Réponses adaptatives des microorganismes eucaryotes du sol aux pollutions métalliques

Frédéric Lehembre

▶ **To cite this version:**

Frédéric Lehembre. Réponses adaptatives des microorganismes eucaryotes du sol aux pollutions métalliques. Sciences agricoles. Université Claude Bernard - Lyon I, 2009. Français. NNT : 2009LYO10348 . tel-00482109v2

HAL Id: tel-00482109

https://theses.hal.science/tel-00482109v2

Submitted on 29 Feb 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre 348-2009 Année 2009

THESE DE L'UNIVERSITE DE LYON

Délivrée par

L'UNIVERSITE CLAUDE BERNARD LYON 1

ECOLE DOCTORALE E2M2

DIPLOME DE DOCTORAT

(arrêté du 7 août 2006)

soutenue publiquement le 14 décembre 2009

par

Frédéric LEHEMBRE

Réponses adaptatives des microorganismes eucaryotes du sol aux pollutions métalliques

Directeur de thèse : Dr Roland MARMEISSE
Co-encadrante: Dr Laurence FRAISSINET-TACHET

JURY :

Mr Roberto GEREMIA, DR CNRS, HDR, Université de Grenoble	Rapporteur
Mr Telesphore SIME-NGANDO, DR CNRS, HDR, Univ. Clermont Ferrand	Rapporteur
Mr Max MERGEAY, Professeur Honoraire, Univ. Libre de Bruxelles	Rapporteur
Mr Roland MARMEISSE, CR CNRS, HDR, Université Lyon1	Examinateur
Mr Bruno COMBOURIEU, Professeur, Université Lyon1	Président

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur L. Collet**
Vice-président du Conseil Scientifique	M. le Professeur J-F. Mornex
Vice-président du Conseil d'Administration	M. le Professeur G. Annat
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. Simon
Secrétaire Général	M. G. Gay

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. Etienne
Faculté de Médecine Lyon Sud – Charles Mérieux	Directeur : M. le Professeur F-N. Gilly
UFR d'Odontologie	Directeur : M. le Professeur D. Bourgeois
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. Locher
Institut des Sciences et Techniques de Réadaptation	Directeur : M. le Professeur Y. Matillon
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. Farge

COMPOSANTES SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. Le Professeur F. Gieres
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. Collignon
Observatoire de Lyon	Directeur : M. B. Guiderdoni
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. Lieto
Institut Universitaire de Technologie A	Directeur : M. le Professeur C. Coulet
Institut Universitaire de Technologie B	Directeur : M. le Professeur R. Lamartine
Institut de Science Financière et d'Assurance	Directeur : M. le Professeur J-C. Augros
Institut Universitaire de Formation des Maîtres	Directeur : M R. Bernard

Remerciements

Ce travail a été réalisé dans le Laboratoire d'Ecologie Microbienne (UMR 5557 CNRS-Université Lyon 1) dirigé par René Bally, au sein de l'équipe Symbiose mycorhizienne. Il a été dirigé par Roland Marmeisse et co-encadré par Laurence Fraissinet-Tachet entre novembre 2006 et novembre 2009.

Je tiens à remercier toutes les personnes qui m'ont aidé, supporté et soutenu pendant toute ma thèse. Mes remerciements vont en priorité à Roland Marmeisse et Laurence Fraissinet-Tachet pour m'avoir accueilli au sein de leur équipe et pour m'avoir transmis tant de connaissances théoriques et techniques qui ont inspiré ce travail. Je vous remercie également pour votre soutien et votre disponibilité tout au long de ces années.

Mes remerciements vont également à Gilles Gay, Delphine Melayah et Patricia Luis pour leurs remarques et leurs idées lors de la préparation de mes présentations orales. Je tiens à remercier les membres de mon comité de pilotage, Damien Blaudez, Jacques Bourguignon et Marc Lemaire pour leurs précieux conseils et leurs soutiens. Merci à Jan Colpaert pour son accueil et sa disponibilité lors des campagnes d'échantillonnages en Belgique.

Je remercie bien sûr toute les personnes qui ont contribué à la réalisation de ce projet: Marie-Christine Verner, Elise David, Sandrine Perrotto, Sophie Cros, Laurie Félix, Benjamin Doix, Charlotte Guillarme et Jessica Baude.

Merci à toutes les personnes de l'équipe 2 : Jeanne, Concetta, Coralie, Laurent, Chrisse pour l'ambiance chaleureuse qui régnait au Batiment Lwoff. Un grand merci aux étudiants du laboratoire d'écologie microbienne qui pour la plupart sont devenus des amis proches. Un grand merci à vous tous : Olivier, Vincent, Karima, Edwige, Marie-Lara, Jean, David, Florient, Amel, Lamiae, Béné, Clo, Tony, Steph, Arnaud et tous les autres.

Enfin un grand merci à ma famille et à mes proches pour m'avoir fait confiance et m'avoir soutenu toutes ces années.

Sommaire

Liste des abréviations

AA	Acides aminés
ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ADNr	ADN ribosomique
ARN	Acide ribonucléique
ARNm	ARN messager
ARNr	ARN ribosomique
ARNt	ARN de transfert
BSA	Albumine sérique bovine
DNase	Désoxyribonucléase
dNTP	2'-déoxynucléoside 5'-trisphosphate
EDTA	Acide Ethylènediamine tétraacétique
EST	Expressed Sequence Tag
kb	kilobase
LB	Luria-Bertani
min	minute
pb	paire de base
PCR	Réaction de Polymérisation en Chaine
rpm	rotation par minute
RT	Réverse Transcription
RT-PCR	Réverse Transcription de l'ARN suivie d'une PCR
SEVAG	Chloroforme/Alcool isoamylique (24:1)
Tris	Tris (hydroxyméthyl)-aminoéthane

Résumé

Les sols pollués par des métaux lourds sont colonisés par des communautés de microorganismes qui ont développé différentes adaptations leur permettant de résister à ces contaminants. Afin d'analyser au niveau moléculaire la diversité de ces adaptations, une approche expérimentale innovante basée sur l'étude du métatranscriptome eucaryote des sols a été utilisée pour comparer les fonctions exprimées au sein d'une communauté de microorganismes eucaryotes colonisant un sol contaminé, anciennement contaminé et non contaminé par des métaux lourds. Les banques d'ADNc eucaryotes construites à partir des ARNm extraits directement de ces sols ont été criblées par séquençage aléatoire de leurs inserts et par complémentation fonctionnelle de mutants de levures sensibles au cadmium. Cette étude a permis d'identifier de nouveaux gènes et de nouveaux mécanismes impliqués dans la résistance au cadmium ainsi qu'un nombre important de nouvelles protéines hypothétiques. Ceci démontre l'intérêt appliqué de cette approche pour la recherche de nouveaux biocatalyseurs et molécules bioactives.

En parallèle, la diversité microbienne eucaryote a été révélée et comparée entre les sols par le clonage-séquençage du gène codant la petite sous-unité ribosomique 18S. Cette étude a mis en évidence une diversité inattendue de microorganismes eucaryotes dans ces sols et une analyse phylogénétique a permis de découvrir un nouveau clade de protistes (Rhizaria, Cercozoa).

Mots clés: microorganismes eucaryotes, diversité taxinomique, diversité moléculaire, métatranscriptome, métaux lourds.

Summary

Heavy metal-polluted soils are colonised by microbial communities which have developed different adaptations that allow them to resist to these pollutants. The objectives of this study were to reveal at the molecular level the diversity of these adaptations. To this aim, we implemented an innovative approach based upon the analysis of soil eukaryotic metatranscriptome to compare resistance mechanisms expressed by eukaryotic microorganisms living in heavy metal contaminated and control non contaminated or formerly-contaminated soils. Eukaryotic cDNA libraries were prepared using mRNA directly extracted from these soils and screened by either systematic sequencing of their inserts or functional complementation of yeast mutants sensitive to cadmium. This study allowed us to characterise novel genes and mechanisms implicated in Cd resistance as well as numerous novel hypothetical proteins. This study also demonstrates the potential of this experimental approach to look for novel biocatalysts as well as novel bio-active molecules.

In addition, eukaryotic molecular diversity was studied by the cloning-sequencing of the gene encoding the 18S ribosomal RNA. This study revealed an unexpected diversity of eukaryotic diversity in the studied soils and allowed us to discover a novel clade of protists (Rhizaria, Cercozoa).

Key words: eukaryotic microorganisms, taxonomic diversity, molecular diversity, metatranscriptome, heavy metal.

Introduction générale

Les écosystèmes sont de plus en plus souvent soumis à des pressions anthropiques qui perturbent durablement leur fonctionnement. C'est le cas des pollutions d'origine industrielle parmi lesquelles la déposition de métaux lourds dans les sols. L'accumulation de ces composés chimiques non dégradables et potentiellement toxiques perturbe et modifie les équilibres compétitifs entre espèces et conduisent à l'élimination de certaines, à l'implantation de nouvelles et en définitive à éventuellement une diminution (ou voire à une augmentation) de la diversité biologique (Amaral Zettler *et al.*, 2002). Les pollutions par de très fortes quantités de polluant conduisent souvent à une élimination directe de nombreuses espèces sensibles par toxicité aigue. A long terme les écosystèmes subissant une pollution chronique élevée se trouvent donc colonisés par des espèces et des individus dont une des caractéristiques principales qui détermine leur présence sur ces sites est leur capacité à résister aux toxiques (Blaudez *et al.*, 2000b ; Colpaert *et al.*, 2000). La résistance aux métaux peut résulter d'un phénomène de plasticité phénotypique par modification du profil d'expression de gènes ou bien de mutations présentes potentiellement au sein de populations d'organismes colonisant des milieux non contaminés. Des études de profils d'expressions de gènes ont permis de refléter la modification de l'expression de gènes d'organismes modèles en condition de stress métallique (Vido *et al.*, 2001 ; Bellion *et al.*, 2006 ;Weber *et al.*, 2006).

Le but de mon projet de thèse a été d'étudier les réponses adaptatives des microorganismes eucaryotes du sol soumis à des pollutions chroniques par des métaux lourds. Ces microorganismes jouent un rôle essentiel dans la vie des sols et contribuent à leur fertilité. Les champignons sont par exemple les principaux décomposeurs primaires de la matière organique et ils jouent aussi un rôle important en interagissant de façon bénéfique (symbiontes mycorhiziens) ou négative (pathogènes) avec les macroorganismes végétaux ou animaux. Les « protistes » quant à eux sont sans doute parmi les principaux régulateurs des populations de microorganismes. A l'image de beaucoup de bactéries, de très nombreux microorganismes eucaryotes ne sont pas cultivables en condition de laboratoire et de nombreuses espèces restent à décrire.

Nous nous sommes donc fixé comme objectif de vouloir révéler et analyser *in situ* la diversité des mécanismes mis en œuvre par l'ensemble d'une communauté de microorganismes eucaryotes pour résister aux métaux lourds et donc leur permettant de

coloniser et d'assurer des fonctions essentielles au fonctionnement biologique d'un sol pollué. Ce travail ne se limite pas à la seule étude d'une ou plusieurs espèces modèles ; il s'intéresse à l'ensemble des microorganismes eucaryotes d'un sol, connus ou inconnus, cultivables ou non cultivables. La réalisation de cet objectif a nécessité la mise en œuvre d'une approche expérimentale de métatranscriptomique développée chez les microorganismes procaryotes et qui a récemment permis de révéler la structure et le fonctionnement de communautés bactériennes des sols (Leininger *et al.*, 2006 ; Urich *et al.*, 2008) et les environnements marins (Poretsky *et al.*, 2005 ; Frias-Lopez *et al.*, 2008 ; Gilbert *et al.*, 2008 ; Poretsky *et al.*, 2009). Cette approche a été utilisée au laboratoire pour l'étude d'une communauté de microorganismes eucaryotes colonisant un sol forestier (Bailly *et al.*, 2007).

Dans le cadre de ma thèse cette approche a été utilisée pour comparer trois communautés microbiennes eucaryotes, l'une colonisant un sol fortement contaminé par des métaux lourds, la seconde un sol anciennement contaminé par des métaux lourds et la troisième un sol témoin non contaminé. Pour chacune de ces communautés une banque d'ADNc a été réalisée dans un plasmide permettant l'expression des gènes clonés dans la levure *Saccharomyces cerevisiae*. Chacune de ces banques a été criblée pour des gènes conférant une résistance au stress métallique par complémentation fonctionnelle de différents mutants de levure sensibles aux métaux étudiés. Les banques du sol non contaminé et du sol contaminé aux métaux lourds ont aussi fait l'objet d'un séquençage systématique de leurs inserts. La répartition des gènes dans différentes classes fonctionnelles a permis de révéler le profil d'expression global des différentes communautés. En parallèle, la diversité taxinomique des communautés eucaryotes présentes dans chacun des sols a été étudiée par l'analyse des séquences ribosomiques 18S amplifiées par PCR à partir des ADN et des ARN environnementaux.

Dans un premier temps, il sera présenté un bilan succinct de nos connaissances concernant la biodiversité des organismes eucaryotes dans l'environnement précédé de notions générales de phylogénie nécessaire à leur classification. Un bilan des nouvelles approches de génomiques environnementales et leurs retombées pour l'étude de la diversité taxinomique et fonctionnelle des microorganismes seront ensuite exposés. Pour finir, nous aborderons le thème des pollutions métalliques et leur impact sur les organismes vivants du sol et nous présenterons les connaissances acquises sur des organismes eucaryotes modèles quant aux mécanismes permettant une résistance accrue aux métaux lourds.

Synthèse bibliographique

1. La phylogénie

1.1 La systématique

Il s'agit de la science des classifications. Elle permet d'identifier et de décrire les êtres vivants présents dans la nature présente et passée. Une fois cet inventaire effectué, elle s'attache à les classer de façon à rendre intelligible leur immense diversité. En biologie, la logique la plus pertinente pour classer les espèces est celle de leur parenté évolutive.

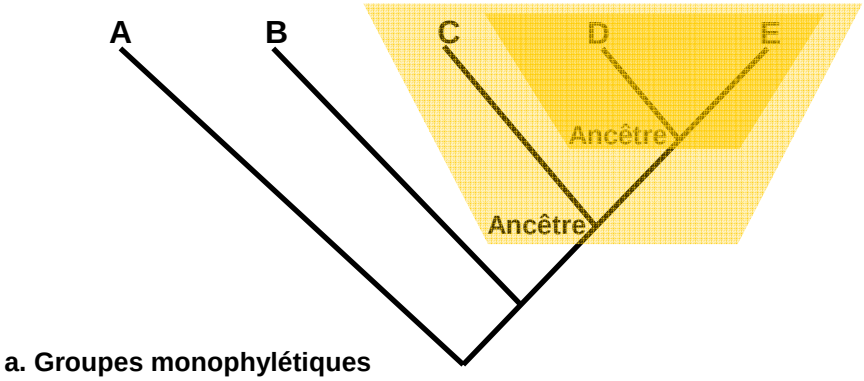
L'importance de la systématique est particulièrement bien illustrée par les sciences de l'environnement. En effet, la liste des espèces et leur abondance respective dans un milieu donné sont d'excellents indicateurs de l'état de santé de celui-ci (Beeby, 2001). Toutes les décisions à prendre en matière de protection des milieux dépendent étroitement d'une bonne connaissance de la biodiversité qui les peuple. De même, la lutte contre les grandes épidémies parasitaires ne saurait progresser sans une bonne connaissance de la systématique des parasites et des espèces vectrices.

1.2 La notion de biodiversité

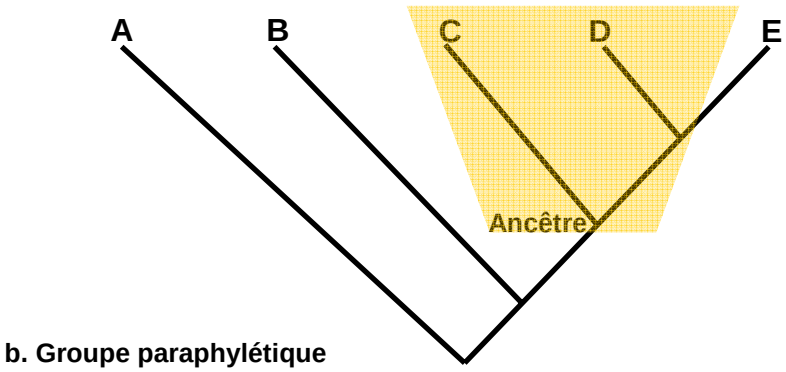
La diversité des êtres vivants, ou biodiversité, peut être perçue de deux manières.

La première résulte d'une approche du milieu naturel. Par exemple, les espèces peuvent être classées suivant le milieu qu'elles occupent ou encore suivant leur capacité à tolérer un stress environnemental. Ces groupes servent à l'écologie et aux sciences de l'environnement, sciences des processus.

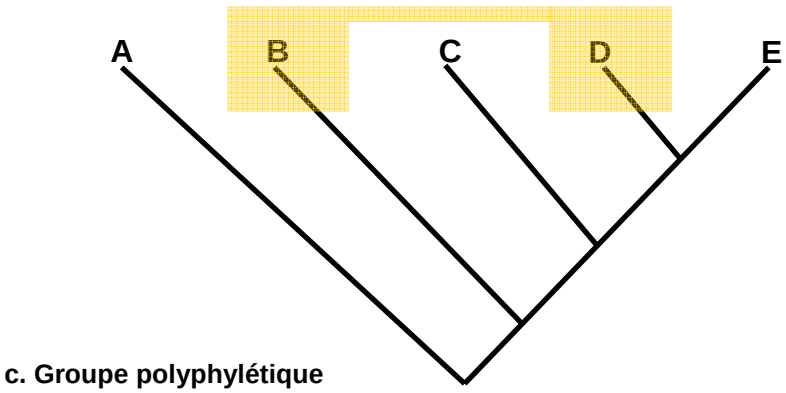
La seconde perception de la biodiversité se réfère à l'histoire pour expliquer sa structure. Cette classification répond à la question, d'où cela vient-il ? En effet, les contraintes d'un même milieu peuvent engendrer l'apparition de structures similaires, convergentes, chez des espèces très éloignées phylogénétiquement. Si les groupes écologiques servent à l'analyse du fonctionnement d'un biotope, on ne peut le comprendre complètement sans intégrer son histoire. Or, cette histoire ne saurait être reconstruite sans l'outil phylogénétique. Pour qui veut expliquer la biodiversité, la dimension historique est incontournable.



a. Groupes monophylétiques



b. Groupe paraphylétique



c. Groupe polyphylétique

Figure 1. a : un groupe monophylétique comprend un ancêtre et tous ses descendants. DE et CDE sont tous deux des groupes monophylétiques. **b :** Le groupe paraphylétique comprend un ancêtre et une partie seulement de ses descendants. L'un des membres du groupe (ici D) est plus proche d'un taxon hors du groupe (E) qu'il ne l'est de ces collatéraux dans le groupe C. **c :** Le groupe polyphylétique comprend des membres (ici BD) sans ancêtre commun dans le groupe (d'après Lecointre et Le-Guyader, 2001).

1.3 Les méthodes de classifications actuelles

Aujourd'hui, il existe des algorithmes de construction d'arbres partant de séquences d'ADN qui exploitent spécifiquement la similitude globale, ou reposent sur une approche probabiliste. Les premiers, les algorithmes phénétiques sont efficaces tant que la dissimilarité reste proportionnelle au degré d'apparentement (ce qui n'est pas toujours le cas). Les seconds, les algorithmes probabilistes, utilisent des probabilités de changements d'un nucléotide en un autre ou d'un acide aminé en un autre pour calculer la vraisemblance des données pour un arbre. Brièvement dit, trois grandes familles d'algorithmes sont donc utilisées aujourd'hui : cladistique, phénétique et probabiliste (maximum de vraisemblance). Elles sont toutes pratiquées à l'aide de l'outil informatique

1.3.1 La cladistique

L'analyse cladistique vise à reconstruire la phylogénie d'un taxon par distinction, au sein d'un caractère, de l'état primitif (plésiomorphe) de l'état dérivé (apomorphe) (Hennig 1950). Ceci n'est valable qu'au sein d'un taxon. On appelle taxon tout groupe d'organismes reconnu et nommé par les taxinomistes, sans niveau précis (ce peut être une variété, une espèce, un ordre, ou même un règne). A ce niveau, seuls les états de caractères dérivés partagés (synapomorphies ou innovations du groupe) sont des signes d'apparentement exclusif ; les regroupements sur base d'états dérivés partagés conduisent donc à la création de groupes monophylétiques.

Pour que la classification soit strictement phylogénétique, elle ne doit contenir que des groupes monophylétiques. Un groupe monophylétique comprend un ancêtre hypothétique et l'ensemble de ces descendants (Fig. 1.a).

Les groupes paraphylétiques et polyphylétiques s'opposent aux groupes monophylétiques en cela qu'il leur manque quelque chose. Au premier, il manque certains descendants de l'ancêtre commun et au second, il manque l'ancêtre commun (Fig. 1.b et 1.c).

Fonder des groupes sur des caractéristiques écologiques, adaptatives, liées à un progrès ou à une augmentation de complexité conduit souvent à construire des groupes paraphylétiques.

1.3.2 La phénétique

Les méthodes phénétiques analysent les données d'une autre manière. A l'opposé de la cladistique, elles tentent de quantifier la ressemblance générale entre organismes ; pour cela

elles calculent un indice de similitude globale entre deux taxons, c'est-à-dire une distance pour chaque couple de taxons. Dans le cas des séquences alignées, cette distance est le nombre de nucléotides différents entre deux espèces (ou d'acides aminés différents dans le cas de séquences protéiques), divisé par le nombre de sites examinés. En somme, il s'agit d'un pourcentage de différences entre les séquences de deux espèces. Les distances sont inscrites dans une matrice.

Il existe plusieurs techniques de construction d'un arbre à partir d'une telle matrice. La plus simple est celle de l'UPGMA (*Unweighted Pair Group Method using Averages*) (Kuhn *et al.*, 1978) qui trouve un seul arbre à partir d'une matrice, par agglomération successive des espèces, des plus proches aux plus éloignées. Elle nécessite l'hypothèse que les séquences évoluent à la même vitesse dans toutes les branches de l'arbre, mais ce postulat est rarement vérifié.

Une autre technique, basée sur un algorithme plus élaboré, est celle dite du Neighbor-Joining (Saitou et Nei, 1987). Elle permet d'introduire un critère de minimisation de la longueur totale de l'arbre. Elle construit un seul arbre, mais en ne choisissant pas nécessairement dès le départ d'agglomérer les espèces les plus proches. A chaque étape d'agglomération, elle choisit en revanche d'agglomérer les espèces dont le regroupement va minimiser la longueur de l'arbre.

De nos jours, les méthodes de distances sont surtout utilisées sur des données moléculaires et non plus sur des données morphologiques. Contrairement à une idée fausse assez répandue, les données de séquences ne doivent pas forcément être analysées à l'aide de distances et peuvent parfaitement faire l'objet d'une approche cladistique.

1.3.3 Les autres méthodes

Hormis ces méthodes les plus couramment utilisées, il existe d'autres méthodes comme les méthodes probabilistes qui établissent un modèle constitué d'un ensemble de paramètres dont le réglage formule différentes hypothèses. Ces hypothèses concernent surtout l'évolution des états de caractères et sont exprimées en termes de probabilités. Comme tout arbre implique des changements d'états de caractères le long de ses branches, toutes les probabilités associées aux transformations impliquées par un arbre donné vont se multiplier et fournir une valeur globale de vraisemblance des données associées à cet arbre. Parmi les arbres possibles, l'arbre choisi est celui dont la vraisemblance des données au vu du modèle

est maximale. Cette méthode fonctionne sur les caractères moléculaires, pour lesquels on peut établir des modèles d'évolution des protéines et des acides nucléiques.

1.3.4 Notions associées à ces méthodes

1.3.4.1 L'attraction des branches longues

Toutes les méthodes sont sujettes à l'artefact d'attraction des branches longues. Cet artefact provient des inégalités du taux d'évolution des caractères entre les lignées analysées. Les espèces qui évoluent plus vite que les autres ont leur branche propre plus longue. On a pu montrer théoriquement et expérimentalement qu'au-delà d'un certain écart de vitesse d'évolution entre les espèces, les espèces qui évoluent plus vite ont plus de chance d'avoir des états de caractères communs par hasard que par ascendance commune, et que le nombre de caractères communs ainsi acquis (homoplasie) devenait supérieur aux caractères qui auraient dû les séparer. Par conséquent, elles sont regroupées ensemble dans l'arbre indépendamment de leurs parentés : c'est ce qu'on appelle l'attraction des branches longues.

1.3.4.2 Le *bootstrap*

La robustesse d'un arbre phylogénétique repose sur la valeur de *bootstrap* (Felsenstein, 1985). Il s'agit de mesurer, sans utiliser d'informations extérieures, le degré avec lequel un jeu de données (le plus souvent de séquences) supporte un regroupement. Le *bootstrap* procède à un tirage avec remise des caractères, en l'occurrence des sites de la séquence, jusqu'à ce que le nombre de sites tirés soit identique au nombre de sites du jeu initial de données. Ce processus est répété un grand nombre de fois (en général 1000 fois). On obtient donc un grand nombre d'arbres, qui représentent en fait un grand nombre de regroupements, et dont certains, bien entendu, sont contradictoires. Le résultat est généralement représenté sous forme d'un arbre consensus dans lequel figurent tous les regroupements majoritairement apparus. Chaque regroupement (chaque nœud) possède alors un pourcentage qui indique la proportion d'arbres issus des tirages qui le présentent. C'est la valeur de *bootstrap*, ou la proportion du *bootstrap* du nœud. On peut aussi présenter l'arbre le plus parcimonieux issu d'une analyse cladistique standard et poser sur les clades leur valeur de *bootstrap*.

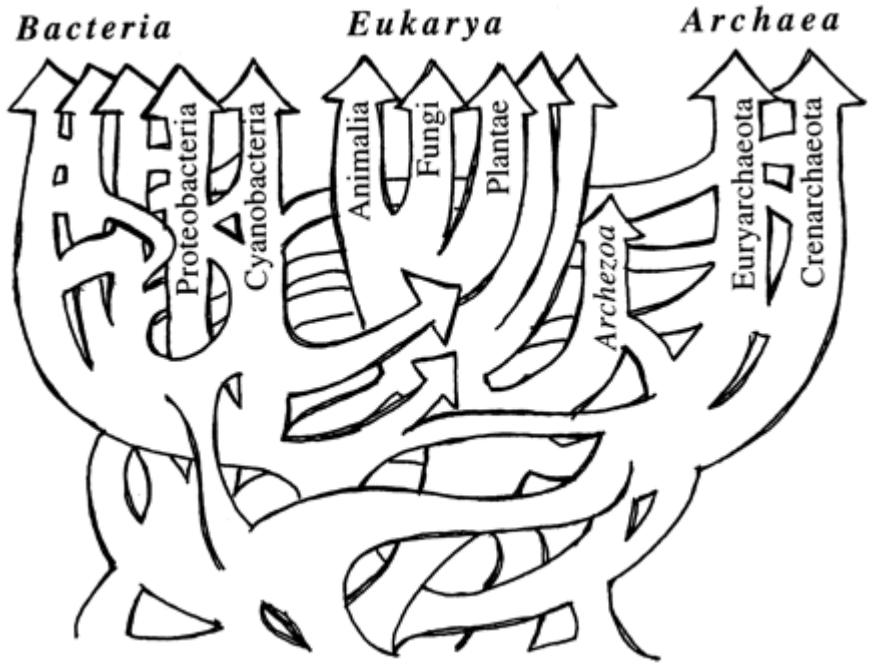


Figure 2: Un arbre réticulé ou en réseau, qui pourrait représenter convenablement l'histoire de la vie. (d'après Doolittle, 1999).

