S rRNA gene and oligonucleotides design.", Nucleic Acids Res, 2009, 37: e109.

[15] Talbi, E.-G. and Zomaya, A.Y., "Grid Computing for Bioinformatics and Computational Biology". Wiley Interscience, 2008, 392 pp.

[16] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science, 270, 1995, 467-470.

[17] Wagner, M., Smidt, H., Loy, A., and Zhou, J., "Unravelling Microbial Communities with DNA-Microarrays: Challenges and Future Directions", Microb Ecol, 53, 2007, 498–506.

[18] Wang, X., and Seed, B., "Selection of oligonucleotide probes for protein coding sequences", Bioinformatics, 19, 2003, 796–802.

[19] Zhu, D., Fofanov, Y., Willson, R.C. and Fox, G.E., "A parallel computing algorithm for 16S rRNA probe design", Journal of Parallel and Distributed Computing, 66, 2006, 1546-1551.

Détermination de sondes oligonucléotidiques pour outils moléculaires à haut-débit : application pour le développement d'une nouvelle approche de capture de gènes pour l'écologie microbienne

Résumé :

Les microorganismes, par leurs fascinantes capacités d'adaptation liées à l'extraordinaire diversité de leurs capacités métaboliques, jouent un rôle fondamental dans les tous les processus biologiques. Ils interviennent notamment au niveau des changements globaux, comme le réchauffement climatique, en partie occasionné par les émissions croissantes de méthane dans l'atmosphère, mais également par les pollutions résultant de la dispersion de molécules comme les Hydrocarbures Aromatiques Polycycliques. Ainsi, les communautés microbiennes vont participer à réduire ou à augmenter les effets délétères de l'anthropisation des écosystèmes. La régulation des changements globaux passe donc par une meilleure connaissance de ces communautés qui doivent être explorées dans leur globalité au sein des environnements. Néanmoins en raison de leur forte complexité, une telle exploration n'est possible qu'en utilisant des outils d'analyse haut-débit. Cependant, l'emploi d'outils moléculaires à haut-débit comme les biopuces à ADN passe par la détermination de sondes combinant à la fois une forte sensibilité, une très bonne spécificité et un caractère exploratoire. Pour concevoir de telles sondes un nouveau logiciel KASpOD a donc été développé. De même, en utilisant des sondes présentant les mêmes caractéristiques, le développement d'une nouvelle approche innovante en écologie microbienne de capture de gènes en solution été entrepris. Cette nouvelle méthode d'enrichissement de gènes d'intérêt couplée à du séquençage haut-débit a été appliquée pour l'exploration des communautés méthanogènes du lac Pavin. Les résultats obtenus montrent la pertinence de l'approche qui assure une meilleure évaluation de diversité de l'écosystème avec notamment l'identification de populations appartenant à la biosphère rare. L'autre ajout majeur de cette approche est qu'elle autorise l'identification de grandes régions d'ADN génomique exploitable pour caractériser de nouveaux gènes ou de nouveaux processus adaptatifs.

Mots clés : *changement global, métagénomique, détermination de sondes, capture de gènes*

Selection of oligonucleotide probes for high-throughput molecular tools : application for a new gene capture method's development for microbial ecology

Abstract :

Microorganisms play a crucial role in all biological processes related to their huge metabolic potentialities. They are involved in global changes such as global warming partially caused by the growing methane emissions in the atmosphere, but also by the release of pollutants such as Polycyclic Aromatic Hydrocarbons. Thus, microbial communities will contribute to reduce or increase the negative effects of human impacts on ecosystems. The regulation of global changes needs a better knowledge of the microbial communities involved in complex environments functioning. Nevertheless, a complete exploration of such environments requires the use of high-throughput tools, due to the extraordinary diversity of microorganisms within the ecosystems. The use of DNA microarrays requires a probe design step allowing the selection of highly sensitive, specific and explorative oligonucleotides. For this purpose, we have developed KASpOD, a new software, allowing the generation of efficient probes dedicated to environmental applications. Using high quality probe sets, an innovative in solution-based gene capture method combined with Next Generation Sequencing, was developed and applied for the exploration of the methanogen communities in lake Pavin, Results showed the relevance of this approach that allows a better evaluation of the methanogen diversity with an efficient detection of populations belonging to the rare biosphere. The other main advantage of this approach is the identification of large regions of genomic DNA, useful for the characterization of new genes or adaptive processes.

Keywords: *global change, metagenomics, probe design, gene capture*