Il faut souligner que la robustesse n'est pas tout. En effet, les indices de robustesse peuvent être eux aussi victimes d'artefact. La robustesse n'est pas la fiabilité. Pour qu'un résultat phylogénétique nouveau puisse être considéré comme fiable, il doit être répété plusieurs fois indépendamment, c'est-à-dire par des chercheurs différents et à partir de jeux de données indépendants.

2. Aperçu de la phylogénie des microorganismes eucaryotes

2.1 L'arbre du vivant

Avant le développement des méthodes moléculaires, il était impossible de connaître les relations phylogéniques connectant les êtres vivants et donc d'élaborer un arbre universel de l'évolution.

Whittaker, en 1969, commença à développer des méthodes moléculaires et à classer les êtres vivants dans 5 règnes : les animaux, les plantes, les champignons, les protistes et les bactéries. (Whittaker, 1969). Il avait également différencié les eucaryotes, organismes contenant une membrane nucléaire, des procaryotes, prédécesseurs supposés des eucaryotes et dépourvus de membrane nucléaire.

La découverte qui a remis en question les notions émises précédemment et qui a permis d'ordonner la diversité microbienne, et donc la diversité biologique, est apparue avec la mise en évidence de séquences moléculaires et le concept que ces séquences peuvent être utilisées pour relier les organismes. Cette formulation a été faîte par Carl Woese (1977) qui, par comparaison des séquences d'ARN ribosomique, a établi un arbre phylogénétique basé sur des séquences moléculaires qui peut être utilisé pour relier tous les organismes et reconstruire l'histoire de la vie (Woese et Fox, 1977 ; Woese, 1987). Woese articule les 3 lignées primaires de l'origine de l'évolution reconnues actuellement, en termes de domaines ou règnes : eucaryotes, bactériens (appelés initialement eubactéries) et archaées (appelé archaebactérie) (Fig. 2).

2.2 Phylogénie et évolution

2.2.1 L'ARNr et autres données moléculaires

Les séquences d'ARNr sont considérées comme de bons « chronomètres moléculaires » de l'évolution. Ces molécules sont codées en abondance par tous les organismes vivants et ont une fonction fondamentale essentielle. De plus, elles possèdent une structure conservée et universelle qui renferme des parties à évolution lente et des parties à évolution rapide. D'autres molécules sont aussi connues pour être de bons « chronomètres moléculaires », comme par exemple, le gène de la grande sous unité de l'ARN polymérase (Hirt *et al.*, 1999) ou les séquences protéiques du facteur d'élongation 2 (Hashimoto *et al.*, 1995) et des tubulines (Edgcomb *et al.*, 2001).

L'analyse des ARNr et d'autres données moléculaires confirment solidement la notion issue du siècle dernier que les organelles majeures des eucaryotes (mitochondries et chloroplastes des Plantae) proviennent de symbiontes bactériens qui ont subi des spécialisations pendant la coévolution avec leurs cellules hôtes (Margulis, 1975). Des comparaisons de séquences confirment que les mitochondries sont des représentants de protéobactéries (groupe incluant *Escherichia* et *Agrobacterium*) et que les chloroplastes proviennent des cyanobactéries (*Synechococcus* et *Gloeobacter*) (Sapp, 1994). Cependant, des arbres basés sur les séquences d'ARNr et d'autres molécules montrent que le noyau eucaryote semble être profondément ancré dans l'histoire de la vie. Les mitochondries et les chloroplastes sont apparus relativement plus tard. Cette évolution tardive est évidente par le fait que les mitochondries et les chloroplastes divergent d'organismes libres situés à la périphérie de l'arbre moléculaire du vivant. De plus, l'eucaryote le plus profondément ancré est dépourvu de mitochondrie (Cavalier-Smith, 1993). Ces organismes, peu étudiés, comme *Giardia*, *Trichomonas* et *Vairimorpha*, contiennent cependant au moins quelques gènes de type bactérien (Bui *et al.*, 1996). Ces gènes peuvent être la preuve d'une symbiose mitochondriale antérieure qui a été perdue ou d'un évènement de transfert de gènes.

Par ailleurs, les représentants modernes d'eucaryotes et d'archaea partagent beaucoup de propriétés qui diffèrent fondamentalement des cellules bactériennes. Un exemple de similarité et de différence est celui de la machinerie transcriptionnelle. Les ARN polymérases des eucaryotes et des archaea se ressemblent plus entre elles qu'elles ne ressemblent à celles des bactéries et, alors que toutes les cellules bactériennes utilisent le facteur sigma pour

réguler l'initiation de la transcription, les cellules eucaryotes et les archaea utilisent des protéines se liant à la boite TATA (Rowlands *et al.*, 1994). Le séquençage du génome de l'archaea *Methanococcus jannaschii* montre cependant que la lignée évolutive des archaea est indépendante de celles des eucaryotes et des bactéries (Bult *et al.*, 1996).

Toutes ces données ont permis d'émettre l'hypothèse d'une origine eucaryote de la vie sur Terre (Pace, 2009).

2.2.2 Limites de la phylogénie moléculaire « classique »

La phylogénie moléculaire basée uniquement sur une simple séquence de gène ou de protéine ne permet pas de déduire les relations les plus profondes dans le domaine des eucaryotes. Cette approche souffre d'un manque de résolution à cause du faible nombre de sites informatifs analysés, de l'hétérogénéité de la composition des séquences étudiées et des problèmes liés à l'attraction des longues branches (Stiller et Hall, 1999 ; Philippe et Germot, 2000 ; Dacks *et al.*, 2002). De plus, dans quelques cas, le gène analysé peut être transféré latéralement entre les lignées étudiées (Andersson, 2005).

Récemment, la disponibilité de quantités énormes de données de séquençage de génome et de projets d'EST (Expressed Sequence Tag) sur une vaste gamme d'eucaryotes a permis d'effectuer une évaluation phylogénétique à partir de "supermatrices" d'un plus ou moins grand nombre de gènes (Baldauf *et al.*, 2000 ; Simpson *et al.*, 2006 ; Bapteste *et al.*, 2002 ; Arisue *et al.*, 2005 ; Rodriguez-Ezpeleta *et al.*, 2007 ; Patron *et al.*, 2007 ; Burki *et al.*, 2007). Malgré tout, quelques erreurs peuvent persister comme l'attraction des longues branches (Bapteste *et al.*, 2002 ; Rodriguez-Ezpeleta *et al.*, 2007).

Des analyses récentes emploient donc des approches objectives de filtrage de données qui isolent et enlèvent les sites ou les taxa qui contribuent le plus à ces erreurs systématiques (Rodriguez-Ezpeleta *et al.*, 2007 ; Brinkmann *et al.*, 2005).

2.3 Les microorganismes eucaryotes

La phylogénie et donc la classification des eucaryotes est aujourd'hui en pleine phase de défrichage, pour ne pas dire en déchiffrage. Les phylogénies moléculaires nous révèlent des parentés insoupçonnées, une fois évacués les artefacts de reconstruction phylogénétique.

Notre compréhension de la biologie des eucaryotes est dominée par l'étude des plantes terrestres, des animaux et des champignons. Cependant, la majorité des eucaryotes en termes de taxons majeurs et probablement aussi en nombre de cellules pures, est composé exclusivement ou principalement de phylums unicellulaires. Un nombre important de ces phylums est encore mal caractérisé.

2.3.1 La cellule eucaryote

Les eucaryotes sont par définition des organismes à cellules complexes. Même "les plus simples" ont des noyaux avec de la chromatine fortement structurée, des introns et de grands complexes de spliceosomes pour les exciser (Collins et Penny, 2005). Ils contiennent aussi des pores membranaires complexes pour contrôler le trafic interne et externe (Jékely, 2005). Le cytoplasme est structuré par un vaste cytosquelette facilitant le trafic intracellulaire, l'endo-et exocytose et la locomotion amiboïde (Cavalier-Smith, 2002). La cellule eucaryote est également composées d'organelles incluant, au minimum, une mitochondrie ou un dérivé de mitochondrie (Embley et Martin, 2006) et un appareil de Golgi pour synthétiser et recycler les protéines (Mironov *et al.*, 2007). Les flagelles des eucaryotes sont de grandes structures extracellulaires complexes ancrées dans le cytoplasme sans rapport avec les structures bactériennes plus simples du même nom (Pazour *et al.*, 2005). Enfin, les cellules eucaryotes ont souvent des formes multiples fortement distinctes, incluant parfois des complexes multicellulaires.

D'autre part, les eucaryotes sont assez uniformes métaboliquement. Ceci, contraste avec les bactéries, dont la diversité métabolique est énorme et probablement en grande partie inconnue (Frias-Lopez *et al.*, 2008). Les eucaryotes dépendent d'anciennes bactéries endosymbiotiques pour leur production d'ATP (mitochondrie et chloroplaste). Ces anciens symbiontes sont probablement universels parmi les eucaryotes existant, bien qu'ils aient été réduits fonctionnellement de multiples fois (Embley et Martin, 2006).

2.3.2 Classification des organismes eucaryotes

Notre compréhension de la phylogénie des eucaryotes a commencé à s'unir autour des données de projets de séquençage à grande échelle (Burki *et al.*, 2007 ; Hackett *et al.*, 2007 ; Rodriguez-Ezpeleta *et al.*, 2005). Cependant, beaucoup de groupes d'eucaryotes, y compris des divisions majeures entières, ne sont que très peu connus (Adl *et al.*, 2005). De plus, la

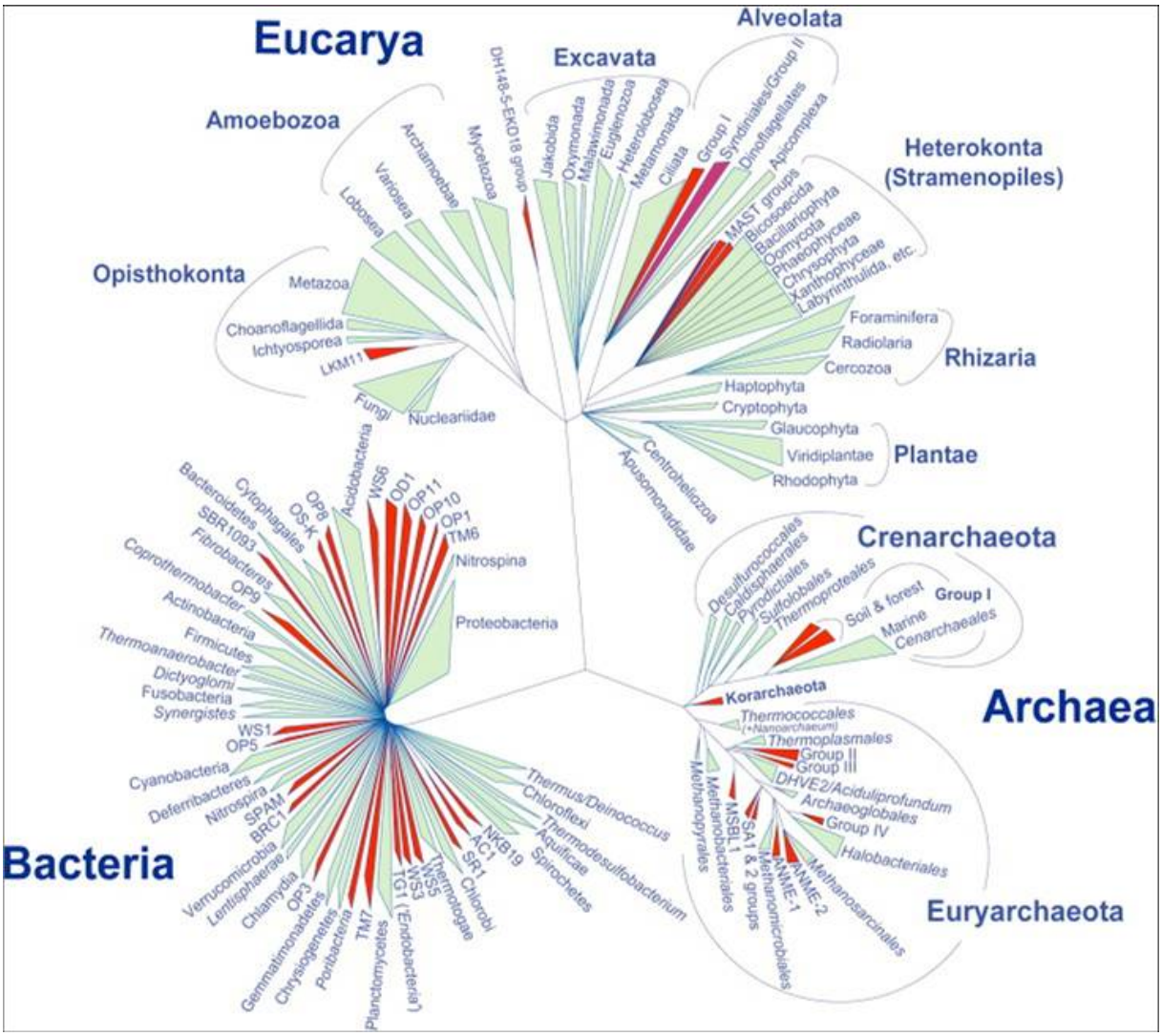


Figure 3: Schéma phylogénétique de l'arbre de vie basée sur les connaissances moléculaires actuelles (petite sous unité de l'ARNr et autres données moléculaires).
Les triangles verts représentent les phyla, les divisions ou les groupes de haut rang taxonomique pour lesquels un membre a été cultivé et/ou décrit correctement (par exemple, beaucoup d'espèces de protistes). Les triangles rouges représentent les divisions ou les lignées fortement divergentes sans espèces cultivées/décrites (d'après Lopéz-Garcia et Moreira, 2008).

Figure 4: Photos d'organismes appartenant au phylum des Opisthokonta. Les lignées des Metazoa (1 à 3), des Choanoflagellida (4), des Nucleariidae (5) et des Fungi (6) sont représentées.

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Fungi	52444	34	225
Metazoa	152977	44	75
Nucleariidae	14	0	0
Ichtyosporea	48	0	0
Choanoflagellida	57	1	0
Autres	26	0	1

Tableau 1: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) de quelques lignées d'Opisthokonta, à la date du 21/09/2009.

compréhension de la diversité cryptique des eucaryotes, supposée énorme, reste très faible (Slapeta *et al.*, 2005). Enfin, nous commençons tout juste à découvrir l'énorme diversité des pico et nano-eucaryotes, de taille bactérienne, découverts dans des banques de clones obtenues par amplification par PCR d'ADN environnemental (Moreira et Lopéz-Garcia, 2002).

Les eucaryotes les mieux étudiés peuvent maintenant être assignés à un des sept ou huit phyla définis par Lopéz-Garcia et Moreira, en 2008 (Fig. 3).

2.3.2.1 Les Opisthokonta

Ce grand phylum regroupe des organismes hétérotrophes, unicellulaires ou pluricellulaires. Les animaux et les champignons représentent les groupes taxinomiques les mieux étudiés (Steenkamp *et al.*, 2006). Chaque lignée a au moins 2 taxons unicellulaires frères ; les ichtyospores et les choanoflagellés dans le cas des animaux, et les nucléariides et les rozellida dans le cas des champignons (Lara *et al.*, 2009) (Fig. 4 et Tableau 1). Ces taxons regroupent des organismes dont le caractère ancestral commun est un flagelle unique postérieur.

Les premières branches des champignons sont celles du groupe paraphylétique des chytridiomycètes (James *et al.*, 2006). Il s'agit d'organismes unicellulaires aquatiques formant des pseudohyphes. Ce sont les seuls champignons avec des flagelles. Les autres champignons sont des organismes multicellulaires à hyphes ou unicellulaires (levures) avec une nutrition par absorption et peuvent produire des fructifications multicellulaires. Les champignons mycorhiziens sont parmi les plus étudiés. Leur symbiose avec les plantes terrestres est probablement apparue très tôt dans l'évolution et a probablement joué un rôle important dans l'invasion du milieu terrestre par les plantes (Wang et Qiu, 2006). Bien que l'arbre phylogénétique fongique soit maintenant clair, le nombre d'espèces non découvertes est probablement immense et est estimé à 95 % de toute les espèces existantes (Vandenkoornhuyse *et al.*, 2002 ; James *et al.*, 2006).

Les ichtyospores sont des parasites ou des symbiontes et possèdent un flagelle et une forme amiboïde (cellule déformable et émettant des pseudopodes). Les choanoflagellés, sont des organismes flagellés aquatiques connu pour leur ressemblance avec les cellules à collerettes (choanocytes) des éponges (Porifera). Ils codent des protéines liées au développement chez les métazoaires (King *et al.*, 2008).

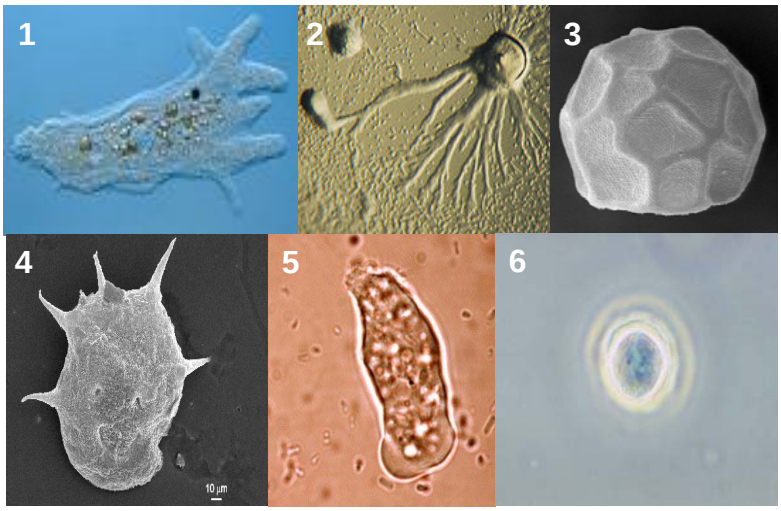


Figure 5: Photos d'organismes appartenant au phylum des Amoebozoa. Les lignées des Flabellinea (1), des Mycetozoa (2), des Lobosea (3 et 4), des Archamoebae (5) et des Centramoebida (6) sont représentées.

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Lobosea	136	0	0
Flabellinea	131	0	0
Archamoebae	40	1	2
Centramoebida	626	0	2
Mycetozoa	346	1	5

Tableau 2: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) des différentes lignées d'Amoebozoa, à la date du 21/09/2009.

2.3.2.2 Les Amoebozoa

Les Amoebozoa incluent plusieurs divisions d'amibes libres (Fig. 5 et Tableau 2) telles que les Mycetozoa (par exemple, *Dictyostelium*: vie libre, mitochondriale) et les Archamoebae (par exemple, *Entamoeba*: parasite de l'Homme, amitochondriale et *Mastigamoeba*: vie libre, amitochondriale) (Bapteste *et al.*, 2002). Le faible nombre d'organismes isolé et le peu de données moléculaires disponibles sur ces organismes, rend leur phylogénie incertaine. Les Amoebozoa ont une forme amiboïde et ont tendance à avoir des pseudopodes en forme de tube, un noyau simple et des mitochondries tubulaires à sillon (Adl *et al.*, 2005). Des formes amiboïdes sont par ailleurs retrouvées dans plusieurs autres super-groupes (Parfrey *et al.*, 2006). Leur taille varie entre quelques microns et plusieurs millimètres, et plusieurs petites formes restent probablement à découvrir. La formation de kyste pour survivre à la dessiccation ou pour envahir l'hôte parasité est répandue. La plupart des taxa sont sous forme libres dans les sols où ils jouent un rôle de prédateurs bactériens. Les amibes communes du sol, appartenant à la famille des *Arcellinidae* (division des Lobosea, genre *Tubulinea*), sont les seules amibes à tests (à coquilles), lesquels sont construits à partir de matériel organique.

Les Opisthokontes et les Amoebozoa font partis des Unikonts c'est-à-dire des eucaryotes uniflagellés. (Cavalier-Smith, 2002)

2.3.2.3 Les Plantae

Egalement appelé Archaeplastida, il s'agit du groupe d'eucaryotes dans lequel la photosynthèse est apparue en premier (Adl *et al.*, 2005). Toutes les espèces de ce groupe sont photosynthétiques à l'exception de quelques lignées parasites ou hétérotrophes mineures. On distingue 3 lignées distinctes, les rhodophytes, les glaucophytes et les chloroplastides (Fig. 6 et Tableau 3). Tous les autres eucaryotes photosynthétiques ont acquis leurs plastes par absorption d'un organisme de ce groupe (endosymbiose secondaire), sauf une exception récente (Nowack *et al.*, 2008), et plusieurs preuves montrent que les chloroplastes des Plantae sont monophylétiques (Reyes-Prieto *et al.*, 2007). La monophylie de leurs génomes nucléaires est plus ambigüe (Rodriguez-Ezpeleta *et al.*, 2005), probablement en raison de l'ancienneté du

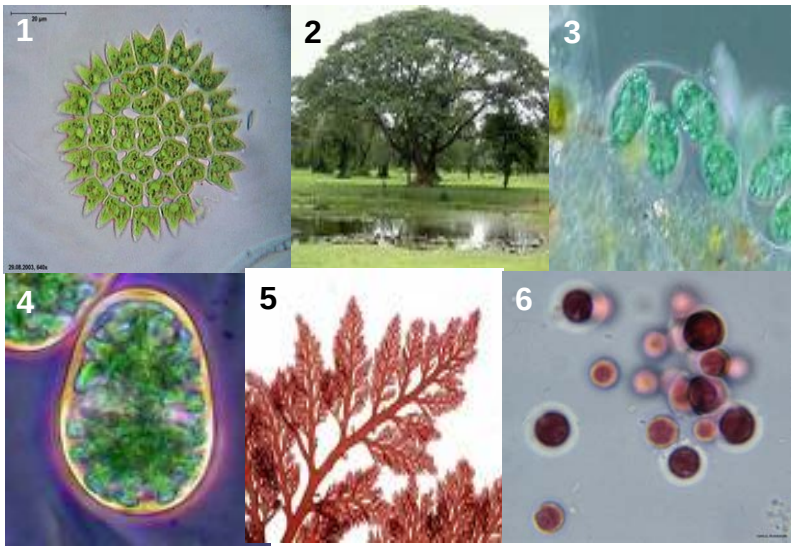


Figure 6: Photos d'organismes appartenant au phylum des Plantae. Les lignées des Chloroplastida (1 et 2), des Glaucophyta (3 et 4) et des Rhodophyta (5 et 6) sont représentées.

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Chloroplastida	101882	13	55
Glaucophyta	16	0	1
Rhodophyta	3203	1	2

Tableau 3: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) des différentes lignées de Plantae, à la date du 21/09/2009.

groupe et/ou de l'échantillonnage moléculaire extrêmement rare des taxa de rhodophytes et de glaucophytes.

Les glaucophytes sont des organismes unicellulaires de forme variable. Ils peuvent être biflagellés et en forme de coque ou non-flagellé et en forme de palme. Leur plaste (cyanelle) ressemble à ceux des algues rouges. Ces dernières possèdent des phycobiliprotéines et des thylacoïdes non entassés et sont dépourvus de chlorophylle b.

Les rhodophytes, plus communément appelés algues rouges, sont pour la plupart des organismes marins et multicellulaires. Certaines de ces grandes algues sont capables de produire une enveloppe extracellulaire de carbonate de calcium qui, à l'œil nu, leur donne un aspect de roche. Deux sous-groupes majeurs sont reconnus, *Bangiophyta* et *Florideophyta*, qui ont probablement tous deux inventé la multicellularité indépendamment.

Les chloroplastides regroupent les plantes terrestres et la plupart des algues vertes. Ils incluent les *Chlorophyta*, les *Ulvophyta*, les *Prasinophyta* et les *Strepsystera*. La multicellularité est apparue plusieurs fois chez les chloroplastides. Les *Chlorophyta*, les *Ulvophyta* et les *Strepsystera* sont composés d'organismes uni et multicellulaires.

2.3.2.4 Les Rhizaria

Une grande partie de ces organismes, mais pas tous, sont de forme amiboïde. Ces derniers ont tendance à avoir des pseudopodes (filose) dirigés et des coquilles (des tests) constituées de matériels divers. Les divisions les plus connues sont les Radiolaires, les Foraminifères et les Cercozoaires (Pawlowski et Burki, 2009) (Fig. 7 et Tableau 4).

Les Radiolaires sont des amibes exclusivement marines possédant "des squelettes" minéralisés internes dont sont émis des filopodes semblables à des bras (axopodia).

Les Foraminifères sont largement distribués dans tous types d'environnements marins, d'eaux douces et terrestres. Ils sont de formes amiboïdes et possèdent des pseudopodes réticulés avec un flux cytoplasmique bidirectionnel. Beaucoup construisent aussi des tests, qui sont organiques ou calcaires et avec une ou plusieurs loges. Les squelettes des foraminifères et des radiolaires ont considérablement contribué à leur conservation depuis le cambrien. Ces tests fossilisés sont utilisés en micropaléontologie comme des marqueurs biostratigraphiques et en paléo-océanographie comme des indicateurs de l'âge et de la profondeur des océans et des anciennes températures de l'eau (Habura *et al.*, 2004b).

Le reste des Rhizaria forme un grand assemblage hétérogène: les Cercozoa. Ce sont des amibes et des flagellés hétérotrophes. Certains ont acquis la photosynthèse par

Figure 7: Photos d'organismes appartenant au phylum des Rhizaria. Les lignées des Radiolaria (1 et 2), des Foraminifera (3 et 4) et des Cercozoa (5 et 6) sont représentées.

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Radiolaria	149	0	0
Foraminifera	928	0	0
Cercozoa	486	0	2

Tableau 4: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) de quelques lignées de Rhizaria, à la date du 21/09/2009.

endosymbiose secondaire d'une algue verte (Archibald, 2005). Les Cercozoa incluent aussi des espèces d'eau douce et/ou terrestre commune, comme les euglyphides, les haplosporides (endoparasites d'invertébrés d'eau douce et marine) et les plasmodiophorides (endoparasites importants de plantes ou d'algues straménopiles) (Cavalier-Smith et Chao, 2003).

2.3.2.5 Les Straménopiles (ou Heterokonta)

Les straménopiles sont caractérisés par des flagelles avec des poils tripartites (straménopiles) en rangées rigides, qui changent complètement le flux autour du flagelle pour permettre à la cellule de se trainer en avant. La plupart possède aussi un second flagelle, plus court, et lisse (d'où le nom alternatif "hétérokonte").

C'est un groupe extraordinairement divers incluant de nombreuses lignées d'unicellulaires hétérotrophes (bicosoecides), de parasites plasmodiales (oomycètes) comme *Phytophthora infestans*, responsable du mildiou de la pomme de terre, d'algues unicellulaires ubiquistes (diatomées et ochromonades) et de grandes algues multicellulaires géantes (xanthophytes et phaephytes) (Fig. 8 et Tableau 5).

L'échantillonnage environnemental suggère l'existence de divisions majeures supplémentaires dans ce groupe, comprenant en grande partie, voir entièrement, des espèces très petites (pico et nano-eucaryotes) (Moreira et Lopéz-Garcia, 2002).

2.3.2.6 Les Alvéolaires

Ils incluent les ciliés, les dinoflagéllés et les apicomplexes. Ils sont caractérisés par la présence d'alvéoles corticales à la base de leur membrane plasmique (Fig. 9 et Tableau 6).

Les ciliés sont très riches en espèces unicellulaires aquatiques caractérisées par une abondance de flagelles et un noyaux dimorphe, le micronoyau (germinatif) et le macronoyau (Prescott, 2000). Ce sont des organismes hétérotrophes pouvant se trouver sous forme libre (*Paramecium sp*.), parasite ou symbiotique.

Les dinoflagellés sont un groupe varié, principalement composés d'organismes unicellulaires avec des plaques caractéristiques et deux flagelles inégaux qui provoquent un mouvement de nage rotatoire unique. Bien que le groupe soit originellement photosynthétique, seulement environ la moitié des espèces existantes l'est toujours. Les dinoflagellés sont des symbiontes importants du corail et d'autres hydrozoaires et sont la principale source d'efflorescence d'algues nuisibles (marées rouges),

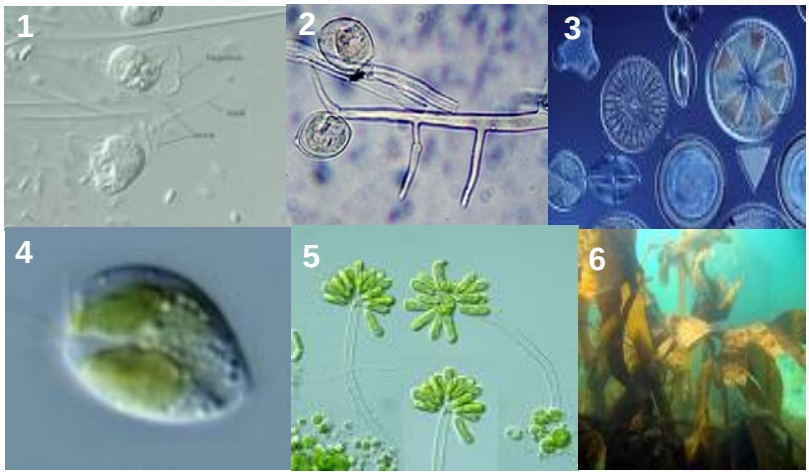


Figure 8: Photos d'organismes appartenant au phylum des Straménopiles. Les lignées des Bicosoecida (1), des Oomycota (2), des Bacillariophyta (3), des Ochromonadaceae (4), des Xantophyceae (5) et des Phaeophyceae (6) sont représentées.

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Bicosoecida	53	0	0
Oomycota	1431	3	4
Bacillariophyta (Diatoms)	1112	2	3
Ochromonadaceae (ochromonads)	207	0	2
Xantophyceae	180	0	0
Phaeophyceae	947	0	0
Autres	794	0	7

Tableau 5: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) de quelques lignées de Straménopiles, à la date du21/10/09

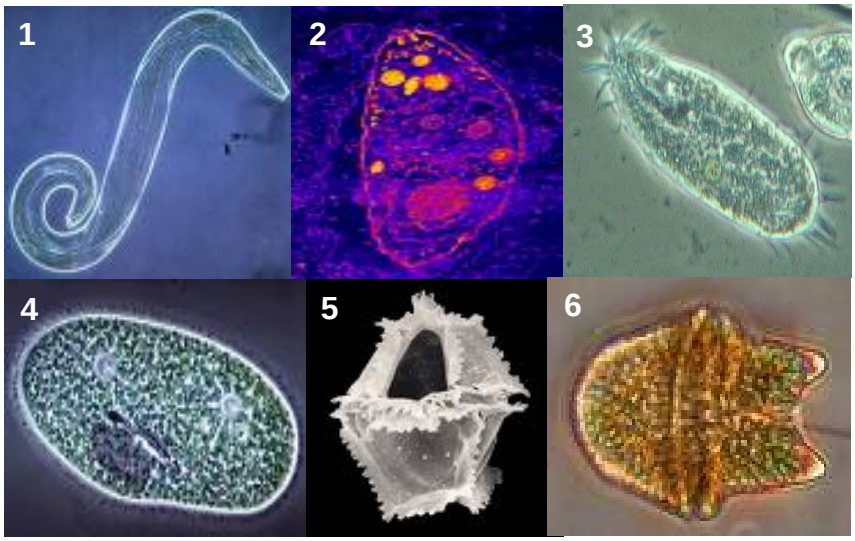


Figure 9: Photos d'organismes appartenant au phylum des Alveolata. Les lignées des Apicomplexa (1 et 2), des Cilita (3 et 4) et des Dinoflagellates (5 et 6) sont représentées.

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Apicomplexa	2652	9	30
Ciliata	1348	2	3
Dinoflagellates	2519	0	1
Autres	176	0	1

Tableau 6: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) de quelques lignées d'Alveolata, à la date du 21/09/2009.

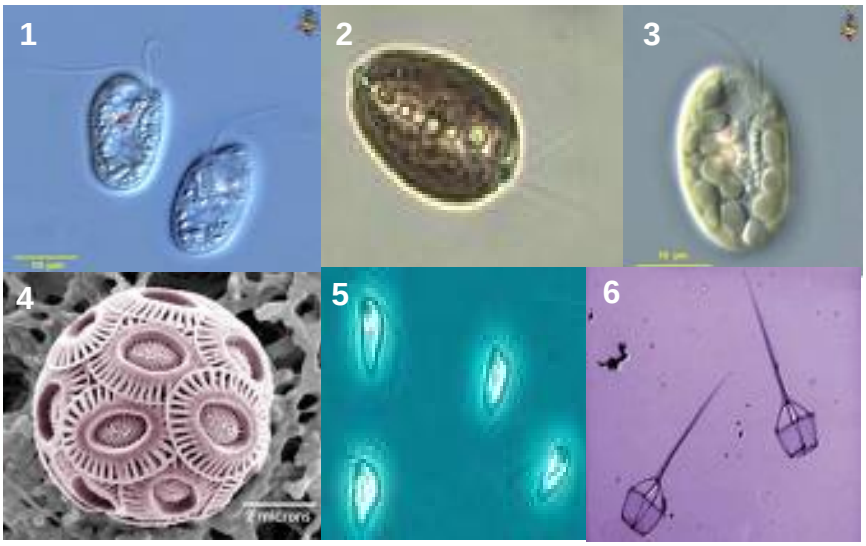


Figure 10: Photos d'organismes appartenant au phylum des Cryptophyta (1 à 3) et des Haptophyta (4 à 6).

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Cryptophyta	180	2	1
Haptophyta	307	0	3

Tableau 7: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) des différentes lignées de Cryptophyta/Haptophyta, à la date du 21/09/2009.

qui produisent quelques unes des plus puissantes neurotoxines connues. Ils possèdent aussi quelques uns des plus grands génomes nucléaires connus (entre 3000 et 215000 Mb), avec de grandes quantités de séquences d'ADN répété (Hackett *et al.*, 2005).

Les apicomplexes sont le groupe frère des dinoflagellés et incluent quelques uns des plus importants agents pathogènes d'invertébrés et de vertébrés. Presque tous sont des parasites obligatoires intracellulaires, comme les agents responsables de la malaria (*Plasmodium spp.*) et de la toxoplasmose (*Toxoplasma sp.*). Leur nom est tiré des caractéristiques de leur complexe apical, qui permet l'attachement et la pénétration initiale de la cellule hôte. Toutes les espèces conservent un plaste rudimentaire (apicoplaste), très probablement originaire d'une algue rouge (Fast *et al.*, 2001).

2.3.2.7 Les Haptophytes et les Cryptophytes

Il s'agit principalement d'algues à chlorophylle c. Cependant, bien que leurs plastes partagent clairement une origine commune avec les plastes des straménopiles, il y a peu de preuves que leurs génomes nucléaires en aient une. Le groupe des Cryptophytes et des Haptophytes, s'il s'agit d'un groupe, est potentiellement très grand. Deux nouvelles lignées majeures ont récemment été découvertes. (Shalchian-Tabrizi *et al.*, 2007 ; Not *et al.*, 2007).

Les Haptophytes sont nommés ainsi pour leur haptonème, un appendice antérieur utilisé pour la capture de proies et l'adhésion. Ils incluent les coccolithophorides, qui sont des organismes unicellulaires couverts d'écailles de carbonate de calcium (coccolithes).

Les Cryptophytes sont des organismes unicellulaires relativement petits (entre 2 et 10 µm de diamètre) et principalement retrouvés dans les eaux froides ou profondes. Ils sont abondants et ubiquistes et sont généralement impliqués dans des endosymbioses provisoires (Patron *et al.*, 2007) (Fig. 10 et Tableau 7).

2.3.2.8 Les Excavés

Les taxa classés comme excavés sont représentés par des organismes unicellulaires possédant un grand sillon creux au niveau de leur extrémité antérieure dans lequel ils prennent au piège des particules de nourriture en suspension à l'aide d'un flagelle (Simpson, 2003). Leur phylogénie est complexe, puisqu'ils ont tendance à avoir des taux d'évolution moléculaire extrêmement rapides. Cependant, la monophylie de ce groupe a été récemment

Figure 11: Photos d'organismes appartenant au phylum des Excavata. Les lignées des Euglenozoa (1 et 2), des Heterolobosea (3), des Jakobida (4 et 5) et des Fornicata (6) sont représentées.

Lignées	Nombre de gènes 18S référencés dans GenBank	Nombre de génomes séquencés	Nombre de génomes en cours de séquençage
Euglenozoa	1039	4	4
Heterolobosea	349	0	1
Jakobida	16	0	0
Fornicata	74	1	3

Tableau 8: Aperçu du nombre de gènes 18S référencés dans GenBank (http://www.ncbi.nlm.nih.gov/) et du nombre de génomes séquencés et en cours de séquençage (http://www.genomesonline.org/) de quelques lignées des Excavata, à la date du 21/09/2009.

démontré (Hampl *et al.*, 2009). Par commodité, ils sont divisés en "excavés avec mitochondries", ce qui inclut les Euglenozoa, les Heterolobosea et les Jakobida et en "excavés sans mitochondries" (Simpson *et al.*, 2006).

2.3.2.8.1 Les excavés avec mitochondries

Les Euglenozoa sont de petites cellules uni- ou biflagellées, dont beaucoup sont des parasites comme les agents responsables de la maladie du sommeil (*Trypanosoma sp.*) et des leishmanioses (*Leishmania sp.*). Certains euglènes ont une forme de vie libre et sont capables d'ingérer des cellules eucaryotes entières. L'espèce *Euglena gracilis* a ainsi acquis un chloroplaste d'algue verte.

Les Heterolobosea sont surtout des amibes nues. Ils sont abondants, ubiquistes et leur importance écologique est mal comprise, mais probablement très importante. Ce sont des prédateurs bactériens du sol ou d'eau douce, et parfois des pathogènes facultatifs de l'homme (*Naegleria fowleri*) (Maclean *et al.*, 2004).

Les Jakobides sont de petits prédateurs de bactéries à forme de vie libre et sont particulièrement réputés pour leur morphologie mitochondriale variable (Lara *et al.*, 2006) (Fig. 11 et Tableau 8).

2.3.2.8.2 Les excavés sans mitochondries

On connaît cet immense groupe, probablement ancestral, seulement comme des unicellulaires vivants dans des habitats anaérobique ou micro-aérobiques, souvent comme commensaux ou parasites. Leur structure cellulaire interne simplifiée et le manque apparent de mitochondries ou d'organelles dérivés de mitochondries ont suscité l'hypothèse d'Archaezoa, suggérant que ceux-ci étaient les restes des premières lignées d'eucaryotes prémitochondriées (Cavalier-Smith et Chao, 1996). Cependant, des gènes ancestraux mitochondriaux ont été trouvés dans leurs génomes nucléaires et des vestiges de mitochondries fortement réduites ont récemment été découverts dans beaucoup de ces organismes (Embley et Martin, 2006). Ainsi l'hypothèse des Archaezoa est maintenant oubliée. Néanmoins, ces taxa apparaissent toujours comme les toutes premières branches dans les arbres moléculaires enracinés (Arisue *et al.*, 2005). Ce positionnement est souvent interprété comme un artefact d'attraction des longues branches (Philippe et Germot, 2000).

2.3.2.9 Position incertaine

En 1999, 230 protistes cultivés avaient un positionnement incertain. En 2005, ce nombre a diminué à 204 (Adl *et al.*, 2005). La plupart d'entre eux sont des petits flagellés hétérotrophes ou des amibes à vie libre, ou des parasites de diverses sortes. Plusieurs seront sans doute classés dans un ou plusieurs groupes décrits ci-dessus. Cependant, des études de PCR suggèrent l'existence de lignées majeures d'eucaryotes non découvertes (Moreira et Lopéz-Garcia, 2002), dont beaucoup sont composées probablement de nano-et pico-eucaryotes. Même soi-disant connues, des espèces peuvent être les représentants uniques de lignées majeures non soupçonnées. Par exemple, des études récentes de PCR montrent que les Apusomonades forment un grand groupe divers pouvant être le groupe frère des Opisthokontes (Kim *et al.*, 2006).

3. La diversité microbienne dans l'environnement

3.1 L'environnement sol : un réservoir de diversité microbienne

Le sol est probablement le plus intéressant de tous les environnements naturels pour les microbiologistes, en ce qui concerne la taille de la communauté microbienne et la diversité d'espèces présentes. Un gramme de sol de forêt contient environ 4.10^7 bactéries, tandis qu'un gramme de sol cultivé ou de prairie contient environ 2.10^9 bactéries (Paul et Clark, 1989). Les champignons constituent également une fraction très importante de la biomasse microbienne. C'est par exemple le cas des forêts tempérées ou boréales dans lesquels la biomasse des seuls champignons symbiotiques a pu être estimée à 1/3 de la biomasse microbienne totale (Högberg et Högberg, 2002).

En se basant sur la cinétique de réassociation de l'ADN, le nombre de génomes bactériens distincts a été estimé entre 2000 et 18000 génomes par gramme de sol (Torsvik *et al.*, 1998; Torsvik et Ovreas, 2002 ; Doolittle, 1999). Ce nombre est une sous estimation. En effet, il se peut que des génomes ne soient pas récupérés, notamment les génomes représentant des espèces rares, et soient exclus de ces analyses. L'extrême hétérogénéité spatiale, la nature polyphasique (incluant les gaz, l'eau et le matériel solide), et les propriétés chimiques et biologiques complexes de l'environnement sol contribuent à la diversité microbienne présente dans un échantillon de sol.

3.2 Caractéristiques de l'environnement sol

Le sol est constitué de particules minérales de tailles, de formes et de caractéristiques chimiques différentes, ainsi que d'organismes vivants (biota) et de composés organiques à différent stades de décomposition. La formation de complexes d'argile et de matière organique et la stabilisation des particules d'argiles, de sables et de limons par la formation d'agrégats sont les caractéristiques structurales dominantes de la matrice sol. Ces agrégats peuvent s'étendre sur environ 2 mm ou plus (macro-agrégats) jusqu'à des fractions de l'ordre du micromètre pour les bactéries et les particules colloïdales. Les microorganismes du sol adhèrent ou s'adsorbent souvent fortement sur les particules du sol comme les grains de sables ou les complexes d'argiles et de matière organique. Leurs micro-habitats sont la surface des agrégats et les espaces complexes (pores) entre et à l'intérieur des agrégats (Foster, 1988).

Le métabolisme et la survie des microorganismes du sol et donc la composition de la communauté microbienne est également fortement influencée par la disponibilité en eau et en nutriments et par d'autres facteurs exogènes comme le pH, la disponibilité en oxygène ou la température. Le sol est également un important réservoir de carbone organique. Il est le siège de la transformation de la plupart de la matière organique, issue des plantes, des animaux et des microorganismes, en humus par une combinaison de processus microbiologiques et abiotiques. Néanmoins, des substances humiques stables restent récalcitrantes au processus de décomposition microbien ; la demi-vie de ces complexes stables de matière organique est approximativement de 2000 ans (Paul et Clark, 1989).

3.3 Accéder à la diversité des microorganismes du sol

Les approches directes de mise en culture ou indirecte moléculaire peuvent être utilisées pour explorer la diversité microbienne présente dans un sol. La mise en culture et l'isolement de microorganismes est la méthode traditionnelle mais, seulement 0,1% à 1% des bactéries du sol sont cultivables en utilisant des méthodes de mise en culture standard (Torsvik et Ovreas, 2002 ; Amann *et al.*, 1995). La diversité microbienne du sol reste principalement inexplorée et seule une infime portion a été caractérisée en utilisant la mise en culture et l'isolement.

Pour contourner les limites des approches de mise en culture, des méthodes moléculaires indirectes basées sur l'isolement et l'analyse des acides nucléiques (ADN et ARN) ont été développées à partir d'échantillons de sol sans mise en culture des microorganismes. Plusieurs protocoles d'isolement de l'ADN microbien du sol ont été publiés (Ogram *et al.*, 1987 ; Steffan *et al.*, 1988 ; Zhou *et al.*, 1996 ; Miller *et al.*, 1999 ; Hurt *et al.*, 2001).

Suivant l'approche utilisée, le degré d'information s'étend de la simple séquence génomique à l'ensemble des gènes réellement exprimés par une communauté de microorganismes du sol.

3.3.1 L'approche de clonage et séquençage

Des études phylogénétiques peuvent être effectuées par amplification par PCR du gène de l'ARNr 16S et 18S, en utilisant des amorces universelles bactériennes ou eucaryotes. Les séquences de gènes d'ARNr 16S et 18S obtenues sont confrontées à celles référencées dans les bases de données (GenBank, Ribosomal Database Project, etc) afin de leur assigner une appartenance à une espèce ou à un groupe taxinomique. Des seuils de similarité sont définis pour les différents marqueurs génétiques possibles. Pour l'ARNr 18S, ce seuil est généralement situé à 97 %.

Cette méthode a permis de mettre en évidence une modification de la composition d'une communauté de micro-organismes eucaryotes dans des sols sous une végétation soumise à des concentrations élevées en CO_2 atmosphérique (Lesaulnier *et al.*, 2008). Elle a également permis de découvrir, à partir de sols archivés, de nombreuses séquences de cercozoa et du 'super-groupe' des Amoebozoa, dont douze sont dispersées dans divers clades (Moon-van der Staay *et al.*, 2006).

Cependant, la majorité des travaux s'intéressant à la diversité eucaryote dans les sols se restreint à des groupes taxinomiques précis, comme les champignons (Fierer *et al.*, 2007 ; O'Brien *et al.*, 2005 ; Porter *et al.*, 2008 ; Yergeau *et al.*, 2007), les nématodes (Yergeau *et al.*, 2007) et les cercozoa (Bass et Cavalier-Smith, 2004). Ces études suggèrent que la diversité et l'abondance des microorganismes eucaryotes est encore largement sous-estimée, et plus particulièrement pour les champignons puisque très peu d'espèces fongiques sont communes à deux sols différents (Fierer *et al.*, 2007)

D'autres études de classification et de comparaison de la diversité microbienne dans différents habitats de sol ont mis en évidence des changements de la structure de la communauté bactérienne suivant l'impact de facteurs environnementaux (Smit *et al.*, 2001 ; Zhou *et al.*, 2002b). Aucune étude de ce type n'a été réalisée sur les communautés de microorganismes eucaryotes.

3.3.2 Les profils électrophorétiques : empreintes moléculaires

Ces méthodes qualitatives et semi-quantitatives sont basées sur l'électrophorèse. Il s'agit de la DGGE/TGGE (électrophorèse en gradient dénaturant linéaire ''chimique'' ou ''thermique'', respectivement) ou encore de la technique RFLP, qui permet un crible préalable des différents taxa avant le séquençage. Elles sont assez résolutives car elles révèlent les taxa spécifiques et abondants dans certaines conditions (Moon-van der Staay *et al.*, 2006) qui peuvent être identifiés par excision et séquençage des fragments séparés par électrophorèse. Cependant, les taxa rares peuvent être masqués, dans la mesure où la bande correspondante peut se confondre avec d'autres, plus intenses.

Ces approches sont particulièrement appropriées à l'étude de modifications de la diversité des communautés en fonction de variables environnementales. Lawley *et al.*, 2004 ont ainsi étudié la diversité des micro-eucaryotes dans différents échantillons de sol de l'antarctique par séquençage des bandes d'intérêt suite à une analyse de polymorphisme des fragments de restriction (RFLP) ou ARDRA (analyse de restriction des fragments d'ADNr) lorsqu'il s'agit de l'ADNr. Cette étude montre que des espèces provenant de tous les 'super-groupes' eucaryotes sont représentées, bien que le nombre d'unités taxinomiques opérationnelles (OTU) révélé par profil RFLP puisse sous-estimer la diversité.

3.3.3 Limites des méthodes moléculaires « classiques »

L'utilisation des ARNr a permis de révéler non seulement de nouvelles espèces mais aussi de nouveaux phylums dont aucun représentant n'a été isolé à ce jour. Cependant, l'estimation de la diversité génétique des microorganismes reste incertaine en raison de plusieurs biais (problèmes de séquençage des ARNr, obtention de séquences chimériques, interprétation erronée de séquences à évolution rapide…). On estime que 9% des séquences d'ARNr 16S répertoriées dans les banques présentent des anomalies (Ashelford *et al.*, 2006).

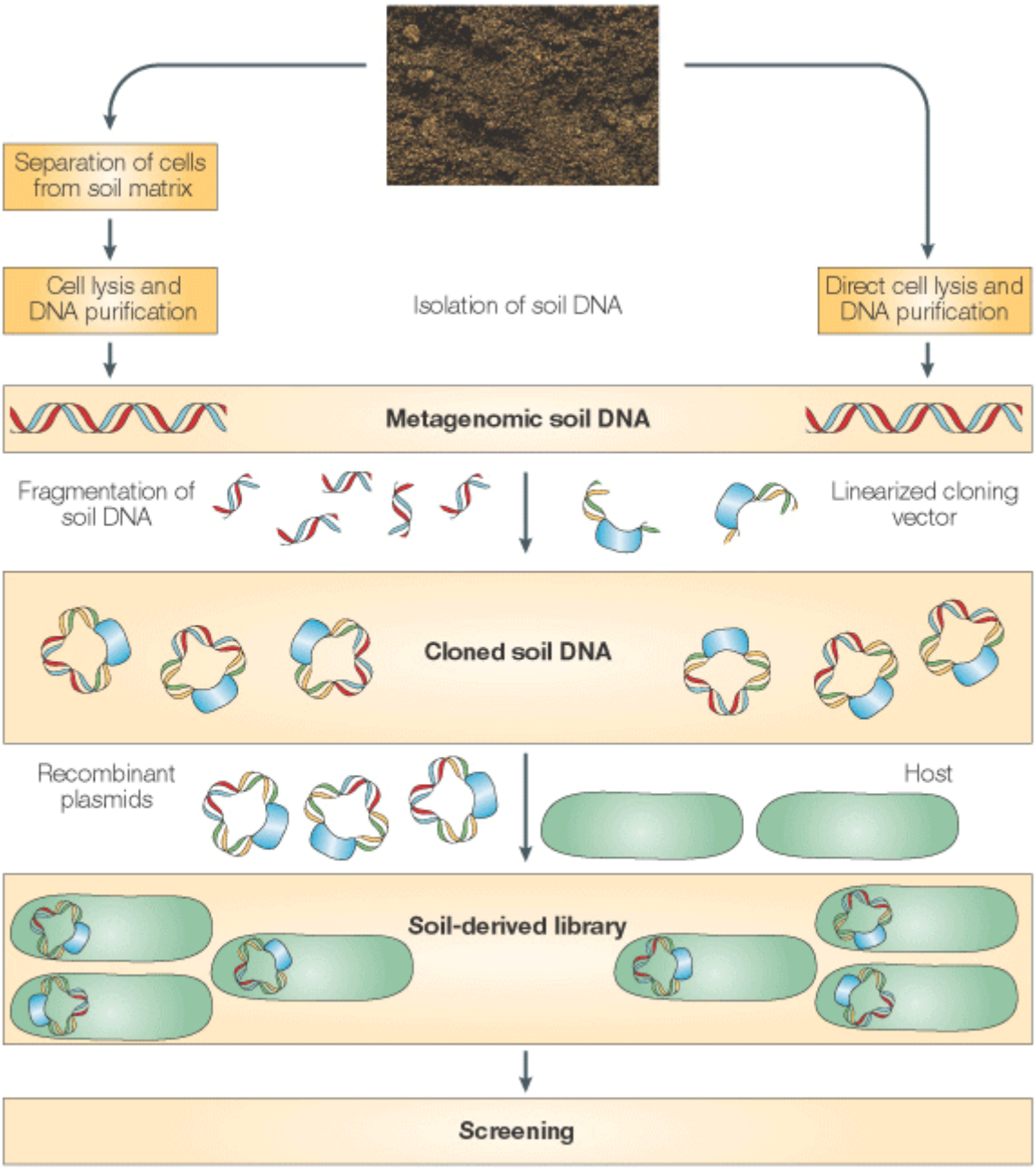


Figure 12: Les différentes étapes de l'approche métagénomique (d'après Daniel, 2005). L'ADN de sol est récupéré soit par la séparation des cellules des particules de sol suivie par la lyse des cellules et l'isolement de l'ADN, soit par la lyse directe des cellules contenues dans le sol. L'ADN récupéré est fragmenté et inséré dans un vecteur de clonage (plasmide, cosmide, fosmide ou BAC). Après l'introduction des vecteurs recombinants dans un hôte de clonage bactérien approprié, des stratégies de sélection peuvent être conçues pour identifier ces clones qui pourraient contenir de nouveaux gènes d'intérêt.

Par ailleurs, la quantification des organismes présents dans un échantillon environnemental à partir de données basées sur la séquence du gène de l'ARNr reste difficile en raison de la variation du nombre de copies génomiques de ce gène suivant l'organisme (Ribosomal RNA Operon Copy number Database, http://rrnbdb.cme.msu.edu, Klappenbach 2000).

Enfin, ces approches moléculaires ne donnent que très peu d'informations sur le rôle et les fonctions assurées par l'ensemble des communautés microbiennes dans l'environnement. Elles ne permettent donc pas d'estimer la diversité fonctionnelle *in situ* de ces microorganismes.

3.3.4 L'approche métagénomique

La métagénomique, définie comme l'analyse génomique indépendante de la mise en culture de tous les microorganismes présents dans une niche environnementale particulière (Handelsman *et al.*, 1998), a été développée afin de découvrir la diversité microbienne d'environnements naturels (Fig. 12).

Cette approche de plus en plus sophistiquée repose sur l'isolement direct d'ADN d'un habitat défini, suivi par un clonage (dans un hôte comme *Escherichia coli*) des génomes complets de toute la communauté microbienne présente dans cet habitat. La banque d'ADN résultante est alors analysée pour des fonctions et des séquences d'intérêt.

La métagénomique permet soit une analyse basée sur la séquence, soit une analyse basée sur la fonction des microorganismes non cultivés. La métagénomique fonctionnelle implique le criblage de banques métagénomiques pour un phénotype particulier, par exemple la tolérance au sel, la production d'antibiotique ou l'activité d'enzymes et ensuite l'identification de l'origine phylogénétique de l'ADN cloné (Dinsdale *et al.*, 2008). D'autre part, les approches basées sur la séquence impliquent un criblage des clones pour les gènes d'ARNr 16S (Liles *et al.*, 2003 ; Rondon *et al.*, 2000) ou 18S (Grant *et al.*, 2006 ; Bailly *et al.*, 2007 ; Urich *et al.*, 2008) fortement conservés afin de rechercher l'origine taxinomique des microorganismes composant la communauté de l'écosystème étudié.

La construction et le criblage de banques métagénomiques dérivées du sol dépendent de plusieurs facteurs : de la composition de l'échantillon de sol ; de l'échantillonnage et du stockage des échantillons de sol ; de la méthode d'extraction d'ADN utilisée pour récupérer de l'ADN de haute qualité ; de la représentativité des ADN de la communauté microbienne

présente dans l'échantillon de sol ; du système vecteur-hôte utilisé pour le clonage et de la stratégie de criblage (Daniel, 2005).

3.3.4.1 Isolement d'ADN à partir du sol

La construction de banques métagénomiques commence par la collecte des échantillons. Comme les échantillons de sol sont hétérogènes, les caractéristiques physiques, chimiques et biotiques comme la taille des particules, le type de sol, la teneur en eau, le pH, la température et le couvert végétal sont utiles pour l'évaluation et la comparaison des résultats des études basées sur le sol. Comme les populations microbiennes sont grandes, les volumes des échantillons peuvent être petits (inférieur à 500g dans la plupart des études) (Miller *et al.*, 1999 ; Henne *et al.*, 1999 ; Rondon *et al.*, 2000). La perturbation du sol lors de l'échantillonnage pourrait changer la composition des communautés microbiennes du sol, par conséquent le temps de stockage et de transport d'un échantillon doit être court.

Les méthodes d'extraction d'ADN peuvent être divisées en deux catégories : par lyse directe des cellules contenues dans l'échantillon suivie par une séparation de l'ADN à partir de la matrice et des débris cellulaires (initié par Ogram *et al.*, 1987) ; ou par séparation des cellules à partir de la matrice sol suivie d'une lyse des cellules (initié par Holben *et al.*, 1988). L'ADN brut récupéré par les deux méthodes est purifié selon des procédures standards. Une plus grande quantité d'ADN est récupérée en utilisant des approches de lyses directes, par exemple, Gabor et al. (2003) enregistrent une réduction de 10 à 100 fois du rendement d'ADN extrait en utilisant l'approche de séparation cellulaire comparée avec l'approche de lyse directe. De plus, 61 à 93% de l'ADN extrait par lyse directe était d'origine eucaryote, alors que plus de 92% de l'ADN extrait par lyse indirecte était bactérien (Gabor *et al.*, 2003). Ce résultat s'explique par la plus grande taille des génomes eucaryotes (compris entre 3 et 215000 Mb) par rapport aux génomes bactériens (compris entre 0,6 et 9,5 Mb) (Vellai et Vida, 1999). Enfin, l'ADN récupéré par lyse indirecte semble être moins contaminé par des composés matriciels comme les substances humiques et présente une taille moyenne plus grande que celle typiquement obtenue par l'approche de lyse directe (Courtois *et al.*, 2001). La lyse indirecte semble être la plus appropriée pour la construction de banques de grands inserts.

3.3.4.2 Systèmes vectoriels

Le choix d'un vecteur de clonage dépend de la qualité de l'ADN de sol isolé, de la taille moyenne des inserts de la banque, du nombre de copies du vecteur exigé, de l'hôte et de la stratégie de criblage qui sera utilisée, donc tout dépend du but de l'étude (Daniel, 2005).

Les banques de petits inserts sont utile pour l'isolement de gènes seuls ou de petits opérons codant de nouvelles fonctions métaboliques (Henne *et al.*, 1999 ; Henne *et al.*, 2000 ; Majernik *et al.*, 2001 ; Knietsch *et al.*, 2003b; Yun *et al.*, 2004; Riesenfeld *et al.*, 2004b). Elles sont généralement construites dans des plasmides.

Les banques de grands inserts sont plus appropriées pour récupérer des voies complexes qui sont codées par des grands groupes de gènes ou pour la caractérisation des génomes de microorganismes de sol non cultivés (Rondon *et al.*, 2000 ; Courtois *et al.*, 2003). L'utilisation de vecteur de type BAC (chromosome bactérien artificiel), cosmide ou fosmide est recommandée.

3.3.4.3 Criblage des banques métagénomiques

Plusieurs techniques ont été utilisées pour identifier et récupérer des gènes à partir de banques de sol. À cause de la complexité du métagénome du sol, des méthodes de criblages à haut débit et sensibles sont exigées. En principe, les cribles de banques peuvent être basés soit sur la séquence nucléotidique (approches basées sur la séquence), soit sur l'activité métabolique (approches basées sur la fonction).

3.3.4.3.1 Approches basées sur la séquence

La PCR est plus communément utilisée pour un criblage des banques d'ADN du sol basée sur la séquence (Zhou *et al.*, 2002b ; Courtois *et al.*, 2001 ; Courtois *et al.*, 2003 ; Liles *et al.*, 2003 ; Precigou *et al.*, 2001). L'hybridation ciblée utilisant des sondes spécifiques a aussi servi à cribler des banques (Knietsch *et al.*, 2003a ; Demaneche *et al.*, 2009). Ces deux approches nécessitent des sondes et des amorces appropriées obtenues à partir de régions conservées de gènes connus et de produits de gène. Leur application est limitée à l'identification de nouveaux membres de familles de gènes connues. Cette approche a été utilisée pour identifier des gènes codant l'ARNr 16S (Liles *et al.*, 2003 ; Rondon *et al.*, 2000) et des enzymes avec des domaines fortement conservés comme les polykétides synthases

(Courtois *et al.*, 2003), les acides gluconiques réductases (Eschenfeldt *et al.*, 2001) et les nitrile hydratases (Precigou *et al.*, 2001).

Le séquençage aléatoire de banques issues de sol est une autre approche pour caractériser l'écosystème sol à un niveau génomique, mais la richesse en espèces présentes exige des efforts de séquençage et d'assemblage énorme.

La technologie de « puce à ADN » a été utilisée pour analyser le métagénome du sol et pour définir le profil des banques métagénomiques (Wu *et al.*, 2001 ; Zhou et Thompson, 2002a ; Sebat *et al.*, 2003 ; Pathak *et al.*, 2009). Par exemple, les gènes codant des réactions clefs dans le cycle de l'azote ont été détectés en utilisant des puces à ADN à partir d'échantillons de sol. Des informations sur la composition et l'activité de la communauté microbienne complexe du sol ont été obtenues (Wu *et al.*, 2001). Cependant, les méthodes de puces à ADN pour la détection de gènes sont 100 à 10000 fois moins sensibles que la PCR (Zhou et Thompson, 2002a). Cette différence pourrait empêcher l'analyse de séquences de microorganismes peu abondants dans le sol. L'amélioration de la sensibilité et de la spécificité de ces puces à ADN est un enjeu pour l'analyse des ADN et des ARN du sol.

3.3.4.3.2 Approches basées sur la fonction

La plupart des méthodes de criblage pour isoler des gènes ou des groupes de gènes codant de nouveaux biocatalyseurs ou la synthèse de petites molécules sont basées sur la détection de l'activité des clones contenus dans la banque. Comme les informations de séquences ne sont pas exigées, c'est la seule stratégie qui a le potentiel d'identifier de nouvelles classes de gènes qui codent des fonctions connues ou inconnues.

Cette approche a été validée par l'isolement de nouveaux gènes qui codent des hydrolases (Henne *et al.*, 1999 ; Rondon *et al.*, 2000 ; Lee *et al.*, 2004 ; Gupta *et al.*, 2002 ; Yun *et al.*, 2004 ; Mayumi *et al.*, 2008 ; Kim *et al.*, 2008), des protéines de résistance au métaux (Mirete *et al.*, 2007), des enzymes de résistance aux antibiotiques (Riesenfeld *et al.*, 2004a) et des antibiotiques (Gillespie *et al.*, 2002; Courtois *et al.*, 2003). La plupart des biomolécules récupérées par criblage fonctionnel de banques de sol sont faiblement proches ou entièrement sans rapport avec celles répertoriées dans les bases de données. Cela confirme que la quantité d'ADN de sol qui a été clonée et examinée représente seulement la partie visible de l'iceberg en ce qui concerne la découverte de nouveaux produits naturels à partir du métagénome du sol.

Une méthode est d'effectuer des tests directs sur colonies pour une fonction spécifique. Par exemple, des colorants chimiques et des substrats d'enzymes insolubles ou dérivés de chromophores peuvent être incorporés dans le milieu de croissance solidifié avec de l'agar pour contrôler les fonctions enzymatiques des clones individuels. La sensibilité de ces cribles permet de détecter des clones rares (Knietsch *et al.*, 2003b ; Gupta *et al.*, 2002).

Une autre technique permettant la détection de clones fonctionnels est l'utilisation de souches hôtes ou des mutants de souches hôtes qui exigent une complémentation hétérologue pour croître dans des conditions sélectives. Un exemple est la complémentation d'une souche d'*E. coli* déficiente pour un transporteur Na^+/H^+ avec une banque de sol, qui a mené à l'identification de deux nouveaux gènes codant des transporteurs Na^+/H^+ à partir d'une banque contenant 1.480.000 clones (Majernik *et al.*, 2001). Bien que les cribles basés sur la fonction aboutissent d'habitude à l'identification de gènes complet (et donc des produits de gènes fonctionnels), cette approche est limitée par l'expression du gène cloné et par le fonctionnement de la protéine codée dans un hôte étranger. Dans la plupart des études, *E. coli* a été utilisé avec succès comme hôte pour des cribles fonctionnels. Récemment, d'autres hôtes bactériens comme *Streptomyces* ou des souches de *Pseudomonas* ont été utilisés pour étendre la gamme de gènes détectables par des cribles fonctionnels (Wang *et al.*, 2000 , Courtois *et al.*, 2003).

Comme l'expression dans des hôtes bactériens est d'habitude limitée aux gènes bactériens et que l'ADN métagénomique de sol, selon la méthode d'extraction, contient une quantité importante d'ADN eucaryote (Gabor *et al.*, 2003), l'utilisation d'hôtes eucaryotes pourrait aussi être envisagée pour les cribles fonctionnels de banques de sol.

3.3.4.4 Augmenter la fréquence de détection de gènes

Le nombre de clones qui doit être criblé pour récupérer les gènes d'intérêt est déterminé par la fréquence des organismes de l'échantillon de sol qui contiennent les gènes désirés. Pour augmenter cette fréquence, des étapes d'enrichissement en microorganismes hébergeant les caractéristiques désirées ont été utilisées avant la construction de certaine banque (Gabor *et al.*, 2004 ; Voget *et al.*, 2003). Dans la plupart des études, des sources de carbone ou d'azote sélectives pour l'espèce microbienne contenant les gènes d'intérêts ont été utilisées comme substrats de croissance. Un inconvénient de l'étape d'enrichissement est la perte de diversité microbienne, puisque les individus à croissance rapide et cultivables du consortium microbien sont d'habitude choisis. Néanmoins, une combinaison d'enrichissement

traditionnel et de technologie métagénomique est un moyen efficace pour augmenter la quantité de clones positifs dans un crible et pour isoler de nouvelles biomolécules quand des échantillons d'habitats complexes tel que le sol sont utilisés comme matériel de départ (Daniel, 2004).

Une autre méthode, la SIP (stable isotope probing), a été utilisée pour enrichir les génomes des individus métaboliquement actifs de la communauté microbienne du sol avant la construction de banques (Radajewski *et al.*, 2003 ; Wellington *et al.*, 2003). Cette technique a permis d'identifier une nouvelle methane monooxygenase à partir d'une banque métagénomique de sol forestier (Dumont *et al.*, 2006).

3.3.4.5 Optimisation de la métagénomique du sol

Les méthodes bioinformatiques qui permettent des comparaisons statistiques de banques sont nécessaires pour déterminer si les différences entre banques sont des artefacts d'échantillonnage et de construction de banques ou sont causées par des changements de la composition de la communauté.

Les différentes méthodes de criblage de banque du sol ont fourni un aperçu de la diversité des communautés microbiennes et ont permis d'identifier de nouvelles biomolécules, mais ces approches ont des forces et des faiblesses. Pour prendre en compte l'énorme diversité de microorganismes du sol, une combinaison d'approches basées sur la séquence et sur la fonction et différents types de banques devrait être utilisée pour explorer le métagénome du sol.

Une troisième stratégie de criblage à haut débit, qui est basée sur l'expression de gènes clonés induite par le substrat (SIGEX) a été présentée pour l'identification et la récupération de gènes qui codent des voies cataboliques (Uchiyama *et al.*, 2005). Cette méthode repose sur le clonage aléatoire d'ADN metagénomique en amont du gène *gfp* (codant une protéine fluorescente verte), plaçant ainsi l'expression de ce gène sous le contrôle de promoteurs présents dans l'ADN métagénomique. Les clones influençant l'expression de la gfp par addition du substrat d'intérêt peuvent être isolés par triage des cellules fluorescentes.

3.3.4.6 Problèmes liés aux eucaryotes

Pour différentes raisons, l'approche métagénomique développée pour les Procaryotes qui est basée sur le clonage de grands fragments d'ADN génomique n'apparaît pas la plus

adaptée à l'étude des génomes eucaryotes. En effet, leurs tailles sont très supérieures à celles des génomes de bactéries et d'archaea et leur densité en gènes est bien plus faible. Ceci rend nécessaire l'étude d'un nombre important de clones pour assurer une bonne représentativité des différents organismes eucaryotes au sein de banques de gènes. De plus, la présence d'introns dans de nombreux gènes eucaryotes et la non-conservation de la machinerie de transcription et des modes de maturation des protéines (glycosylation, sécrétion) entre Eucaryotes et Procaryotes n'autorise pas un criblage direct (phénotypique) des clones bactériens recombinants pour la production de métabolites ou d'enzymes. Enfin, seule une fraction des gènes est exprimée de façon constitutive.

3.3.5 L'approche métatranscriptomique

Tandis que les disciplines de génomique et de métagenomique étudient, respectivement, le potentiel génomique d'un organisme particulier ou d'une communauté microbienne, la transcriptomique et la métatranscriptomique s'intéressent au sous-ensemble de gènes qui sont transcrits dans certaines conditions environnementales. En conséquence, la transcriptomique et la métatranscriptomique sont des outils puissants pour capturer instantanément les gènes essentiels pour la survie dans des niches écologiques particulières.

De nombreuses études de profils d'expression ont été effectuées au cours des dernières décennies pour aborder des questions centrales en microbiologie, fournissant un aperçu précieux dans la compréhension de la corrélation entre certains phénotypes, comme la résistance aux radiations, la pathogénie ou la résistance aux chocs thermiques, et l'expression de gènes (Qiu *et al.*, 2008 ; Liu *et al.*, 2003 ; Audia *et al.*, 2008).

L'intérêt des chercheurs pour les gènes transcrits sous des conditions environnementales spécifiques les a conduits à faire des efforts substantiels pour développer des méthodes d'analyses. Des techniques d'extraction directes d'ARNm de bactéries, d'archaea et d'eucaryotes à partir d'échantillons environnementaux ont été récemment publiées (Bailly *et al.*, 2007; Poretsky *et al.*, 2005). Dans les deux études, des banques d'ADNc générés à partir d'ARN environnementaux ont été construites et 119 et 400 clones ont été respectivement séquencés. La plupart des séquences obtenues n'avaient aucune similarité avec n'importe quelle séquence de protéine précédemment déposée dans les bases de données publiques. Ainsi, ces études démontrent le potentiel de découvrir de nouvelles protéines. En outre, une capacité moindre de séquençage est exigée pour l'analyse des transcrits en comparaison avec les analyses d'ADN génomique ou métagénomique. Ceci est particulièrement vrai pour la

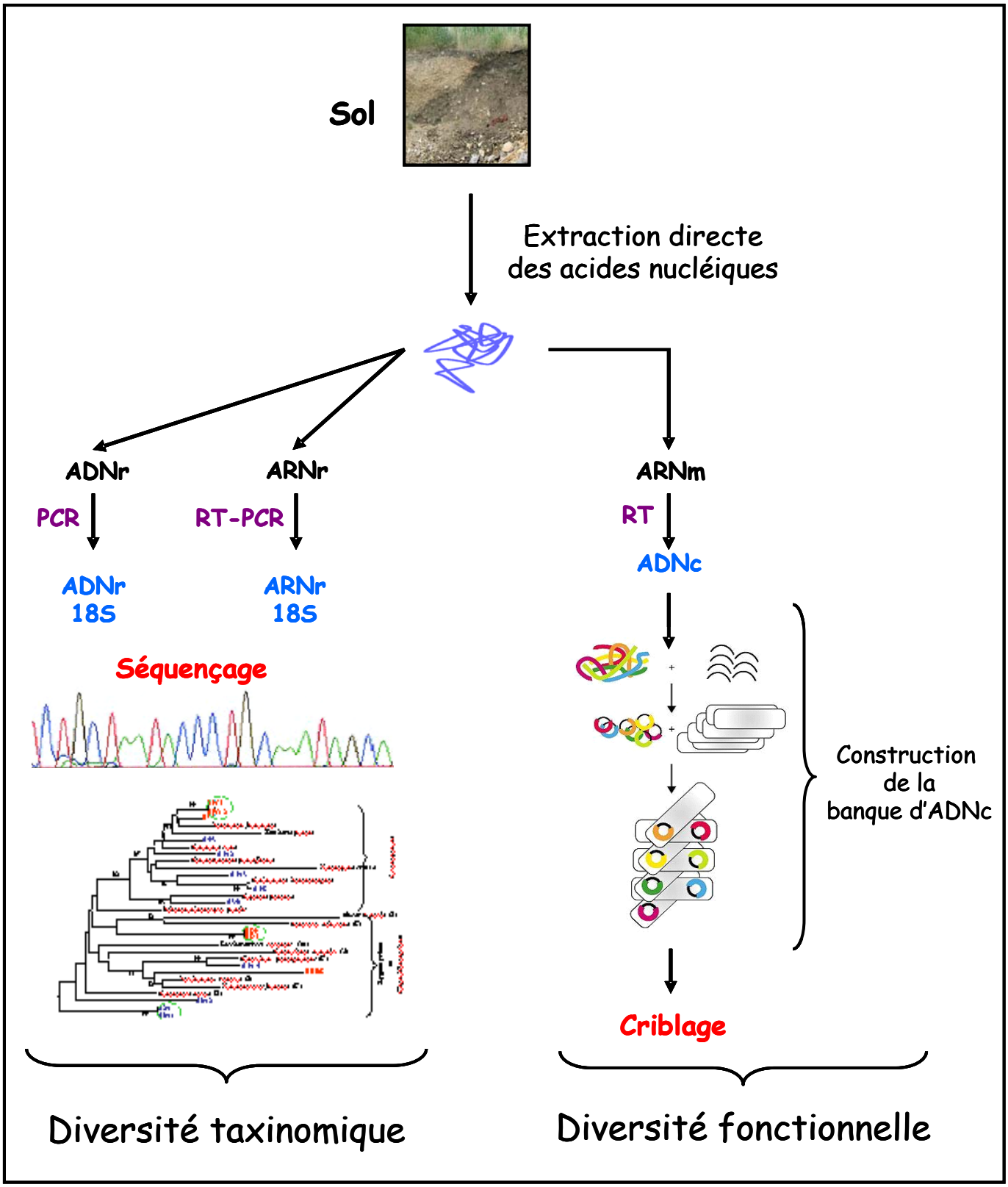


Figure 13: Les différentes étapes de l'approche métatranscriptomique.

faible densité de séquences codantes contenue dans les génomes eucaryotes. Cependant, un effort de séquençage important est effectué afin d'être quantitatif et d'identifier des transcrits rares.

3.3.5.1 Synthèse d'ADNc et construction de banques métatranscriptomiques

La préparation des ADNc pour le séquençage implique essentiellement trois étapes après avoir récupéré l'échantillon environnemental: l'extraction des ARN totaux, l'élimination des ARNr et donc l'enrichissement en ARNm et la synthèse d'ADNc.

Comme avec l'approche métagénomique, le choix de l'échantillon est crucial pour la découverte d'enzymes. Idéalement, l'habitat à partir duquel l'échantillon a été récupéré doit représenter un enrichissement naturel en organismes effectuant l'activité d'intérêt. Des précautions supplémentaires sont à prendre en ce qui concerne les ARN. Ces derniers sont notamment très sensibles à la dégradation par des ribonucléases ubiquistes. Quelques transcrits ont un temps de vie de moins d'une minute (Poretsky *et al.*, 2005). En conséquence, les échantillons doivent être immédiatement congelés ou stockés dans un tampon préservant l'ARN. Les méthodes de lyse directe de Hurt et al. (2001) ou de Bailly et al. (2007) qui permettent le recouvrement simultané de l'ADN et de l'ARN à partir de sols de compositions différentes sont bien adaptées à ce type d'approche.

Après l'extraction de l'ARN, l'enrichissement en ARNm est souhaitable puisqu'ils ne représentent qu'un faible pourcentage de l'ARN total contenu dans une cellule (de 1 à 5% chez les bactéries ; McGrath *et al.*, 2008). Plusieurs stratégies d'enrichissement en ARNm et d'amplification des transcrits existent (Poretsky *et al.*, 2005 ; Bailly *et al.*, 2007 ; Gilbert *et al.*, 2008 ; Frias-Lopez *et al.*, 2008). Les ARNm eucaryotes peuvent être spécifiquement sélectionnés car ils sont poly-adénylés à leur extrémité 3' (Grant *et al.*, 2006; Bailly *et al.*, 2007).

Les ADNc sont ensuite synthétisés par transcription réverse et, suivant le criblage effectué, liés dans un vecteur (plasmide) et clonés dans un hôte bactérien (*E. coli*) (Grant *et al.*, 2006 ; Bailly *et al.*, 2007 ; McGrath *et al.*, 2008) (Fig. 13).

Plateforme	Million de pb/ run	Coût/ base ($ US)	Longueur moyenne lue (pb)
Sanger (ABI 3730xl)	0,07	0,1	700
Pyroséquençage 454-Roche (GS-FLX-Titanium)	400	0,003	400
Séquençage Illumina (GAII)	2.000	0,0007	35
Séquençage ABI SOLiD3	20.000	n.d	35

Tableau 9: Résumé des technologies de séquençage disponibles actuellement (d'après Hugenholtz et Tyson, 2008).

3.3.5.2 Criblage des banques métatranscriptomiques

Une façon de cribler des banques métatranscriptomiques consiste à utiliser la PCR ou l'hybridation ciblée, mais comme pour l'approche métagénomique leur application est limitée à l'identification de gènes connus ou de nouveaux membres de familles de gènes connues.

Une approche originale, réalisée pour la première fois par Bailly *et al*. en 2007, consiste à exprimer des ADNc environnementaux dans un hôte hétérologue eucaryote. Ces ADNc ont préalablement été liés dans un vecteur plasmidique binaire. Une complémentation fonctionnelle d'un mutant de levure (*Saccharomyces cerevisiae*) auxotrophe pour l'histidine avec des EST (Expressed Sequence Tag) environnementales d'origine fongique a ainsi pu être réalisée.

Une approche complémentaire consiste à effectuer un séquençage aléatoire massif des banques générées. Des efforts de séquençage et d'assemblage moins conséquents, par rapport à l'approche métagénomique, sont nécessaires pour recouvrir l'information contenue dans ces banques d'ADNc.

Un progrès considérable a été réalisé pour l'analyse efficace des profils d'expression plus complexes avec le développement de technologies de séquençage de nouvelle génération (par exemple, 454 de Roche, SOLEXA de Illumina et SOLiD3 de ABI). Ces nouvelles technologies ne permettent pas seulement le séquençage direct d'ADN ou même d'ADNc sans aucune étape de clonage (Medini *et al.*, 2008), mais aussi d'augmenter le débit du nombre de paires de base séquencées par « run » et de diminuer le coût par base séquencée (Tableau 9). Le premier rapport d'utilisation du pyroséquençage 454 pour étudier le métatranscriptome d'une communauté microbienne complexe a été publié en 2006 (Leininger *et al.*, 2006). En utilisant cette technologie de séquençage Leininger *et al.*, (2006) ont montré que les transcrits d'archaea codant l'enzyme clef de l'oxydation de l'ammoniaque (*amoA*) étaient plus abondants dans les sols que ceux de la version bactérienne, suggérant ainsi que les archaea sont dominants numériquement dans le sol pour oxyder l'ammoniaque. Les 25 Mb de données de séquences obtenues à partir de ce transcriptome ont ensuite été analysées par Urich *et al.* (2008). Ils ont annoncé que 8 % des transcrits supérieurs à 250 pb pourraient être identifiés comme des ARNm. Cette étude a démontré comment des technologies de séquençage à haut débit peuvent être appliquées facilement pour avoir accès aux informations stockées dans des transcrits connues et inconnues qui ont été isolées directement d'environnements complexes comme le sol.

3.4 La diversité microbienne dans d'autres écosystèmes

Les approches moléculaires et plus particulièrement l'approche métagénomique et métaranscriptomique ont également été utilisées pour étudier la diversité microbienne des milieux aquatiques et de l'appareil gastro-intestinal des vertébrés et des invertébrés (Lopéz-Garcia et Moreira, 2008).

3.4.1 L'intestin d'insecte

Plusieurs études de métagénomique ont été effectuées sur l'intestin postérieur de termites « supérieures » s'alimentant de bois. Les termites sont connues comme des organismes économiquement importants pour dégrader le bois et ayant des rôles environnementaux essentiels dans le recyclage du carbone et comme sources éventuelles de catalyseurs biochimiques qui peuvent être utilisés dans la conversion du bois en biocarburants (Warnecke *et al.*, 2007). Des données significatives ont récemment suggéré que les bactéries symbiotiques résidant dans l'intestin postérieur des termites jouent un rôle fonctionnel dans l'hydrolyse de la cellulose et du xylane (Tokuda et Watanabe, 2007). Pour mieux apprécier la grande diversité des mécanismes biologiques responsables de la dégradation de la lignocellulose, une analyse métagénomique du microbiota de l'intestin postérieur de l'espèce consommatrice de bois *Nasutitermes* a été effectuée afin d'identifier le jeu complexe de gènes bactériens généralement utilisés pour l'hydrolyse de la cellulose et du xylane (Warnecke *et al.*, 2007). L'ADN de la communauté microbienne contenue dans l'intestin postérieur a été extrait, cloné et séquencé. Un total de 1750 gènes d'ARNr bactériens a été amplifié par PCR et l'identification a permis de répartir les bactéries dans 12 phyla et 216 phylotypes. Après PCR, le genre *Treponema* comprenant le taxon *Fibrobacter* est le plus fréquemment récupéré, avec 68 % de gènes séquencés dont 13% appartiennent aux phylotypes *Fibrobacters*. Une analyse conservatrice basée sur des techniques d'alignement global a permis d'identifier plus de 100 modules de gène analogues aux domaines catalytiques des glycosides hydrolases. Cette étude a aussi illustré d'autres fonctions potentiellement importantes de la population microbienne de l'intestin postérieur des termites « supérieures » s'alimentant de bois, comme le métabolisme de l'hydrogène, l'acétogénèse réductrice du dioxyde de carbone et la fixation d'azote (Warnecke *et al.*, 2007).

3.4.2 L'intestin de l'homme

Les communautés microbiennes occupent toutes les surfaces du corps humain avec un nombre de cellules microbienne total environ 10 fois supérieure à celui de cellules humaines (Kurokawa *et al.*, 2007). Le colon distal a été identifié comme l'écosystème bactérien naturel le plus peuplé, englobant plus de cellules bactériennes que toutes nos communautés microbiennes combinées (Frank et Pace, 2008). Il est alors évident que le microbiota gastro-intestinal humain est essentiel; il confère des fonctions métaboliques qui sont absentes chez l'hôte, comme les stratégies pour améliorer la récolte d'énergie des produits alimentaires ingérés, la synthèse de vitamines essentielles et la dégradation de polysaccharides complexes de plantes (Kurokawa *et al.*, 2007). En effet il n'est pas rare que des déséquilibres de la structure de la communauté microbienne intestinale soient induits et/ou causent des maladies comme la maladie de Crohn (inflammation de l'intestin), des allergies, l'obésité et des cancers (Kurokawa *et al.*, 2007).

Le premier effort d'exploration du métagénome de l'intestin humain a été entrepris en 2005 par les membres du laboratoire Relman et l'Institut pour la Recherche Génomique (TIGR) dans le but de découvrir la diversité de la microflore gastro-intestinale (Eckburg *et al.*, 2005). Afin d'identifier les caractéristiques génomiques communes à tous les microbiomes d'intestins humains, Kurokawa *et al*. (2007) ont récemment effectué une analyse métagénomique comparative à partir d'échantillons fécaux de 13 individus sains d'âges divers, y compris des enfants en bas âge non sevrés. L'analyse du métagénome de l'intestin humain s'inscrit dans la continuité du projet de séquençage du génome humain. Beaucoup de résultats sont prévus pour le Projet de Microbiome Humain (HMP), y compris l'identification de nouveaux biomarqueurs pour la santé et la médecine, de nouvelles enzymes capables de dégrader les xénobiotiques, et en fin de compte une compréhension plus complète des besoins nutritionnelles de l'homme (Turnbaugh *et al.*, 2007).

3.4.3 Les milieux aquatiques

De nombreuses études moléculaires basées sur la petite sous unité de l'ARNr 18S ont permis de révéler de nouvelles unités taxinomiques eucaryotes fonctionnelles (Moon-van der Staay *et al.*, 2001 ; Lopez-Garcia *et al.*, 2001 ; Dawson et Pace, 2002 ; Stoeck *et al.*, 2006 ; Stoeck *et al.*, 2003 ; Habura *et al.*, 2004a ; Johnson *et al.*, 2004). Certaines représentent une

nouvelle diversité dans des groupes déjà connus alors que d'autres ne semblent reliés à aucune lignée déjà décrite.

Le premier projet de séquençage à grande échelle (ou séquençage massif aléatoire ou random shotgun sequencing) a été effectué par l'Institut Craig J. Venter en 2004 dans lequel les fragments d'ADN séquencés ont été récupérés à partir de la communauté microbienne de la mer des Sargasses, une région intensivement étudiée de l'Océan Atlantique près des Bermudes et limitée en substance nutritive (Venter *et al.*, 2004). Un séquençage aléatoire de plus de 1,6 milliards de paires de base d'ADN a mené à la découverte de 1,2 millions de nouveaux gènes. 794.061 gènes ont été assignés comme étant des protéines hypothétiques conservées dont les fonctions sont inconnues. La mer des Sargasses a été choisie pour l'analyse métagénomique parce qu'il a été supposé qu'elle possédait une communauté relativement simple. Cela s'est avéré ne pas être le cas. En effet, l'analyse a révélé la présence de 1800 espèces bactériennes, dont 148 appartiennent à de nouveaux phylotypes. La communauté n'était pas assez simple pour permettre d'assembler les génomes microbiens.

Tyson et al. (2004) ont pour leur part choisi d'étudier une communauté beaucoup plus simple, celle du système d'écoulement d'eau acide d'une mine du Richmond (La Montagne de Fer, Californie), un des environnements les plus extrêmes sur Terre. Dans cet environnement le microbiota existe sous forme de biofilms roses qui se forment sur la surface de l'eau de la mine. Le biofilm a un pH de 0,83, une température de 43°C et contient de hautes concentrations de fer, de zinc et de cuivre (Tyson *et al.*, 2004). Une approche de séquençage massif aléatoire a révélé 384 gènes d'ARNr 16S dont les extrémités 5' et 3' ont été séquencées (Baker *et al.*, 2003). La simplicité de la structure de la communauté a permis à Tyson *et al.* (2004) de séquencer presque toute la microflore. L'analyse métagénomique de la communauté AMD a abouti à la reconstruction presque complète des génomes de *Leptospirillum* du groupe II et de *Ferroplasma* de type II.

Quelques années auparavant, Béjà *et al.* (2001) avaient déjà réalisé une étude de l'environnement marin par construction d'une banque métagénomique dans un vecteur fosmidique. Des études phylogénétiques des gènes d'ARNr 16S et un séquençage complet des fosmides ont été réalisés. De cette façon, ils ont découvert l'existence d'une nouvelle fonction dans une bactérie non cultivée. En effet, un clone de la banque possédait le gène d'ARNr 16S de SAR86 (un groupe de séquences retrouvées dans beaucoup d'océans, mais sans représentant en culture axénique) et un gène codant pour une protéine semblable à l'halorhodopsine a été trouvé dans le même clone. Une étude approfondie a montré que cette protéine (nommé protéorhodopsine) utilise la lumière pour produire un gradient de proton à

travers la membrane cellulaire. Ainsi, il a été montré qu'un groupe inconnu de bactéries, SAR86, possédait une nouvelle fonction (la phototrophie) (Béjà *et al.*, 2001).

En 2008, Friaz-Lopez *et al.* ont produit plus de 50 Mb d'ADNc par pyroséquençage 454 (Frias-Lopez *et al.*, 2008). Gilbert *et al.* suivirent peu après avec plus de 300 Mb de données de séquences en utilisant la deuxième génération de pyroséquençage appelée GS-FLX (Gilbert *et al.*, 2008).

Finalement, si des banques métagénomiques ou métatranscriptomiques de deux ou plusieurs communautés différentes sont disponibles, elles peuvent être utilisées pour comparer l'abondance relative des gènes avec des fonctions données et relier cela aux conditions particulières de chaque environnement. Ceci a été réalisé par Tringe *et al.* (2005), qui ont comparé les banques de la Mer des Sargasses à celle de carcasses de baleines reposant à plus de 500m de profondeur dans deux océans différents et à celle d'un sol agricole du Minnesota. Cette comparaison a immédiatement suggéré des hypothèses écologiques intéressantes concernant le chimiotactisme (Tringe *et al.*, 2005). Récemment, Poretsky *et al*. (2009) ont pu obtenir des informations détaillées sur les réponses métaboliques et biogéochimiques d'une communauté microbienne à la contrainte solaire en comparant des données métatranscriptomiques d'eau de surface du Pacifique prélevée le jour et la nuit (Poretsky *et al.*, 2009).

3.5 Conclusion

Il est clair que le petit nombre d'environnements jusqu'ici étudié avec des méthodes moléculaires ne permet qu'une compréhension rudimentaire du monde microbien naturel. Cependant, les méthodes basées sur les séquences fournissent maintenant une façon d'examiner la biodiversité rapidement et sous tous les aspects. Si nous voulons comprendre le monde vivant qui nous entoure (la biosphère), il est important, même essentiel, d'entreprendre une étude représentative de la diversité microbienne dans l'environnement. Une classification complète du biotope microbien de la Terre est inutile et, bien sûr, impossible. Néanmoins, un aperçu représentatif pourrait être réalisé avec un effort modeste au vue des nouvelles technologies de séquençage automatisé. L'analyse de 1000 clones (pour détecter les types de génome les plus abondants) de 100 environnements chimiquement différents serait comparable à un effort de séquençage de l'ordre d'un simple génome microbien. Dernièrement, un consortium international, Terragenome (Vogel *et al.*, 2009), fédère les

efforts de recherche afin d'aboutir au séquençage complet du génome de tous les microorganismes du sol.

Les questions sont importantes et nombreuses: avec quels types d'organismes partageons-nous cette planète et en quoi dépendons-nous d'eux? Quels modèles devrions-nous choisir pour les études des processus environnementaux au laboratoire? Comment pouvons-nous tirer des informations et des ressources de ce vaste fond de biodiversité? Pouvons-nous utiliser la distribution des microbes pour cartographier et contrôler la chimie de la Planète? Y a t-il des embranchements plus profonds dans l'arbre de la vie que les lignées que nous connaissons? (Pace, 1997).

Les opportunités de découverte de nouveaux organismes et le développement de ressources basées sur la diversité microbienne sont plus grandes que jamais auparavant. Les séquences moléculaires ont finalement donné aux microbiologistes une façon de définir leurs sujets, par la phylogénie moléculaire. Les séquences sont aussi la base des outils qui permettront aux microbiologistes d'explorer la distribution et les rôles des organismes dans l'environnement. La microbiologie peut maintenant être une science entière; l'organisme peut être étudié dans l'écosystème (Pace, 1997).

4. Impact d'une pollution métallique sur les organismes eucaryotes du sol

4.1 Les métaux lourds

Les métaux sont définis comme « des éléments qui conduisent l'électricité, avec un aspect métallique, malléable et formant des cations et des oxydes basiques » (Atkins et Jones, 1997). Usuellement, les termes 'métal' et 'métal lourd' sont utilisés en référence à l'élément pur et à toutes les formes chimiques (spéciation) dans lesquelles l'élément peut exister (Duffus, 2002). Le terme 'métal lourd' fait référence aux éléments métalliques ou dans certains cas aux métalloïdes caractérisés par une grande masse atomique (>100) ou une densité supérieure à $5 g.cm^{-3}$ (Adriano, 1986) et qui sont fréquemment associés à une toxicité ou une pollution. Cette définition correspond à 54 éléments du tableau périodique, mais ils n'ont pas tous une importance biologique. En se basant sur leur solubilité sous conditions physiologiques, 17 de ces métaux peuvent être disponibles pour les cellules vivantes et ont

une influence sur les plantes et les écosystèmes en général. Parmi ces métaux, le fer, le zinc, le cuivre, le molybdène et le manganèse ont une grande ou une faible importance comme éléments traces. Quelques éléments, comme le cadmium, n'ont pas de fonctions physiologiques connues et sont toxiques pour les plantes et les microorganismes (Schutzendubel et Polle, 2002)

Le terme 'métal lourd' est très approximatif (basé sur la densité, la masse atomique, le numéro atomique ou d'autres propriétés chimiques) et est souvent mal employé comme un synonyme d'espèce métallique toxique et nuisible, supposant que les termes 'lourd' et 'toxique' sont identiques. L'usage du terme métaux lourds a été sérieusement débattu à cause de certains métaux toxiques qui ne sont pas particulièrement 'lourds' (par exemple l'aluminium), et de quelques métaux lourds qui sont indispensables à la vie en faible quantité (Cu, Zn, Ni, Co…). D'autres éléments hautement toxiques ne sont même pas des métaux à strictement parler, mais des métalloïdes (Arsenic) ou des semi métaux (Sélénium). Pour ces raisons, le terme 'élément trace' a été proposé comme alternative (Fergusson, 1990). Cependant, nous utiliserons le terme 'métal lourd' car il est bien établi dans la littérature biologique et parce que le travail réalisé dans cette thèse porte sur le Zn et le Cd.

4.2 Effets au niveau cellulaire

La plupart des métaux sont des éléments de transition formant des cations avec une haute capacité à se lier aux atomes d'oxygène, d'azote et de soufre (Nieboer et Richardson, 1980) et en particulier aux résidus cystéines, inhibant les activités de nombreuses enzymes (Van-Assche et Clijsters, 1990). Les métaux lourds peuvent bloquer les groupes fonctionnels de molécules biologiques importantes ; déplacer et/ou replacer des ions essentielles dans les biomolécules ; ils peuvent induire un changement de la conformation, la dénaturation et l'inactivation d'enzymes et la disruption de tous les types de membranes cellulaires (Ochiai, 1987). Les métaux sont aussi impliqués dans la production d'espèces réactives de l'oxygène (ROS) (Schutzendubel et Polle, 2002), qui causent des dommages pour une grande gamme de biomolécules.

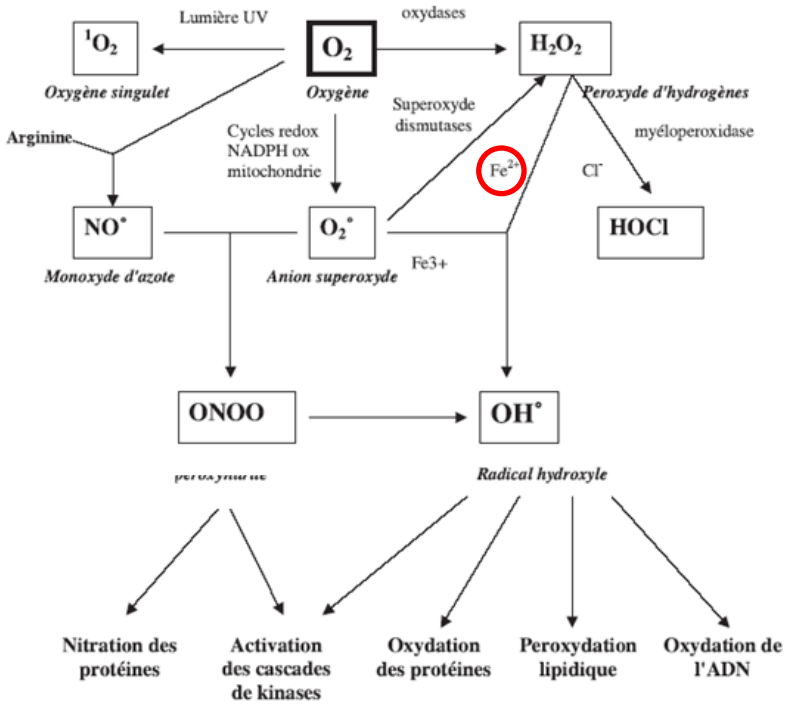


Figure 14: Origine des différents radicaux libres et espèces réactives de l'oxygène impliqués en biologie. En déplaçant le Fe^{2+} des protéines, le Cd amplifie la production de ces composés (d'après Favier, 2003).

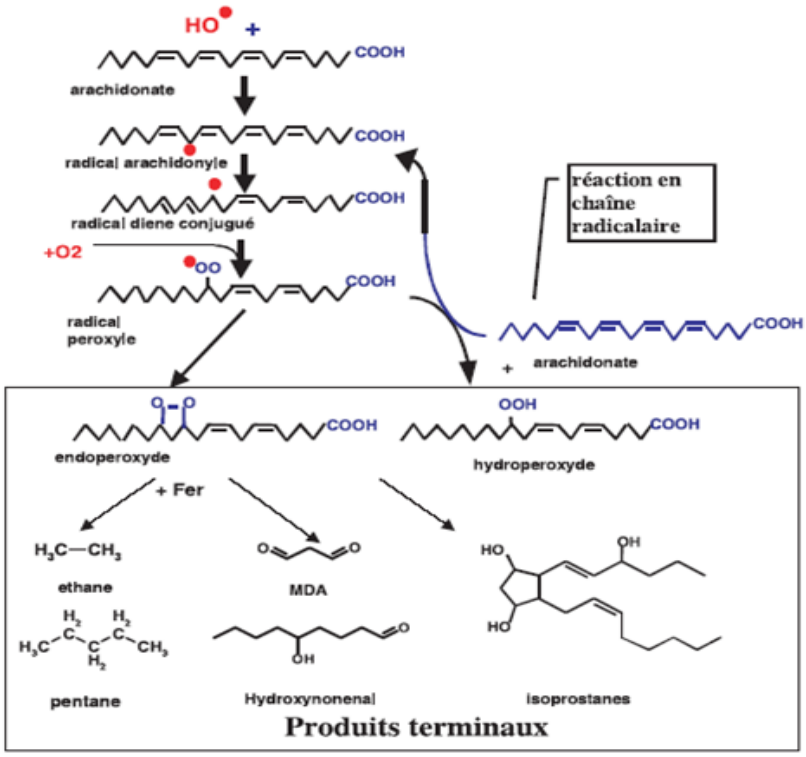


Figure 15: Mécanisme de la peroxydation des acides gras polyinsaturés et nature des produits terminaux formés (d'après Favier, 2003). MDA, malonedialdéhyde.

4.2.1 Altérations des membranes cellulaires

La condition préalable à la toxicité d'un métal est son contact avec les composants cellulaires. Il apparaît donc clairement que la membrane plasmique est le premier site d'action. Les membranes sont affectées de différentes manières. Le Cd inhibe de façon compétitive les ATPases dépendantes du magnésium en formant un complexe avec l'ATP, privant ainsi l'enzyme de son substrat. Cette inhibition peut affecter l'efflux d'ions H^+ ainsi que le potentiel transmembranaire. La fluidité et la composition des lipides membranaires sont également affectées, ce qui a des effets directs sur la perméabilité membranaire (Garcia *et al.*, 2005). Enfin, le Cd est susceptible de déplacer et de libérer les ions Fe^{2+} des protéines qui vont pouvoir générer, via des réactions de type Fenton, la formation d'espèces réactives de l'oxygène tel que le peroxyde d'hydrogène, les ions superoxydes et les radicaux hydroxyles (Stohs et Bagchi, 1995) (Fig. 14). Le radical hydroxyle ($OH^{\cdot}$) initie le processus de peroxydation membranaire. Les acides gras insaturés des membranes sont alors transformés en produits comportant des radicaux et des groupements hydroperoxydes qui, à leur tour, vont favoriser les réactions radicalaires au niveau d'autres constituants cellulaires (Fig. 15). Ces dommages cellulaires causés par les ROS sont regroupés sous le terme de stress oxydant (ou stress oxydatif).

4.2.2 Interactions avec les acides nucléiques

En général les métaux interagissent indirectement avec les acides nucléiques. D'une part, en générant des espèces réactives de l'oxygène qui vont induire des mésappariements entre les deux brins de l'ADN, provoquer des cassures de l'ADN simple brin, ainsi que leur dégradation (Waisberg *et al.*, 2003; Liu *et al.*, 2009) (Fig. 16). D'autres part, en agissant sur le fonctionnement d'enzymes impliquées dans le métabolisme des acides nucléiques par remplacement d'ions comme le Ca^{2+}, le Zn^{2+} et le Fe^{2+}, ils vont conduire indirectement à une altération de l'information génétique en influant sur la fidélité de la réplication (Kunkel et Loeb, 1979).

Ces exemples d'effets au niveau cellulaire ont été observés chez des cellules de mammifères (souris et homme) suite à une intoxication aigue au cadmium. D'autres conséquences biologiques ont également été observées sur ce type de cellules (Fig. 17). Ces effets sont en partie retrouvés chez d'autres organismes eucaryotes présents dans le sol, tels

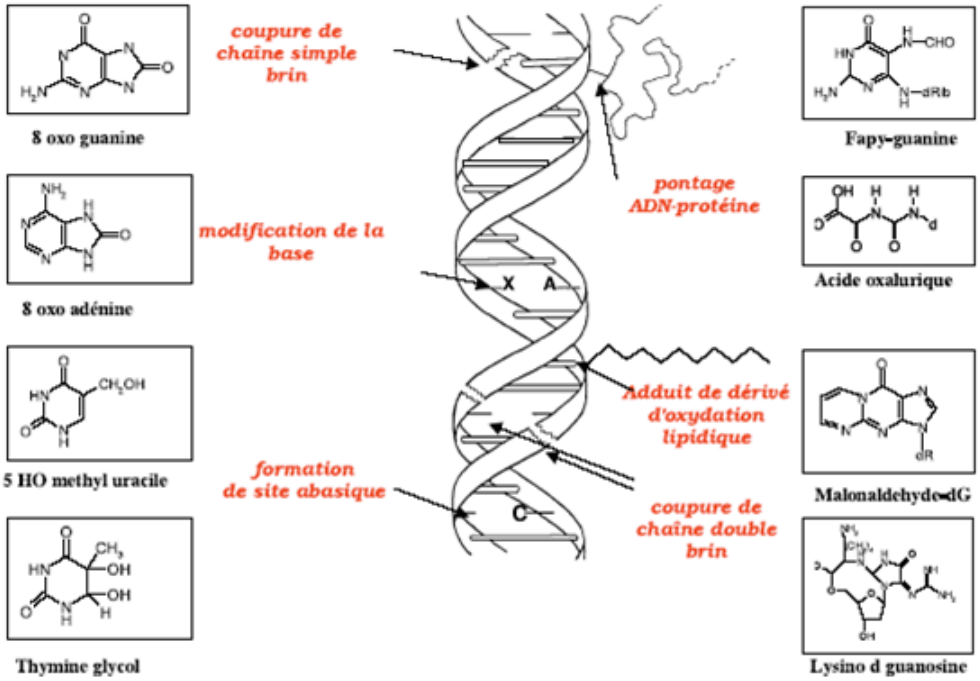


Figure 16: Lésions de l'ADN induites par attaque radicalaire. Les bases puriques et pyrimidiques sont modifiées par les radicaux libres (encadré) (d'après Favier, 2003).

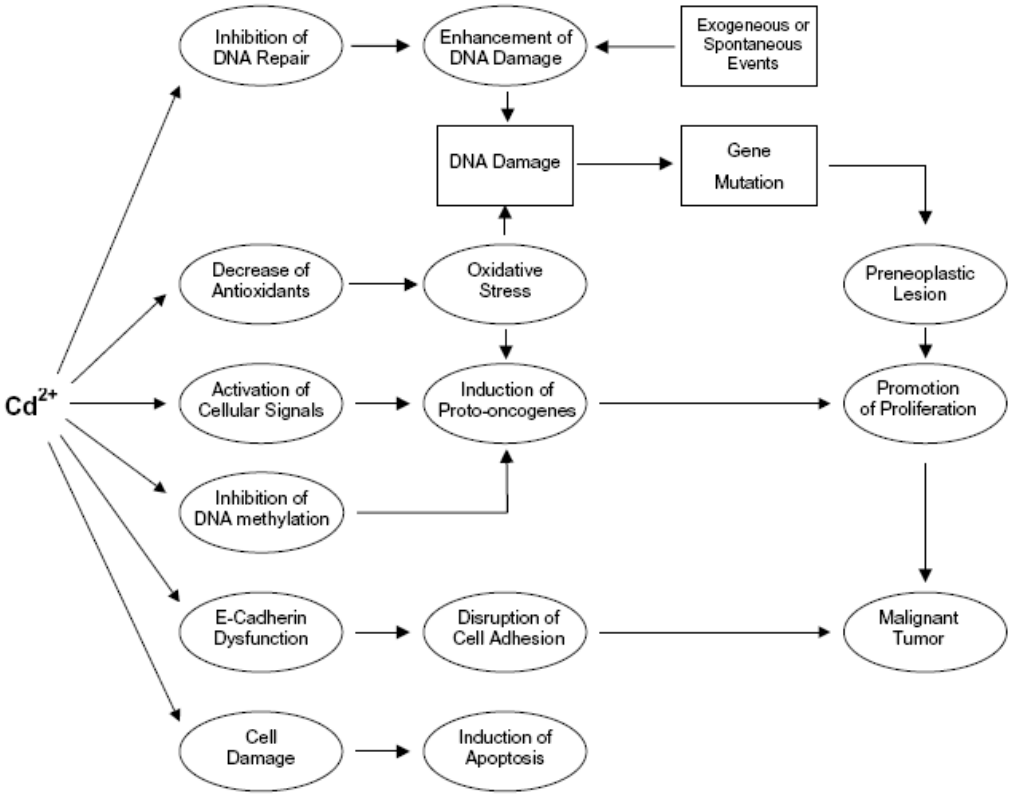


Figure 17: Résumé des effets causés par le cadmium chez les mammifères (d'après Waisberg *et al.*, 2003).

que les plantes, les champignons et les protistes (Deckert, 2005 ; Ruotolo *et al.*, 2008 ; Gallego *et al.*, 2007 ; Watanabe et Suzuki, 2002).

4.3 Toxicité et adaptation

L'extrême toxicité générée par une pollution métallique induit une forte pression de sélection, entraînant l'adaptation de certaines populations d'organismes tolérantes aux métaux. Ce phénomène a bien été décrit pour les procaryotes (Mergeay *et al.*, 2003) et les eucaryotes (Colpaert *et al.*, 2000; Blaudez *et al.*, 2000b).

4.3.1 Toxicité et biodisponibilité des métaux

La toxicité d'un métal dépend de sa biodisponibilité, c'est-à-dire de sa capacité à être transféré du compartiment sol à un organisme vivant sous forme ionique (Juste, 1988). La biodisponibilité est influencée par la concentration en métal, par les facteurs physico-chimiques du sol (pH, taux de matière organique, taux d'argile) et par des facteurs biologiques propres à chaque organisme vivant (capacité de biosorption, bioaccumulation, solubilisation). Il est donc important de mesurer la quantité biodisponible, extractible plutôt que la quantité totale de métaux présente dans le sol. Mais, la biodisponibilité reste rarement mesurée et rend les résultats difficilement comparables.

4.3.2 Toxicité et adaptation chez les plantes

Les premières études s'intéressant à des populations d'organismes eucaryotes tolérantes aux métaux ont été réalisées chez les plantes (Bradshaw et McNeilly, 1981 ; Al-Hiyali *et al.*, 1990). La capacité à survivre sur des sols métallifères n'est pas très répandue dans le règne des plantes (Antonovics *et al.*, 1971) et il est remarquable de constater que dans une large zone géographique les mêmes espèces de plantes développent des populations métallophytes, même quand différents métaux sont responsables de la toxicité. Ceci est bien connue pour quelques plantes monocotylédones, notamment chez les plantes herbacées (Schat *et al.*, 2000). Cette évolution s'expliquerait par la présence de 0,1 à 0,5% d'individus tolérants aux métaux dans les populations de plantes herbacées non adaptées (Al-Hiyali *et al.*, 1993),

probablement dû à un taux de reproduction élevé et à une production de nombreux descendants.

Les plantes ligneuses ne sont pas considérées comme des colonisateurs primaires des sols pollués aux métaux (Schat et al., 2000). Leurs longs cycles de reproduction, ne leurs permettraient pas d'avoir un potentiel adaptatif suffisamment fort pour la tolérance aux métaux (Meharg et Cairney, 2000). De plus, les arbres dépendent beaucoup plus de leurs champignons ectomycorhiziens (ECM) associés que les herbacées de leur symbiontes mycorhiziens à arbuscules, indépendamment de la pollution du sol. En effet, quand les champignons ectomycorhiziens se font rares, les plantes ligneuses colonisent plus lentement l'environnement (Nara, 2006a ; Nara, 2006b). Cependant, des espèces d'arbres à symbiose ectomycorhizienne, comme les bouleaux, les pins et les saules, sont capable de coloniser des sites fortement pollués par des métaux. Par conséquent, les arbres résistent à cette toxicité à cause de leur grande plasticité phénotypique et à travers leur association avec un partenaire fongique ectomycorhizien bien adapté (Wilkinson et Dickinson, 1995). De nombreux champignons (mycorhiziens ou pas) et bactéries de la rhizosphère tolérants aux métaux sont connus pour augmenter la « fitness » des plantes sur un sol contaminé aux métaux (Vivas *et al.*, 2006 ; Kozdroj *et al.*, 2007).

4.3.3 Toxicité et adaptation chez les champignons

Comme pour les plantes, les carpophores de certaines espèces de champignons sont de moins en moins abondants lorsque le degré de pollution augmente, alors que d'autres espèces ne sont pas affectées et l'abondance de leurs carpophores augmente (Rühling et Söderström, 1990). Des études récentes (Colpaert *et al.*, 2000) ont aussi montré des variations génétiques aussi bien aux niveaux inter- qu'intraspécifiques pour la tolérance des champignons ectomycorhiziens au Cd. Ces mêmes études ont montré que des isolats de *Suillus luteus* issus d'un sol contaminé au Cd étaient généralement plus résistants à ce métal que ceux issus d'un sol non contaminé. Néanmoins, chez *Paxillus involutus*, des isolats tolérants ont pu être prélevés à partir de sols non contaminés et aucune corrélation entre le niveau de tolérance et le degré de contamination du sol d'origine n'a pu être observée (Blaudez *et al.*, 2000b).

Les champignons sont capables d'accumuler du Cd dans leurs carpophores en quantité variable suivant l'espèce (jusqu'à 5.7 $\mu g.g^{-1}$ de poids sec pour *S. luteus*). Des études ont montré que le rapport de concentration en Cd dans la biomasse fongique divisée par celle du sol peut varier de 2 à 1000 suivant l'espèce (maximum pour *Amanita muscaria*) (Gast *et al.*,

1988). En condition axénique, où le métal est ajouté sous forme de sel soluble, ce rapport atteint 200 et 80 pour des isolats de *Suillus bovinus* non tolérants et tolérants au Cd, respectivement (Colpaert et Van-Assche, 1992). Dans ce cas, la tolérance semble donc se traduire par l'acquisition d'une capacité accrue à exclure le Cd de la cellule. Ainsi une forte capacité d'accumulation ne reflète pas nécessairement une forte tolérance.

Cependant, l'analyse des carpophores ne reflète pas toujours l'activité des mycéliums souterrains et d'autres indicateurs de l'activité fongique en présence de Cd sont utilisés, comme les mesures d'émission de CO_2 d'échantillons de sol reflétant l'activité biologique globale, ou le dosage de l'ergostérol (marqueur du groupe des champignons). Mais aucune variation de ces indicateurs en réponse à la présence de métaux n'a été observée (Barajas-Aceves *et al.*, 2002). Ces résultats s'expliqueraient par le remplacement des mycéliums d'espèces non tolérantes par ceux d'espèces tolérantes.

4.3.4 Toxicité et adaptation chez les protistes

Très peu d'études portent sur la toxicité des métaux sur les protistes dans les sols (Bowers *et al.*, 1997 ; Campbell *et al.*, 1997 ; Diaz *et al.*, 2006). Dans chaque cas, des espèces de ciliés du genre *Colpoda*, bien représentées dans les sols, ont été utilisés. Des études ont montré des variations intraspécifiques de sensibilité aux métaux (Xu *et al.*, 1997) et des variations suivant la localisation géographique. Les différences de tolérance aux polluants toxiques parmi des souches ou même des individus et des clones d'une même espèce sont des phénomènes bien connus (Forbes, 1998), et illustrent l'importance de l'interaction génotype-environnement dans les réponses aux toxiques. Une autre donnée importante concerne la diminution de la toxicité du Cd en présence de Zn en concentration faible ou modérée (Diaz *et al.*, 2006). Ceci a également été démontré chez l'espèce de protiste d'eau douce *Tetrahymena pyriformis* (Chapman et Dunlop, 1981). Dans tous les cas, un antagonisme entre le Zn et le Cd a été suggéré.

Toutefois, Diaz et al. (2006) ont démontré que les protistes du sol étaient plus résistants au Cd et au Zn que ceux d'autres habitats, probablement parce qu'ils sont adaptés à un habitat dans lequel la pollution métallique persiste. Les protistes sont donc des organismes actifs dans les sols pollués par des métaux et contribuent à leur fertilité. Ils sont en effet connus pour être des prédateurs directs de bactéries et de champignons et des proies de nématodes (Ekelund *et al.*, 2002), mais également capables d'augmenter la capture d'azote par les plantes et ainsi d'augmenter leur biomasse (Bonkowski, 2002).

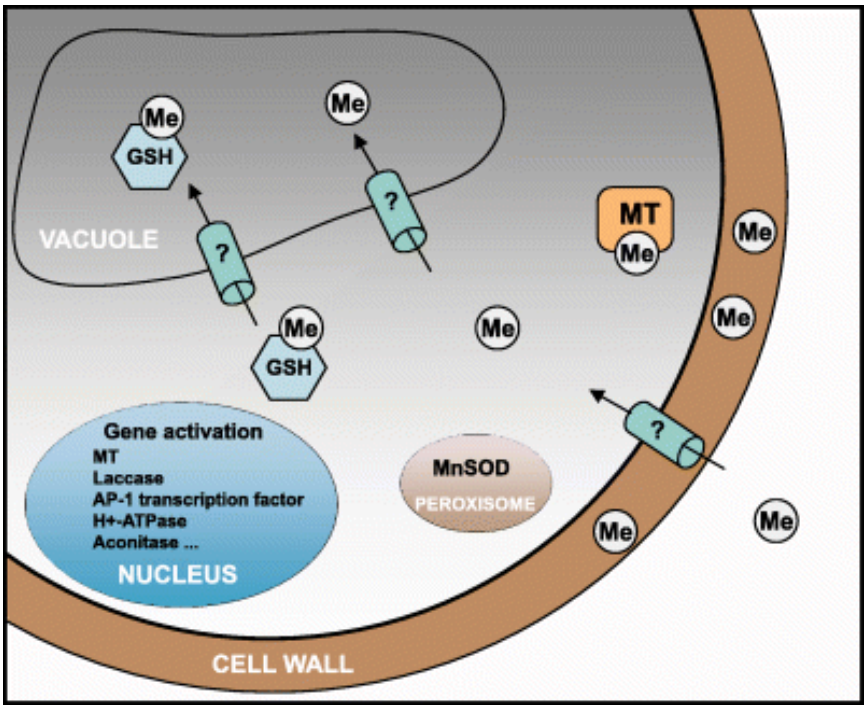


Figure 18: Mécanismes cellulaires et moléculaires potentiellement impliqués dans la tolérance métallique chez le champignon ectomycorhizien *Paxillus involutus* (Bellion *et al.*, 2006). Me, métal ; MT, métallothionéine ; GSH, glutathion ; MnSOD, superoxyde dismutase manganèse-dépendant (d'après Lanfranco, 2007).

La compréhension des interactions complexes entre les microorganismes et les plantes dans les sols pollués par des métaux lourds devraient augmenter le rendement de phytorémédiation des métaux.

4.4 Mécanismes microbiens eucaryotes de tolérance aux métaux

Les termes de «tolérance» et «résistance» sont utilisés de façon interchangeable et sans distinction claire dans la littérature. Ils sont souvent basés sur les capacités d'un organisme à croître sur un milieu de culture contenant une concentration biologiquement active en métal. De nombreuses études se rapportant au sujet, et détaillées ci-dessous, ont montré que seul un petit nombre de gènes majeurs impliqués dans ces mécanismes ont été mis en évidence bien que sans doute de nombreux autres gènes mineurs pourraient aussi être impliqués (Hall, 2002) Les mécanismes de tolérance ont été étudiés en détail chez des modèles tels que les levures *Saccharomyces cerevisiae* et *Schizosacchomyces pombe*, ainsi que chez la plante *Arabidopsis thaliana*. Récemment, la présence de mécanismes extra-et intracellulaires complexes a clairement été démontrée par une combinaison d'approches microbiologiques, biochimiques et de biologie moléculaire principalement appliquées à des systèmes modèles, comme *Paxillus involutus* (Blaudez *et al.*, 2000a), après une exposition aux ions de cadmium (Fig. 18). Ces mécanismes impliquent la liaison non spécifique aux parois cellulaires, des systèmes intracellulaires de chélation du métal, des réponses au stress oxydatif et la modulation d'expression de gènes (Bellion *et al.*, 2006). Quelques données sont également disponibles pour les protistes (Diaz *et al.*, 2006).

4.4.1 Complexation et précipitation extracellulaire

De nombreux composés extracellulaires peuvent se complexer ou précipiter avec les métaux lourds. Le mécanisme essentiel par lequel les plantes sont capables de tolérer l'aluminium implique l'excrétion racinaire d'acide organique tel que l'acide oxalique (Ma *et al.*, 2001). Les oxalates d'aluminium non toxiques sont ensuite accumulés dans les feuilles de la plante. Lors d'une exposition au Cd, la levure *S. cerevisiae* augmente la synthèse d'enzymes impliqués dans le métabolisme des carbohydrates (Vido *et al.*, 2001). Ce résultat

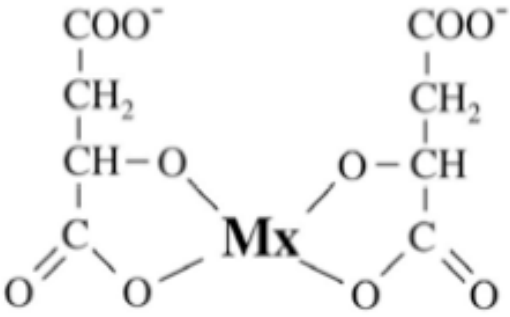


Figure 19: Structure du chélat d'un ion métallique avec des acides organiques ; exemple du malate.

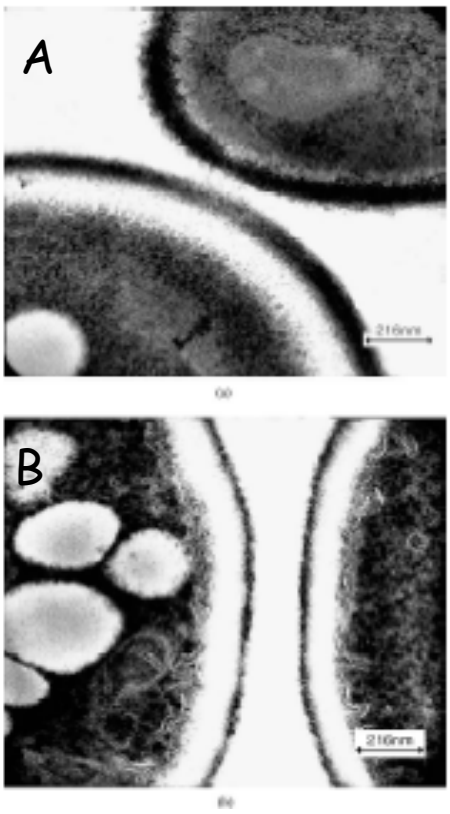


Figure 20: Influence de l'épaisseur de la paroi cellulaire sur l'adsorption du Cd. Coupes de parois cellulaires de 2 souches de *Saccharomyces cerevisiae* observées au microscope électronique à transmission. Une souche à paroi épaisse (A) adsorbe mieux le Cd qu'une souche à paroi fine (B) (d'après Park *et al.*, 2003)

suggère que l'activation des enzymes du métabolisme énergétique (glycolyse, cycle de Krebs) doit être nécessaire pour garantir la production et l'excrétion d'acides organiques (Fig. 19).

Les champignons ectomycorhiziens sont connus pour excréter des acides di- et tri-carboxyliques. Dans de nombreuses études, l'augmentation de l'efflux d'acide oxalique est corrélée à la tolérance aux métaux chez les champignons ECM (Fomina *et al.*, 2005 ; Ahonen-Jonnarth *et al.*, 2000 ; Cumming *et al.*, 2001). Cependant, ce mécanisme n'est pas retrouvé chez toutes les espèces d'ECM et est dépendant du métal (Meharg, 2003).

Par ailleurs, d'autres composés comme la glomaline, une protéine synthétisée et excrétée par un champignon mycorhizien à arbuscule (Gonzalez-Chavez *et al.*, 2004) est capable de séquestrer des ions métalliques, comme le Cu, le Pb et le Cd, présents en forte concentrations dans des sols pollués.

4.4.2 Fixation aux parois cellulaires

La paroi cellulaire est le premier site d'interaction entre les métaux et la cellule microbienne. Les interactions physico-chimiques (échange ionique, adsorption, complexation, précipitation, et cristallisation) responsables de l'association des espèces métalliques aux parois cellulaires portent le nom de biosorption.

La paroi des champignons est composée de polymères comprenant des glucanes, de la chitine et des polymères de galactosamines ainsi que quelques protéines. Elle possède donc un grand nombre de sites de liaison potentiels pour les métaux tels que des groupes carboxyles, amines, hydroxyles, phosphates et thiols.

Chez *S. cerevisiae*, la biosorption du Cd a pu être mise en évidence par microscopie électronique à transmission ou à balayage et analysée par spectroscopie à énergie dispersive. Certaines souches résistantes aux métaux présentent souvent un épaississement de leur paroi corrélé positivement à leur capacité d'adsorption du Cd (Park *et al.*, 2003) (Fig. 20).

Turnau *et al.*, (1994) ont révélé que la tolérance des champignons ectomycorhiziens aux métaux était associée à la formation de pigments pariétaux. Jacob *et al.*, (2004) ont montré une augmentation (de 3.9 fois) des transcrits de laccase (enzyme catalysant la synthèse de mélanine à partir de substrats phénoliques) chez *Paxillus involutus* lors d'une exposition au Cd. Cette augmentation favoriserait ainsi la séquestration du métal sur les pigments de la paroi.

4.4.3 Transport et homéostasie

Le phénomène de tolérance résultant d'une augmentation de la capacité d'adsorption des métaux n'est pas une caractéristique universelle. Certaines souches de levure résistantes au Cd peuvent aussi accumuler à l'intérieur de la cellule des quantités considérables de Cd probablement grâce à des mécanismes de séquestration intracellulaire très efficaces.
Un autre mécanisme de tolérance consiste aussi en l'efflux de cations métalliques vers le milieu extérieur. Des études menées chez les levures *S. cerevisiae* et *Candida albicans* ont montré que les transporteurs membranaires CAD2 et CaCRP1 appartenant à la famille de pompes CPx-ATPases (retrouvés chez de nombreuses bactéries) sont impliqués dans l'efflux de Cd et confèrent une résistance au Cd (Shiraishi *et al.*, 2000 ; Weissman *et al.*, 2000).

Dans le cas de *Paxillus involutus*, la diminution de l'accumulation et de la distribution du ^{109}Cd dans les différents compartiments intracellulaires à basse température suggère que le transport du Cd à travers les membranes est un processus actif (Blaudez *et al.*, 2000a). Dans cette étude, l'utilisation d'un protonophore, dépolarisant la membrane cellulaire, inhibe partiellement la capture du Cd. Par contre, l'inhibition de canaux K^+ ne diminue pas l'accumulation et la compartimentation du Cd dans le mycelium de *Paxillus involutus*, tout comme l'inhibition des H^+/ATPase. Ces résultats démontrent que le système de transport du Cd ne dépend pas du gradient de K^+ et exclut l'implication de H^+/ATPase dans l'efflux de protons. Enfin, l'inhibition de canaux Ca^{2+} diminue l'accumulation de Cd ; ces canaux pourraient donc jouer un rôle dans le transport du Cd à travers les membranes.

4.4.4 Les thiols cellulaires

Une fois entré dans la cellule, le Cd se lie à des composés essentiels des voies de détoxication présents chez de nombreux organismes (animaux, végétaux, microorganismes). Il s'agit de thiols cellulaires non protéiques, tel que le glutathion réduit (GSH), les phytochélatines (PCs) et des thiols cellulaires protéiques, tel que les métallothionéines (MTs).

4.4.4.1 Le glutathion réduit

Le GSH est le composé thiolé le plus abondant dans les cellules eucaryotes. Son potentiel rédox très bas (-240mV) fait de lui un excellent réducteur de radicaux libres. La nature nucléophile du groupement SH lui permet aussi de chélater les ions métalliques. Des

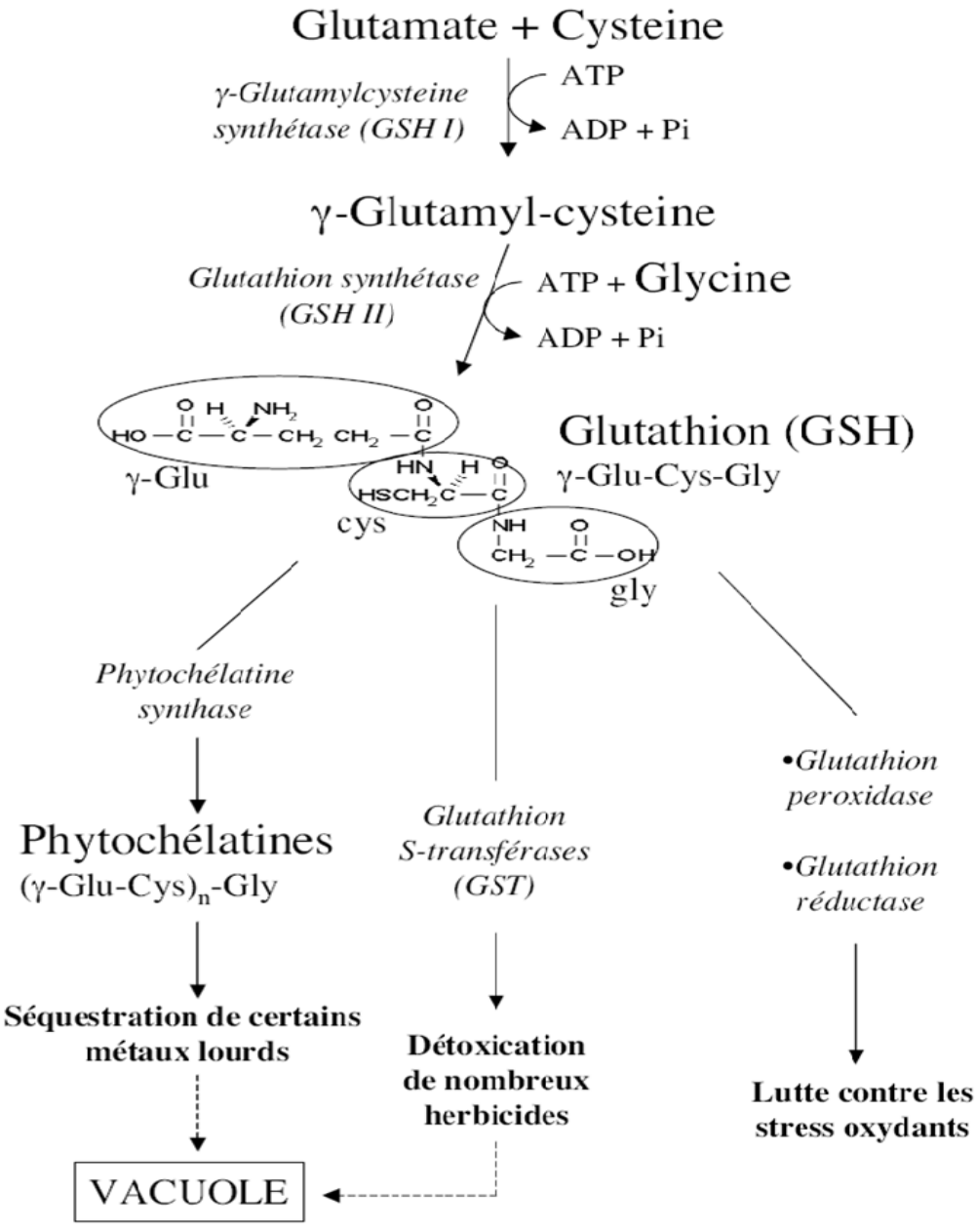


Figure 21: Vue générale du métabolisme du glutathion et ses rôles chez les organismes vivants.

études menées chez *S. cerevisiae* ont montré une relation entre l'exposition au Cd et l'augmentation de la transcription d'enzymes impliqués dans la voie d'assimilation des sulfates et dans la biosynthèse de la cystéine précurseur du glutathion (Vido *et al.*, 2001). Fauchon *et al.*, (2002) ont aussi démontré que des cellules de levure, en condition de stress au Cd, étaient capables de réguler l'utilisation des acides aminés soufrés (Cys et Met) en synthétisant et en utilisant préférentiellement des isoformes de certains enzymes de la glycolyse dont les séquences sont naturellement pauvres en ces deux acides aminés afin sans doute de les économiser pour la synthèse de GSH (Fig. 21).

Des études menées chez des champignons ectomycorhiziens comme *Paxillus involutus* et *Laccaria laccata* ont montré une augmentation de la concentration en glutathion et en ses précurseurs biosynthétiques (γ-GluCys) lors d'un stress au Cd (Courbot *et al.*, 2004). Toutefois, l'induction de l'expression de la cystéine synthase n'a pas été observée chez *P. involutus* lors d'un stress au Cd. Il s'agit d'un enzyme clé de la voie de biosynthèse de la cystéine dont une forte activité amplifie la synthèse de molécules séquestrant le Cd tels que le GSH ou les PCs. Par contre, une diminution de l'activité d'autres enzymes utilisant la cystéine (cystathionine synthase) et une diminution de la synthèse de protéines riches en cystéines (hydrophobines) ont été observées (Jacob *et al.*, 2004). Ce mécanisme alternatif permettrait de rediriger les cystéines vers la production de molécules piégeant le Cd.

4.4.4.2 Les phytochélatines

Les phytochélatines constituent une famille de petits peptides riches en cystéines capables de fixer les ions de métaux lourds via leurs groupes SH. La structure générale des PCs est $[\gamma\text{-GluCys}]_n\text{-Gly}$ (n = 2 à 5) (Fig. 22). Les PCs sont synthétisées de façon enzymatique, par la phytochélatine synthase, à partir de γ-glutamylcystéine qui est un composé intermédiaire de la voie de biosynthèse du GSH et qui possède des propriétés propres de chélation des métaux (Cruz-Vasquez *et al.*, 2002). Les PCs se rencontrent chez les plantes, les algues, quelques espèces de champignons ainsi que chez des invertébrés.

Les premières PCs fongiques ont été isolées de la levure *Schizosaccharomyces pombe* exposée à un stress au Cd (Murasugi *et al.*, 1981). Ces molécules n'existent pas chez *S. cerevisiae*. Une séquence similaire au gène *cad1*, codant la PC synthase chez *Arabidopsis*, a été trouvée dans le génome de *S. pombe* et la délétion de cette séquence conduit à une déficience en PCs et à une sensibilité au Cd (Murasugi *et al.*, 1981).

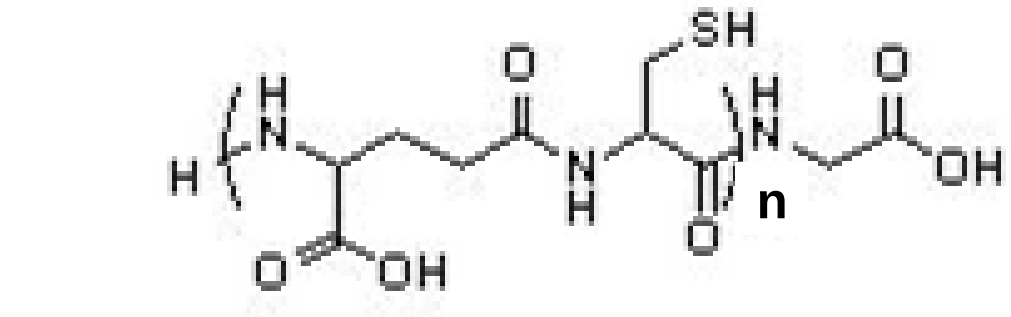


Figure 22: Structure primaire d'une phytochélatine (n = 2-5).

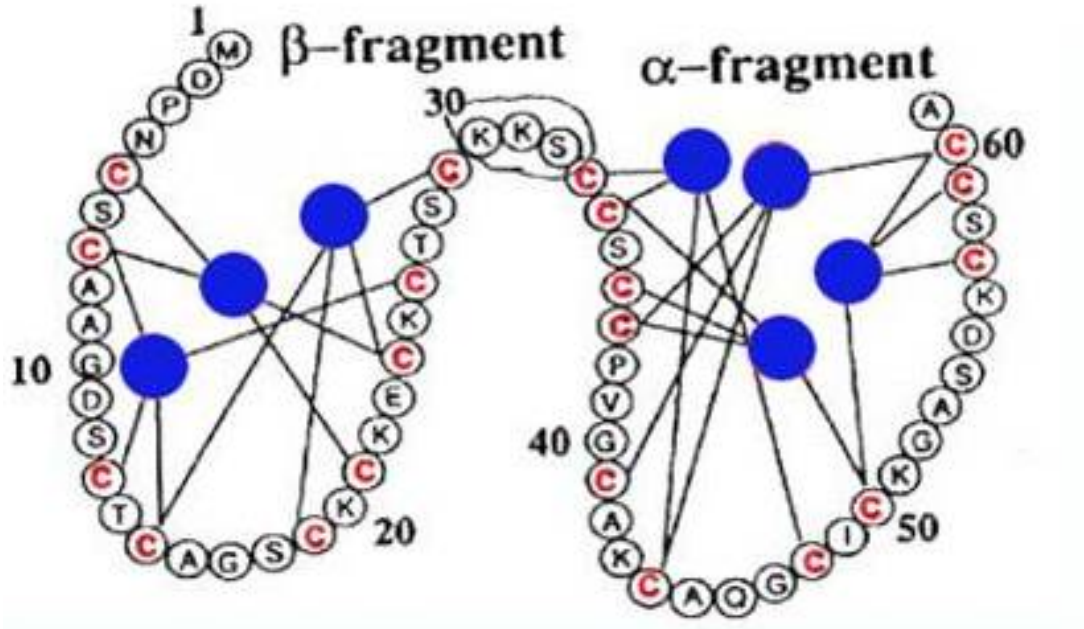


Figure 23: Vue shématique d'une métallothionéine animale. Les 2 domaines de la protéine (fragments alpha et bêta) sont capables de lier respectivement 4 et 3 atomes de Cd (en bleu) par coordination avec les groupements thiolates des cystéines (en rouge).

4.4.4.3 Les métallothionéines

Les métallothionéines quant à elles sont des peptides de faible masse moléculaire (25 à 60 acides aminés) issus de la traduction d'un ARN messager (Cobbett et Goldsbrough, 2002). Elles sont riches en cystéine et chélatent les ions métalliques par coordination avec les thiolates. Les métallothionéines ont été classées en différentes familles sur la base de leur séquence protéique et plus particulièrement suivant le motif dessiné par l'arrangement de leurs résidus cystéines (http:// www.expasy.org/cgi-bin/lists?Metallo.txt) (Fig. 23).

D'un point de vue évolution, les MTs ou les polypeptides ressemblant aux MTs ont été trouvés dans toutes les branches de l'arbre de vie, avec une remarquable conservation de la structure fonctionnelle à travers les phyla. Une hypothèse suggère que les structures des MT ancestrales ont évolué sous la pression sélective d'environnement dans lequel des métaux toxiques et des radicaux libres étaient particulièrement abondants (Coyle *et al.*, 2002). L'origine polyphylétique des MTs modernes s'expliquerait par leur spécialisation, au cours de l'évolution, en termes de fonction et de capacité de liaison aux métaux dans chaque forme de vie pour s'adapter aux différentes niches environnementales et/ou aux conditions métaboliques endogènes spécifiques. Deux groupes principaux de MTs ont été proposés sur la base de leur capacité de liaison aux métaux: les Cu-thionéines et les Zn-thionéines.

Chez *S. cerevisiae*, les deux gènes codant les MTs (*cup1* et *crs5*) ne sont exprimés qu'en présence de Cu, la présence de Cd n'induit pas leur synthèse (Culotta *et al.*, 1994, Vido *et al.*, 2001). Récemment, Crs5 a été caractérisé comme étant capable de lier le Zn mieux que le Cu (Pagani *et al.*, 2007). De même, un gène pouvant coder une MT, appelé *zym1*, a été cloné chez *Schizosaccharomyces pombe* et semble être exprimé en présence de Cu et de Zn mais pas de Cd (Borrelly *et al.*, 2002). Il semblerait que le rôle des MTs dans la détoxication du Cd serait d'une importance moindre comparé à celui des PCs. Toutefois, la surexpression de MTs dans *S. cerevisiae* lui confère bien une résistance au Cd (Kuroda et Ueda, 2006). De plus, un champignon, *Candida glabatra*, connu pour produire à la fois des PCs et des MTs pour la détoxication des métaux, produit seulement des PCs en réponse à un stress au Cd (Mehra et Winge, 1991).

Il y a très peu d'informations concernant le rôle des MTs chez les champignons mycorhiziens. Des gènes codant pour des MTs putatives ont été découverts lors de séquençages systématiques de gènes exprimés chez les champignons ectomycorhiziens

Pisolithus tinctorius (Voiblet *et al.*, 2001) et endomycorhiziens *Glomus intraradices* (Stommel *et al.*, 2001) et *Gigaspora margarita* (Lanfranco *et al.*, 2002).

De même, les données concernant les MTs de protistes sont rares. Elles concernent exclusivement des espèces du genre *Tetrahymena*, chez lesquelles deux principales sous-familles de MTs ont été caractérisées (Diaz *et al.*, 2007); la sous-famille 7a qui inclut 6 Cd-thionéines et la sous-famille 7b constituée par 3 Cu-thionéines. Les deux sous-familles diffèrent par leur profil d'induction par des métaux lourds (principalement par le Cd ou le Cu, mais pas exclusivement) (Boldrin *et al.*, 2002 ; Diaz *et al.*, 2007; Dondero *et al.*, 2004 ; Shang *et al.*, 2002) et par le motif dessiné par leurs résidus cystéines (Diaz *et al.*, 2007). Les MTs de ciliés présentent des caractéristiques exclusives par rapport aux MTs « classiques ». Ce sont des protéines plus longues (96-181 acides aminés, 10-19 kDa), avec un contenu en résidus Cys plus important (22-54) et contenant pour plusieurs d'entres elles des acides aminés aromatiques et des résidus histidine (Diaz *et al.*, 2007).

4.4.5 Compartimentation vacuolaire

La compartimentation joue un rôle essentiel dans les mécanismes de tolérance et de détoxication des métaux lourds en évitant leur libre circulation dans le cytoplasme et en les concentrant dans un espace réduit où ils n'induisent pas de « stress biologique ».

Chez *S. cerevisiae*, le Cd est transporté et séquestré dans les vacuoles sous forme de Cd-$(GSH)_2$ par la perméase spécifique YCF1. Il s'agit d'un transporteur de type ABC, présentant des homologies avec les protéines de la famille des MRP (multi-drug resistance-associated protein) animales (Li *et al.*, 1997). L'importance de ce système de détoxication est démontrée par l'hypersensibilité au Cd de *S. cerevisiae* consécutive à la délétion du gène *ycf1* (Wemmie *et al.*, 1994).

Un homologue de ce transporteur membranaire a été identifié dans les vacuoles d'un mycelium de *Paxillus involutus* comme étant impliqué dans la translocation (et la séquestration physique) de complexes Cd-$(GSH)_2$ (ou Cd-(γ-$GluCys)_2$) dans les vacuoles (Courbot *et al.*, 2004). Dans les conditions de pH acide de la vacuole (environ 5,4), il est vraisemblable que ces complexes soient dissociés et que les molécules de GSH et de γ-GluCys soient dégradées par des hydrolases, ce qui permet à la cellule de récupérer des cystéines sous forme réduite. Les acides aminés résultants de cette dégradation sont redirigés vers le cytoplasme. Les métaux quant à eux sont alors pris en charge par des acides

organiques comme le citrate, le malate ou l'oxalate et sont rendus inactifs en formant des cristaux par complexation (Sanità-di-Toppi et Gabbrielli, 1999).

Trois transporteurs intracellulaires potentiels de zinc ont été identifiés chez *S. cerevisiae*. Ces transporteurs sont trois membres de la famille CDF (cation diffusion facilitator), il s'agit de Zrc1, Cot1 et Msc2. Le gène *zrc1* a été défini comme un déterminant de résistance au zinc ; la surexpression de *zrc1* aboutit à la capacité accrue des cellules à tolérer de hautes concentrations de zinc (Kamizono *et al.*, 1989). Le gène *cot*1 a été isolé d'une façon semblable à *zrc*1, c'est-à-dire, comme suppresseur de la toxicité au cobalt et, plus tard, pour conférer une résistance au zinc (Conklin *et al.*, 1992 ; Conklin *et al.*, 1994). La délétion de *zrc1* ou *cot1* aboutit à une plus grande sensibilité à un excès de zinc, supportant le rôle potentiel de ces gènes dans la compartimentation du zinc.

Enfin, la sensibilité au Zn de mutants de levures H^+-ATPase prouve que l'acidification de la vacuole est nécessaire à la compartimentation vacuolaire du Zn et que son transport dépend d'un antiporteur Zn^{2+}/H^+ (Eide *et al.*, 1993 ; Nishimura *et al.*, 1998).

4.4.6 Système de détoxication antioxydatif

La plupart des cellules est équipée de systèmes antioxydatifs efficaces composés à la fois de mécanismes non enzymatiques (GSH, acide ascorbique) et enzymatiques (*peroxydase*, *superoxyde dismutase* et *catalase*).

Vido *et al.*, (2001) ont montré que des souches de *S. cerevisiae* délétées pour les *superoxydes dismutases* cytosolique et mitochondriale (SOD1 et SOD2) sont hypersensibles au Cd. Ils ont également observé que plusieurs systèmes antioxydatifs représentés par les gènes ahp1 (*alkyl hydroperoxyde réductase*) et tsa (*thioperoxydase*) sont induits par le Cd.

Le Cd peut aussi contribuer indirectement au stress oxydatif en affectant l'équilibre thiol-rédox cellulaire. Il est capable de se fixer sur les thiorédoxines (TRX), probablement au niveau du site actif dithiol, ce qui inhibe l'activité de ces dernières. L'analyse de la réponse protéomique de *S. cerevisiae* à un stress au Cd, a révélé que les deux systèmes cellulaires de maintien du statut redox, glutathion et thiorédoxine, étaient significativement induits (Vido *et al.*, 2001). De plus, les souches délétées à la fois pour les gènes de 2 thiorédoxines cytosoliques (TRX1 et TRX2) ou pour la thiorédoxine réductase (TRR1) sont hypersensibles au Cd (Vido *et al.*, 2001). Enfin, le régulateur transcriptionnel YAP-1 de *S. cerevisiae*, dont la délétion entraîne une hypersensibilité au Cd (Lee *et al.*, 1999) et dont la surexpression

entraîne une hyper-résistance à ce métal (Wu et Moye-Rowley, 1994), est impliqué dans le contrôle de plusieurs gènes de réponse au métal tels que *ycf1* et *gsh1*, et également dans l'induction de gènes de défense antioxydants tels que *trx* et *ccp1*.

L'ensemble de ces mécanismes pourrait se rencontrer chez les champignons filamenteux. En effet, Jacob *et al.*, (2001 ; 2004) ont montré non seulement une induction de la SOD en réponse au Cd chez *P. involutus*, reflétant une augmentation du taux de $O_2^{-\cdot}$ mais aussi que AP-1, homologue de YAP-1, est surexprimé chez *P. involutus* lors d'un stress au Cd.

4.5 Conclusion

La toxicité directe (altération membranaire, enzymatique) et indirecte (stress oxydant) des métaux, et en particulier du Cd, sur les microorganismes eucaryotes induit de nombreux mécanismes capables d'y remédier. Ces mécanismes, dits de tolérance, prennent en charge ce composé toxique à tous les niveaux cellulaires, depuis sa complexation extracellulaire jusqu'à sa séquestration intra-vacuolaire. Cette réponse, innée ou adaptative, entraîne la sélection d'espèces résistantes ou tolérantes et est susceptible de modifier la biodiversité des microorganismes eucaryotes des sols pollués aux métaux lourds. Les microorganismes eucaryotes se développant dans les sols pollués constituent donc un réservoir de biodiversité encore peu exploité. Leur étude, par des approches métagénomique et métatranscriptomique, pourrait permettre d'identifier des mécanismes ou des gènes nouveaux impliqués dans la détoxication et ouvre potentiellement une nouvelle voie pour la biorémédiation des sols pollués par les métaux lourds.

-Chapitre 1-

Molecular diversity of soil eukaryotic communities, different molecules (18S rDNA *versus* cDNA) tell different stories

Ce chapitre fait l'objet d'une publication soumise à la revue *Applied and Environmental Microbiology*.

Le but de cette publication est d'estimer et de comparer la diversité taxinomique et fonctionnelle des microorganismes eucaryotes présents dans 5 sols forestiers, d'une part, par l'analyse des séquences partielles d'ADNr et d'ARNr 18S obtenues par une approche de clonage/séquençage et d'autre part, par l'analyse des ARNm poly-adénylés rétrotranscits obtenues par une approche métatranscriptomique. Il s'agit de la première étude comparative de la diversité eucaryote des sols basée sur ces approches. Un regard critique a ainsi pu être porté sur les résultats générés et notamment sur les proportions relatives des différents groupes taxinomiques au sein de chaque jeu de séquences.

Cette publication incorpore une partie de mes travaux de thèse mais aussi une partie de la thèse de Coralie Damon (co-auteur), qui étudie les conséquences de pratiques forestières, et plus particulièrement l'impact du couvert végétal (hêtre versus épicéa) sur la diversité taxinomique et fonctionnelle de communautés de microorganismes eucaryotes du sol. Des données de la thèse de Julie Bailly (Bailly *et al.*, 2007), qui s'intéressait à la diversité taxinomique et fonctionnelle des microorganismes eucaryotes présents dans un sol sous pin maritime, sont aussi incluses dans cette publication. Ma contribution à ce travail concerne l'analyse de deux sols prélevés sous des pins sylvestres, l'un non contaminé, l'autre anciennement contaminé par des métaux lourds (durant près d'un siècle). Ces deux sols font l'objet d'analyses complémentaires présentées dans le chapitre 3.

Quelques détails concernant le protocole d'extraction des acides nucléiques du sol, qui ne figurent pas dans cette publication, sont décrits ci-dessous.

En effet, l'étape clé de ces approches concerne l'extraction et la purification des acides nucléiques (ADN et ARN) à partir de sols afin d'obtenir des ADN et des ARN en quantité et de qualité suffisante pour effectuer des manipulations de biologie moléculaire (rétro-transcription, amplification par PCR, clonage, construction de banques ribosomiques et métatranscriptomiques).

Un protocole d'extraction mis au point précédemment au laboratoire (Bailly *et al.*, 2007) a été testé et optimisé. Les différents essais effectués ont confirmé que le broyage préalable du sol à sec dans de l'azote liquide à l'aide d'un broyeur de roches préalablement

	Matière organique (g/kg)	SDS	Guanidine isothiocyanate
Paal	4,9	1,85 %	100 µM
Lommel	15,9	1,7 %	400 µM
Balen	30,9	1,6 %	600 µM

Tableau 10: Quantité de matière organique contenue dans les échantillons de sols prélevés sur les sites de Paal, Lommel et Balen et quantités de SDS et de guanidine isothiocyanate utilisées lors de l'extraction des acides nucléiques.

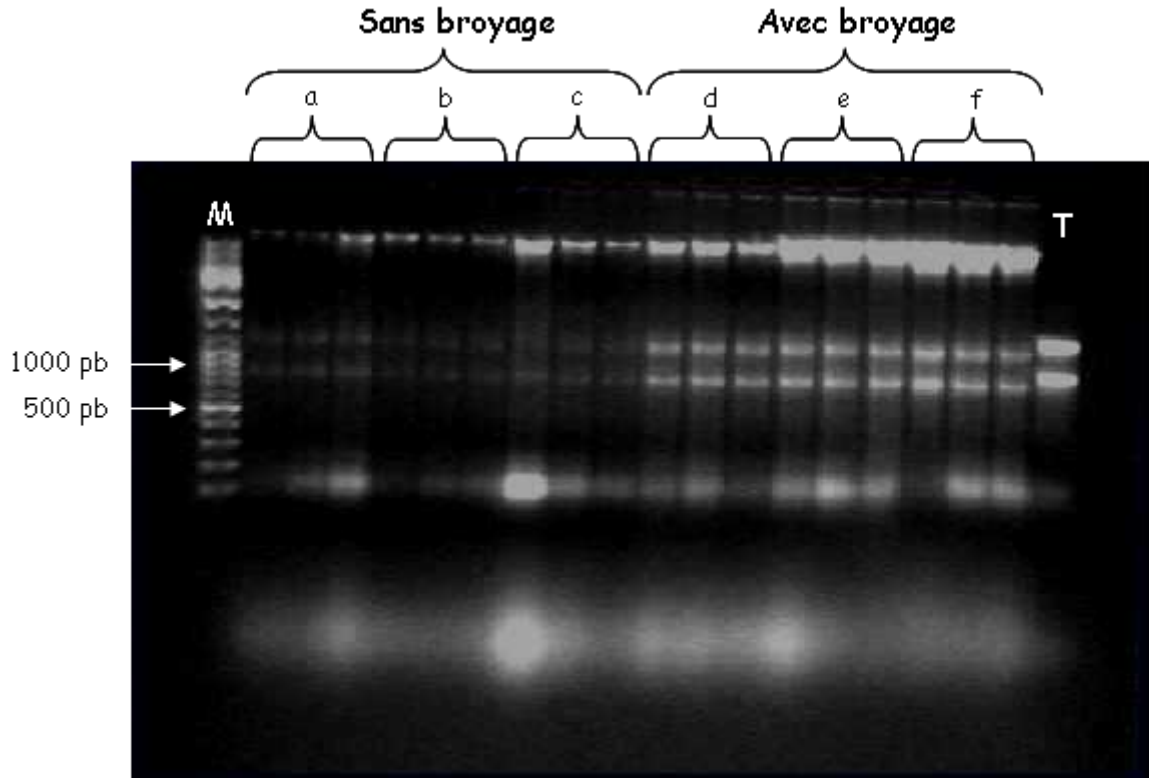


Figure 24: Extraction des acides nucléiques du sol de Lommel avec (d-f) ou sans (a-c) broyage préalable du sol à l'aide d'un broyeur de roche. 20 µL d'extrait sont déposés dans chaque puits.
a: extraction à partir d'1g de sol avec 1,7% de SDS et 400 µM de guanidine isothiocyanate.
b: extraction à partir d'1g de sol avec 1,8% de SDS et 200 µM de guanidine isothiocyanate.
c: extraction à partir de 0,5g de sol avec 1,7% de SDS et 400 µM de guanidine isothiocyanate.
d: extraction à partir de 0,5g de sol avec 1,85% de SDS et 100 µM de guanidine isothiocyanate.
e: extraction à partir de 0,5g de sol avec 1,8% de SDS et 200 µM de guanidine isothiocyanate.
f: extraction à partir de 0,5g de sol avec 1,7% de SDS et 400 µM de guanidine isothiocyanate.
M: marqueur de poids moléculaire
T: extrait d'ARN d'*Hebeloma cylindrosporum*.

refroidi à -70°C permettait d'augmenter de manière significative les rendements d'extraction des acides nucléiques (Fig. 24). Une durée de broyage de 3 minutes semble optimale pour les 5 sols de natures physico-chimiques différentes. Une autre amélioration a porté sur la quantité de sol utilisée par extraction, réduite à 0,5 g au lieu de 1 g afin de favoriser l'homogénéisation du sol dans le tampon d'extraction (Fig. 24). De plus, la seconde étape de broyage, à l'aide de billes de verres et d'un broyeur à billes (microdismembrator, Braun Biotech), permet également d'augmenter la surface de contact entre la matrice sol et le tampon d'extraction et donc d'améliorer le rendement d'extraction. L'optimisation la plus flagrante concerne la composition du tampon d'extraction. Suivant le type de sol, des volumes différents de tampon de lyse (contenant du SDS) et de solution dénaturante (contenant de la guanidine isothiocyanate) ont été utilisés (Fig. 24). Une corrélation entre la quantité de matière organique contenue dans l'échantillon de sol et la quantité de guanidine isothiocyanate utilisée a pu être mise en évidence (Tableau 10). Enfin, l'ajout d'une étape supplémentaire de précipitation au chlorure de lithium pendant une nuit permet d'augmenter significativement le rendement d'extraction des ARN.

Cependant, tous les acides nucléiques extraits des différents sols ne se prêtent pas immédiatement à des manipulations de biologie moléculaire. En effet, comme le suggère la couleur brune de certains extraits, une partie des acides humiques contenus dans certains échantillons de sol sont vraisemblablement co-extraits avec les acides nucléiques et inhibent les réactions enzymatiques effectuées ultérieurement. Une étape de purification sur une micro-colonne d'exclusion de taille contenant du Sephadex G-50 (ProbeQuant, Amersham) s'est montrée particulièrement efficace pour éliminer une bonne partie de ces inhibiteurs. Toutefois, l'utilisation d'un « mime moléculaire » de l'ADN comme la BSA (Bovin Serum Albumin) est indispensable lors des étapes de rétro-transcription et d'amplification par PCR afin de piéger les inhibiteurs enzymatiques restant.

Toutes ces optimisations ont permis de construire 5 banques ribosomiques d'ADNr et d'ARNr 18S et 5 banques d'ADNc environnementales à partir de 5 sols forestiers dont l'analyse et la comparaison sont présentées dans cette publication. Cette méthode a également permis la construction d'une banque d'ADNr et d'ARNr 18S et d'une banque d'ADNc environnementale à partir d'un sol forestier pollué par des métaux lourds. L'analyse de ces banques de sol pollué est décrite dans le chapitre 3.

Molecular diversity of soil eukaryotic communities, different molecules (18S rDNA *versus* cDNA) tell different stories.

Frédéric Lehembre,[1]✖ Coralie Damon,[1]✖ Julie Bailly,[1] Elise David,[1]✝ Jacques Ranger,[2] Jan Colpaert,[3] Laurence Fraissinet-Tachet[1] and Roland Marmeisse[1*]

Ecologie Microbienne, UMR CNRS, USC INRA, Université de Lyon, Université Lyon 1, 69622 Villeurbanne, France,[1] Biogéochimie des Ecosystèmes Forestiers, INRA centre de Nancy, 54280 Champenoux, France,[2] and Universiteit Hasselt, Centrum voor Milieukunde, 3590 Diepenbeek, Belgium[3]

✖ These two authors contributed equally to the work

✝ Present address: Laboratoire d'Eco-Toxicologie, Unité de Recherche Vigne et Vins de Champagne, Stress et Environnement, Université de Reims Champagne-Ardenne, Reims, France

* Corresponding author, mailing address: Ecologie Microbienne, UMR CNRS, USC INRA, Université de Lyon, Université Lyon 1, Bâtiment Lwoff, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France. Phone: +33 472448047. Fax: +33 472431643; E-mail: **roland.marmeisse@univ-lyon1.fr**

Running title: MOLECULAR DIVERSITY OF SOIL EUKARYOTES

ABSTRACT

Diversity of microbial communities is estimated by sequencing neutral markers such as ribosomal genes amplified from DNA extracted from environmental samples. This approach has rarely been applied to soil eukaryotes and does not give access to the functional diversity of these organisms, which play important roles in biogeochemical cycles. We compared the taxonomic diversity of different sets of eukaryotic nucleic acid sequences retrieved from five forest soils. Two sets were composed of partial 18S ribosomal genes, amplified from either soil-extracted DNA (rDNA) or reverse-transcribed RNA (rRNA). The third set was obtained following a metatranscriptomic approach. Soil polyadenylated mRNAs were converted into cDNAs which were sequenced. All major eukaryotic phyla were represented within the rDNA and rRNA datasets and the 94 to 157 sequences obtained from each soil did not saturate the true diversity at each of the sites. Preliminary phylogenetic analyses suggested the presence of potentially new, deep-branching, phyla in the amebozoa and the cercozoa. In contrast, almost all cDNA sequences were attributed either to fungi, animals or plants. The almost complete absence of cDNA sequences from unicellular eukaryotes is likely an artefact resulting from the scarcity of protein sequences from these organisms in databases. The relative proportions of each sequence dataset vary unpredictably across taxonomic groups and forest sites. Despite these pitfalls, with nearly 50% of the soil cDNA sequences encoding proteins of known functions, the metatranscriptomic approach appears promising to study at the molecular level key soil biological process and how they respond to environmental changes.

INTRODUCTION

Amplification and sequencing of the nuclear small subunit ribosomal gene (18S gene) using DNA extracted from various environments has revealed an unexpected diversity of unicellular eukaryotes. Molecular surveys carried out using "universal" or phylum-specific 18S PCR primers have, in particular, revealed the existence of novel and yet widespread deep-branching phyla of unicellular eukaryotes Lopez-Garcia *et al.*, 2001; Moon-van der Staay *et al.*, 2001; Not *et al.*, 2007 and also a higher than expected species diversity within known phyla Bass et Cavalier-Smith, 2004; Cavalier-Smith et von der Heyden, 2007; Porter *et al.*, 2008; Viprey *et al.*, 2008. These molecular surveys also point to a high diversity of eukaryotes in environments traditionally considered as hostile to eukaryotic life, such as deep-sea hydrothermal vents Edgcomb *et al.*, 2001, anoxic basins Stoeck *et al.*, 2006 or hypersaline waters Fomina *et al.*, 2005. Finally, these molecular surveys have also been used to rank taxonomic and trophic groups according to their relative abundance thus highlighting the importance of sometime underestimated mechanisms such as parasitism in ecosystem functioning Guillou *et al.*, 2008; Lefevre *et al.*, 2008.

Most of these studies, which made use of "universal 18S primers", have been conducted in aquatic environments either marine or continental and very few of them on soils. Indeed, soil biologists tend to specialize on specific eukaryotic groups and very seldom try to appreciate and to quantify the global soil eukaryotic diversity. There is for example, a very large body of data specifically reporting, using phylum-specific primers, on the molecular diversity of e.g. soil fungi O'Brien *et al.*, 2005; Porter *et al.*, 2008; or nematodes Floyd *et al.*, 2002. On the opposite, other major eukaryotic phyla known to thrive in soils (such as the cercozoa or the excavates) are largely ignored if we except the studies specifically dealing with their taxonomy. One possible reason, which refrains soil microbiologists from studying microbial eukaryotic communities as a whole is that it is not possible to easily separate soil microorganisms *sensu stricto*, by eg filtration, from fine plant roots and metazoa in the millimetre range. Therefore, as indeed illustrated by the few studies on the subject, studies on the global soil eukaryotic communities encompass the entire eukaryotic domain, including the phyla with multicellular organisms O'Brien *et al.*, 2005; Tringe *et al.*, 2005; Lesaulnier *et al.*, 2008; Urich *et al.*, 2008.

Global studies on soil eukaryotic diversity seem necessary to compare basic biological processes between soils and how they are impacted by various environmental factors. One emerging experimental approach used to address functional diversity of complex microbial

communities is metatranscriptomic. This technique relies on the extraction of labile, short-lived "environmental RNA" and their conversion into cDNA which are sequenced. Distribution of the cDNAs by phylogenetic origin and into functional groups reflects "who is doing what" within the studied microbial community. Initial examples of this approach did not attempt to separate prokaryotic from eukaryotic microorganisms. cDNAs from the eukaryotic origin were recovered in all studies but they always represented a very small fraction of the sequences, thus preventing a detailed analysis of eukaryote activities in the studied environments Frias-Lopez *et al.*, 2008; Gilbert *et al.*, 2008; Poretsky *et al.*, 2009.

Metatranscriptomic analysis of a given environment can however be limited to eukaryotes thanks to the polyadenylation of the 3'-end of their mRNA which allows their specific isolation from a mixture of prokaryotic and eukaryotic RNA. Synthesis of "environmental eukaryotic cDNAs" was initially reported by Grant et al. Grant *et al.*, 2006 for aquatic and activated sludge samples and by Bailly et al. Bailly *et al.*, 2007 for a forest soil. In the latter study, the taxonomic distribution of the three studied DNA molecules (18S rDNA amplified from either soil DNA or reverse-transcribed soil rRNA and cDNA from reverse transcribed mRNA) did not match, with for example an almost complete absence of cDNA attributed to unicellular eukaryotes which were otherwise abundantly represented among the 18S rDNA sequence dataset.

The main objectives of the present study were (i) to extend the eukaryotic metatranscriptomic approach to different forest soils which differed with respect to e.g. plant cover and organic matter content and (ii) to determine whether the discrepancies between the distribution of different DNA molecules (rDNA *vs* cDNAs) between taxonomic groups obey to foreseeable specific rules. This latest issue is of importance for the interpretation and evaluation of metatranscriptomic data in the context of ecosystem functioning.

MATERIALS AND METHODS

Study sites and soil sampling. Five forest stands located in four different places in France (Truc Vert and Breuil sites) and Belgium (Lommel Sahara and Paal sites) were sampled. The Truc Vert (TV) site is a monospecific *Pinus pinaster* stand located in South West France (44° 43’ N, 1° 15’ W) on a stabilized coastal sand dune. More details about this site and the sampling procedure are given in Bailly et al. Bailly *et al.*, 2007. The Breuil site, located in the Morvan region in the centre of France (47° 18’ N, 4° 4’ E) was planted in 1976 by different and separate monospecific stands of coniferous and broadleaved tree species. Soil sampling was performed in a spruce (*Picea abies*) (BS) and a beech (*Fagus sylvatica*) (BB) stands. The

Lommel Sahara (LS, 51° 14' N, 5° 15' E) and Paal (PZ, 51° 04' N, 5° 10' E) sites are *Pinus sylvestris* (with a few *Betula sp.* trees) stands on poor sandy soils from the Campine region in North East Belgium (Limburg province). The LS site, reforested in 1975, was contaminated by heavy metals dispersed by a now-dismantled pyrometallurgical zinc smelter Vangronsveld *et al.*, 1995. These five study sites differ with respect to climate, substratum, vegetation, humus types and soil physico-chemical features (Table 1).

In each site, between 14 and 27 soil samples were collected, sieved (2 mm mesh size) to eliminate coarse litter and root fragments and pooled to form composite samples which were stored frozen at -70°C. In TV, soil samples (10x10x20 cm lxLxh) were collected underneath the fruit bodies of 25 different species of saprotrophic or symbiotic basidiomycetes Bailly *et al.*, 2007. In the LS and PZ stands, cores (5 cm in diameter X *ca* 20 cm in depth) were regularly collected along two *ca* 20 m transects. In the BB and BS stands, 8 cm in diameter cores were collected every meter on a regular sampling grid. In these last two stands only the most organic-rich upper layer (5-8 cm) was collected.

Nucleic acid extraction and purification. Nucleic acid (DNA and RNA) extraction from soil was performed as described in Bailly et al. Bailly *et al.*, 2007 with some minor modifications. Modifications concerned the final concentration of guanidine isothiocyanate and SDS in the extraction buffer which varied between 100-600 µM and 1.6-1.85% respectively depending on the soil. Soil DNA was recovered by ethanol precipitation from the supernatant of the LiCl solution used to precipitate the RNA. Polyadenylated eukaryotic mRNA was purified from total soil RNA by affinity capture on paramagnetic beads coated with poly-dT oligonucleotides as described in the Dynabeads Oligo (dT) kit instruction manual (Dynal). Ribosomal RNA, which did not bind to the beads, was recovered by ethanol precipitation.

Construction of the cDNA libraries. First and second strand cDNAs were synthesized from purified polyadenylated mRNA using the SMART cDNA Library Construction Kit (Clontech). PCR-amplified cDNAs were size-fractionated to remove cDNA smaller than 400pb (CHROMA SPIN-400 DEPC-H_2O Columns, Clontech), digested by *Sfi*1 and ligated into the *Sfi*1-digested pDNR-LIB plasmid vector (Clontech) for the PZ, LS, BB and BS soils or into pYESfi-URA3 plasmid for the TV soil cDNAs Bailly *et al.*, 2007.

Amplification and cloning of the 18S rDNA A ca 520 bp-long fragment located at the 5'-end of the eukaryotic 18S rDNA gene was amplified by PCR from both soil DNA and reverse-transcribed soil 18S rRNA using primers Euk1A (CTGGTTGATCCTGCCAG) and Euk516R (ACCAGACTTGCCCTCC) described by Diez et al. Diez *et al.*, 2001. Amplification using *Taq*-DNA polymerase and cloning of the 18S rDNA from the TV soil

were described in Bailly et al. Bailly *et al.*, 2007. For the four others soils, the PCR mixtures (25µL) contained 200nM of each primer, 200µM of each dNTP, 1mM $MgCl_2$, 0.25mg.mL^{-1} of bovine serum albumin, 0.625U of *Pfu* DNA polymerase and the appropriate buffer (Fermentas). Each amplification using 10-100 ng of soil DNA or 50 ng of reversed transcribed soil rRNA was performed for the lowest number of cycles (between 23 and 27) to minimise PCR artefacts.

Amplification products of the expected size from eight different PCR tubes were pooled and isolated from an agarose gel (Nucleospin Extract kit, Macherey-Nagel), ligated in the plasmid pCR-Blunt II-TOPO (Zero Blunt TOPO PCR Cloning kit, Invitrogen) that was used to transform electro-competent DH10B *E. coli* cells (Invitrogen).

Sequencing and sequence analysis. Sequencing and sequence analysis of clones from the different TV rDNA and cDNA libraries were performed as described by Bailly et al. Bailly *et al.*, 2007. In the present study, we generated and analysed an additional 68 cDNA sequences from the TV site. Sequencing of LS, PZ, BB and BS libraries was performed by Agowa Company (Berlin, Germany) using universal primers M13-20F or M13Rev for the 18S rDNA inserts cloned in pCR-Blunt II-TOPO and primer M13-21F for the cDNA inserts cloned in pDNR-LIB. Sequence accession No. are given in Table 1.

Sequences were manually corrected and edited. BLAST (blastn or blastx) searches Altschul *et al.*, 1997 were performed against various sequence databases at NCBI (http://www.ncbi.nlm.nih.gov/). 18S rDNA were analyzed with the rRNA Database Project CHECK_CHIMERA program (http://rdp8.cme.msu.edu/). Potential chimeras were further analysed by blasting separately the two dissimilar segments of the sequences against GenBank. Functional and taxonomic annotation of protein coding sequences was performed manually on the basis of the Blastx searches by looking not only at the 'best hit' but at the different 'best hits' that can correspond to sequences from different taxonomic groups and also by considering different criteria: expect value, percents of identity and similarity, length of the alignment.

Multiple alignments of rDNA sequences were performed using the rRNA Database of the SILVA Project Pruesse *et al.*, 2007 using SINA (http://www.arb-silva.de/aligner/) and edited using SeaView Galtier *et al.*, 1996. Phylogenetic trees were computed and drawn using one of the method implemented in Phylo_win Galtier *et al.*, 1996.

Data analysis. Rarefaction curves for the 18S rDNA sequences were computed using S. Holland's Analytical Rarefaction version 1.3 software (http://www.uga.edu/strata/software/). The abundance-based richness estimators S_{Chao1} and S_{ACE} were computed for different

subsamples of different sizes drawn from the entire 18S rDNA data set as described by Kemp & Aller Kempf et Aller, 2004 and as implemented at http://www.aslo.org/lomethods/free/2004/0114a.html. Pearson's Chi-square statistical test was used to test differences between proportions of the different molecular markers (rDNA, rRNA and mRNA) within a taxonomic group in each of the sites.

RESULTS

Taxonomic affiliation of environmental sequences. Total soil RNA was successfully extracted from between 90 and 250 g of soil from the five forest sites, with yields ranging from 0.36 $\mu g.g^{-1}$ of soil (TV site) to 1.1 $\mu g.g^{-1}$ of soil (BB, BS and PZ sites). PolyA-mRNA were converted into cDNAs which were cloned to give cDNA libraries whose titre ranged from 9.6 10^5 colony forming units (cfu, BB site) to 15 10^6 cfu (LS site). Between 116 and 189 cDNAs were sequenced from their 5' end for each library; none of the sequences were homologous to rDNA ones. Partial 18S ribosomal sequences were amplified from soil DNA (referred to as rDNA) as well as from reverse-transcribed 18S soil rRNA (referred to as rRNA). ca 96 cloned 18S rDNA and 96 cloned rRNA were sequenced for each soil. Unexpectedly, between 17% and 34% of the rDNA and between 10% and 24% of the rRNA sequences turned out not to be 18S rDNA genes. Several of these sequences were homologous to bacterial protein-coding genes.

Each sequence (either cDNA, rDNA or rRNA), was affiliated to one of 8 large Eukaryotic phyla as defined in e.g. Lopez-Garcia & Moreira Lopéz-Garcia et Moreira, 2008: fungi, metazoa, plantae, amoebozoa, alveolata, heterokonta, rhizaria and excavata (the last 5 groups are sometime collectively referred to as "protists" in the text) (Fig. 1). Two additional categories were created for the cDNAs. Those which coded for proteins having similar levels of identity/similarity to proteins from different phyla were binned in a "multiple affiliation category", while those that were homologous to prokaryotic protein-coding genes were binned in the "bacteria" category. The latter sequences may represent genuine bacterial genes cloned by chance or, eukaryotic genes whose homologous sequences have, at present, only been reported from prokaryotes. If we do not consider plant sequences, whose abundance reflect the amount of fine roots which passed through the sieve, for all five sites the sequence dataset was dominated by sequences from fungi and animals (Fig. 1). The two best-represented protist groups were the rhizaria (with sequences from cercozoa exclusively) and the alveolata (ciliates, apicomplexa, perkinsea and dinoflagelates) while sequences from heterokonta (e.g. oomycetes) and excavata (euglenozoa exclusively) were rare or even absent from several

datasets. This global trend should however be nuanced as, for example, in the case of the BB site, 18S rDNA sequences from alveolata were 3.6 times more abundant that those from fungi. On the contrary at the PZ site 18S rDNA sequences from alveolata were absent from the dataset.

The relative proportions of each of the three studied sequences (18S rDNA, 18S rRNA and cDNA) in each of the different phyla were usually different but no clear systematic pattern was observed except for systematic excess of cDNA relative to 18S sequences in the fungi and a deficit in the metazoa (for statistical supports, see Fig. 1). Concerning the 18S rDNA/rRNA ratio, for a particular phylum in a specific site it can either be close to one or alternatively high (eg for the rhizaria in LS or the metazoa in BB) or low (eg the fungi in BB). The most striking result was the almost complete absence of cDNA attributed to protists except for the amoebozoa.

Diversity of eukaryotic communities. 18S sequences (both rDNA and rRNA) were clustered using both 96% and 98% identity threshold to define molecular operational taxonomic units (MOTUs; Table 2). As illustrated for the fungi (Fig. 2), the 18S dataset was globally characterised by a low level of redundancy. At the 98% threshold, within sites, between 6% (TV) and 13% (PZ) of the MOTUs were represented by both rDNA and rRNA sequences. For the three most represented phyla, fungi, animals and cercozoas, the percentage of MOTUs represented by both rDNA and rRNA sequences was greater for animals (6-38%) and lower for the cercozoas (0% in 4 of the sites, 14% at TV). Globally, only 17 % of the MOTUs were represented by 3 or more sequences (either DNA or RNA) in the different sites and the percentage was again highest for the animals (30%). Altogether, these data resulted in rarefaction curves that clearly did not reach an asymptotic value (Fig. 3) thus indicating that we undersampled both the global soil eukaryotic diversity and the diversity of each of the different eukaryotic phyla at the different sites. Based on the calculation of richness estimators S_{ACE} and S_{Chao1} we could estimate that we observed between 31% (in TV, considering the value of S_{ACE}) and 53% (in BB, considering S_{Chao1}) of the 18S phylotypes present in the soil samples (Table 2).

Sequences from each MOTU were ranked according to their percentage of identity to the most similar sequence present in GenBank (either annotated or not). The distribution of these figures differed markedly between eukaryotic phyla (Table 3). For fungi and metazoas, three quarters of the sequences found a homologue in GenBank with more than 95% of identity. This figure was significantly lower in the case of the protist phyla as exemplified by

the amoebozoa for which, on the contrary, almost three quarters of the sequences found a homologue in GenBank with less than 95% of identity.

Although the amplified fragment of the 18S rDNA gene does not allow for the construction of robust molecular phylogenies with highly supported branches, most of our environmental sequences could be placed within clades which grouped together annotated sequences from organisms belonging to the same phylum. Therefore, despite some low percentages of identity to known sequences, most of the 18S environmental sequences generated in this study do not probably identify new very deep-branching eukaryotic phyla. Potential exceptions are two groups of three and four phylotypes within the amoebozoa (Fig. 4) which cluster exclusively with additional environmental sequences from soil or fresh water. Another exception is a group of 8 phylotypes which define a well supported clade affiliated to the cercozoas (supplementary figure S1).

Annotation of environmental cDNA sequences. Based on Blastx searches, cDNA sequences were classified as coding for (i) proteins of known functions, (ii) for proteins of unknown functions with homologues in databases or (iii) for new hypothetical proteins. After excluding short (<400bp) sequence reads to minimise the contribution of non-coding 3' UTRs and of short, poorly-conserved C-terminal polypeptide ends, the latter category was equally represented (ca 30%) in four of the cDNA libraries (Fig. 5). In the library from BS it represented ca 40% of the cDNAs. On the opposite, cDNA encoding proteins of known functions consistently represented ca 50% of the five datasets.

DISCUSSION

The results obtained demonstrate that a metatranscriptomic approach focussing specifically on eukaryotic organisms, with a low background contamination with prokaryotic mRNA, is feasible on a wide range of soils, including organic matter-rich horizons of temperate forest soils. Eukaryote-specific metatranscriptomic analysis will be important to study, at the gene level, major soil processes that are essentially carried out by soil eukaryotic microbes such as plant-derived organic matter degradation. In addition, eukaryote-specific metatranscriptomic analysis will also facilitate the analysis of functions expressed by low-abundance soil eukaryotic taxonomic (e.g. the Oomycetes) or functional groups whose mRNA will not be diluted further by prokaryotic ones.

In this study we cloned the cDNAs and sequenced them using the Sanger chemistry. This strategy presents the advantage of having at our disposal the cDNA clones which, if

containing full-length ORF, can be used to produce the corresponding proteins for functional studies. The protocol can however certainly be adapted for pyrosequencing by following for example the strategy described by Frias-Lopez et al. Frias-Lopez *et al.*, 2008 who amplified antisense mRNA using T7-RNA polymerase prior to the synthesis of cDNAs by random priming.

We found that *ca* 50% of the cDNA clones coded for proteins of known functions. This figure suggests that moderate to deep sequencing of eukaryotic metatranscriptomes could indeed be very informative to decipher the biological processes carried out by these organisms directly in soils. This figure can be explained by the fact that eukaryotic metatranscriptomes, as for eukaryotic transcriptomes in general, are enriched in conserved housekeeping proteins which can be identified even in poorly studied phylogenetic lineages. Another potential reason is that soil eukaryotic biota are dominated by animals and fungi, two intensively studied eukaryotic groups on which concentrate most of the current genome sequencing efforts (see the Genome Online Database: http://www.genomesonline.org/gold.cgi). In this respect, it is likely that the sequencing of metatranscriptomes from aquatic microbial assemblages enriched in representatives from e.g. the stramenopiles, the alveolates, the cercozoas or the euglenids Stoeck *et al.*, 2006; Lefevre *et al.*, 2008 could result in far higher percentages of gene falling in the category "new hypothetical proteins" due to the current low coverage of these phyla in term of genome sequences available.

The current paucity, or almost complete absence (eg for the cercozoas), of protist protein sequences deposited in databases contrasts to the abundance of data for fungi, metazoas and plants. This certainly explains, for a large part, the almost complete absence of cDNA sequences attributed to the former groups although these groups were systematically detected among the 18S rDNA sequences. cDNA sequences from protists, if present, must therefore have been wrongly placed in the other taxonomic categories (Fig. 1). This hypothesis is however not sufficient to explain other discrepancies in the sequence data set.

One discrepancy concerns the within-clade differences between rDNA and rRNA datasets. In microbial ecology, PCR amplification from reverse-transcribed rRNA extracted from environmental samples has been proposed to better describe the diversity of the metabolically active organisms within a community Muttray et Mohn, 1999; Mills *et al.*, 2005. The situation could be more complex with eukaryotes as the number of rDNA gene copies varies widely between taxa as does the number of nuclei per unit of cytoplasm. Furthermore, we observed that for a given clade (eg, the fungi, Fig. 1) the relative proportions

of 18S rRNA and of cDNAs do not always vary in the same direction, a phenomenon not expected for two classes of molecules supposed to both describe physiological activities.

Another obvious discrepancy concerns the almost systematic under-representation of the metazoa among the cDNAs although they are abundantly represented in the rDNA and rRNA dataset (Fig. 1). We could tentatively attribute this observation to a phylum-specific feature; the molar ratio 18S rRNA over mRNA could be systematically higher in animals than in fungi. Apart from this hypothesis, which needs to be tested, another possible explanation is that the primer set used to amplify the 18S sequences could preferentially amplify sequences from specific taxonomic groups to the detriment of others. This may not only result from mismatches at the 18S PCR primer binding sites, but also from length variations and the occurrence of secondary structures which could affect amplification yield. Due to extensive variations in the sequence of the 18S gene across the eukaryotic domain Nickrent et Sargent, 1991; Choe *et al.*, 1999; Van de Peer *et al.*, 2000, it appears that a single PCR primer set that amplify a segment or most of the sequence may not indeed accurately describe the full diversity of eukaryotic communities Jeon *et al.*, 2008.

In term of global taxonomic diversity, for almost all of the studied soils, we recovered 18S sequences from the different major eukaryotic phyla with a clear dominance of the opistokonts (fungi plus animals). All animal sequences were related to taxonomic groups typical of the micro- mesofauna (nematodes, acari, collembolas, millipedes…, Supplementary Fig. S2) while basidiomycete sequences dominated the fungal sequence pool as already observed in temperate forest soils colonised by ectomycorrhizal plants O'Brien *et al.*, 2005. For the "protist" groups, sequences from the alveolates (essentially ciliates and parasitic apicomplexas; Supplementary Fig. S3) and the rhizaria (cercozoas only; Supplementary Fig. S1) dominated while heterokonta and excavata (euglenozoas) were the least represented. As expected for soils, sequences from major photosynthetic phyla (eg chlorophyta, diatoms) were undetected although unicellular algae do occur on the soil surface. Similarly, we did not detect sequences from oomycetes (stramenopiles), which include numerous soil-borne and widespread plant parasites. Different parameters, such as the absence of overlap between the rDNA and rRNA sequence data and the shape of the rarefaction curves indicate that additional sequencing effort combined with the use of group-specific PCR primers are needed to fully describe the extent of eukaryotic diversity in the five studied soils.

As pointed out in the introduction, soil is a neglected environment in term of global surveys of its eukaryotic diversity. We could therefore envisage that it hosts new, yet

undescribed, lineages which have evolved to adapt to soil specific features such as periodic desiccation. Despite the limited phylogenetic information contained in the portion of the 18S gene that we sequenced, some of our sequences in the cercozoas and the amoebozoas are potential candidates for such lineages. These sequences have low similarities (down to ca 70-75 % identity) to sequences from known organisms and cluster together as well as with other environmental sequences in the case of the amoebozoa (Fig. 4) in phylogenetic analyses. Furthermore, they have been amplified independently from different soils, thus excluding the fact that they could represent PCR artefacts such as chimeras. These short sequences can be used as templates to design group-specific oligonucleotides which, in combination with universal primers will be used to amplify the full-length 18S rDNA gene or even the full-length rDNA nuclear cluster for the construction of robust molecular phylogenies Porter *et al.*, 2008. These phylogenies should confirm the affiliation of these divergent clusters of sequences to either the cercozoas or the amoebozoas. Group-specific oligonucleotides can also be used as probes for FISH to characterise morphologically these new microbes Simon *et al.*, 1995; Massana *et al.*, 2006.

ACKNOWLEDGEMENT

This study was supported by grants from the PNETOX programme of the French ministry for ecology and sustainable development; the ANR Ecoger programme (project microger); the ANR Biodiversity programme (project 2006 Fundiv) and the ANR Blanc programme (project 2006 EUMETATOX). Sampling in Belgium was facilitated by a France-Flanders region Tournesol bilateral exchange programme.

We would like to thank the DTAMB platform of the Institut Fédératif de Recherche 41 and Jérôme Briolay for access to specific equipments and helpful advices as well as to Marie-Christine Verner for excellent technical help.

REFERENCES

1. **Alexander, E., A. Stock, H. W. Breiner, A. Behnke, J. Bunge, M. M. Yakimov, and T. Stoeck.** 2009. Microbial eukaryotes in the hypersaline anoxic L'Atalante deep-sea basin. Environ. Microbiol. **11:**360-81.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389-402.
3. **Bailly, J., L. Fraissinet-Tachet, M. C. Verner, J. C. Debaud, M. Lemaire, M. Wesolowski-Louvel, and R. Marmeisse.** 2007. Soil eukaryotic functional diversity, a metatranscriptomic approach. Isme J. **1:**632-42.
4. **Bass, D., and T. Cavalier-Smith.** 2004. Phylum-specific environmental DNA analysis reveals remarkably high global biodiversity of Cercozoa (Protozoa). Int. J. Syst. Evol. Microbiol. **54:**2393-404.
5. **Cavalier-Smith, T., and S. von der Heyden.** 2007. Molecular phylogeny, scale evolution and taxonomy of centrohelid heliozoa. Mol. Phylogenet. Evol. **44:**1186-203.
6. **Choe, C. P., U. W. Hwang, and W. Kim.** 1999. Putative secondary structures of unusually long strepsipteran SSU rRNAs and its phylogenetic implications. Mol. Cells **9:**191-9.
7. **Diez, B., C. Pedros-Alio, and R. Massana.** 2001. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. Appl. Environ. Microbiol. **67:**2932-41.
8. **Edgcomb, V. P., D. T. Kysela, A. Teske, A. de Vera Gomez, and M. L. Sogin.** 2002. Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. Proc. Natl. Acad. Sci. USA **99:**7658-62.
9. **Floyd, R., E. Abebe, A. Papert, and M. Blaxter.** 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. **11:**839-50.
10. **Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. Delong.** 2008. Microbial community gene expression in ocean surface waters. Proc. Natl. Acad. Sci. USA **105:**3805-10.
11. **Galtier, N., M. Gouy, and C. Gautier.** 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12:**543-8.

12. **Gilbert, J. A., D. Field, Y. Huang, R. Edwards, W. Li, P. Gilna, and I. Joint.** 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One **3:**e3042.

13. **Grant, S., W. D. Grant, D. A. Cowan, B. E. Jones, Y. Ma, A. Ventosa, and S. Heaphy.** 2006. Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. Appl. Environ. Microbiol. **72:**135-43.

14. **Guillou, L., M. Viprey, A. Chambouvet, R. M. Welsh, A. R. Kirkham, R. Massana, D. J. Scanlan, and A. Z. Worden.** 2008. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environ. Microbiol. **10:**3349-65.

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

16. **Kempf, P., and J. Aller.** 2004. Estimating prokaryotic diversity: when are 16S rDNA libraries large enough? Limnology and Oceanography: Methods **2:**114-125.

17. **Lefevre, E., B. Roussel, C. Amblard, and T. Sime-Ngando.** 2008. The molecular diversity of freshwater picoeukaryotes reveals high occurrence of putative parasitoids in the plankton. PLoS One **3:**e2324.

18. **Lesaulnier, C., D. Papamichail, S. McCorkle, B. Ollivier, S. Skiena, S. Taghavi, D. Zak, and D. van der Lelie.** 2008. Elevated atmospheric CO2 affects soil microbial diversity associated with trembling aspen. Environ. Microbiol. **10:**926-41.

19. **Lopéz-Garcia, P., and D. Moreira.** 2008. Tracking microbial biodiversity through molecular and genomic ecology. Research in Microbiology **159:**67-73.

20. **Lopez-Garcia, P., F. Rodriguez-Valera, C. Pedros-Alio, and D. Moreira.** 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature **409:**603-7.

21. **Massana, R., R. Terrado, I. Forn, C. Lovejoy, and C. Pedros-Alio.** 2006. Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. Environ. Microbiol. **8:**1515-22.

22. **Mills, H. J., R. J. Martinez, S. Story, and P. A. Sobecky.** 2005. Characterization of microbial community structure in Gulf of Mexico gas hydrates: comparative analysis of DNA- and RNA-derived clone libraries. Appl. Environ. Microbiol. **71:**3235-47.

23. **Moon-van der Staay, S. Y., R. De Wachter, and D. Vaulot.** 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature **409:**607-10.

24. **Muttray, A., and W. Mohn.** 1999. Quantitation of the population size and metabolic activity of a resin acid degrading bacterium in activated sludge using slot-blot hybridization to measure the rRNA:rDNA ratio. Microbial Ecology **38:**348-357.

25. **Nickrent, D. L., and M. L. Sargent.** 1991. An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. Nucleic Acids Res. **19:**227-35.

26. **Not, F., K. Valentin, K. Romari, C. Lovejoy, R. Massana, K. Tobe, D. Vaulot, and L. K. Medlin.** 2007. Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. Science **315:**253-5.

27. **O'Brien, H. E., J. L. Parrent, J. A. Jackson, J. M. Moncalvo, and R. Vilgalys.** 2005. Fungal community analysis by large-scale sequencing of environmental samples. Appl. Environ. Microbiol. **71:**5544-50.

28. **Poretsky, R. S., I. Hewson, S. Sun, A. E. Allen, J. P. Zehr, and M. A. Moran.** 2009. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. Environ. Microbiol. **11:**1358-75.

29. **Porter, T. M., C. W. Schadt, L. Rizvi, A. P. Martin, S. K. Schmidt, L. Scott-Denton, R. Vilgalys, and J. M. Moncalvo.** 2008. Widespread occurrence and phylogenetic placement of a soil clone group adds a prominent new branch to the fungal tree of life. Mol. Phylogenet. Evol. **46:**635-44.

30. **Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner.** 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. **35:**7188-96.

31. **Schadt, C. W., A. P. Martin, D. A. Lipson, and S. K. Schmidt.** 2003. Seasonal dynamics of previously unknown fungal lineages in tundra soils. Science **301:**1359-61.

32. **Simon, N., N. LeBot, D. Marie, F. Partensky, and D. Vaulot.** 1995. Fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes to identify small phytoplankton by flow cytometry. Appl. Environ. Microbiol. **61:**2506-13.

33. **Stoeck, T., B. Hayward, G. T. Taylor, R. Varela, and S. S. Epstein.** 2006. A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. Protist **157:**31-43.

34. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin.** 2005. Comparative metagenomics of microbial communities. Science **308:**554-7.

35. **Urich, T., A. Lanzen, J. Qi, D. H. Huson, C. Schleper, and S. C. Schuster.** 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One **3:**e2527.

36. **Valster, R. M., B. A. Wullings, G. Bakker, H. Smidt, and D. van der Kooij.** 2009. Free-living protozoa in two unchlorinated drinking water supplies, identified by phylogenic analysis of 18S rRNA gene sequences. Appl. Environ. Microbiol. **75:**4736-46.

37. **Van de Peer, Y., S. L. Baldauf, W. F. Doolittle, and A. Meyer.** 2000. An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. J. Mol. Evol. **51:**565-76.

38. **Vangronsveld, J., F. Van Assche, and H. Clijsters.** 1995. Reclamation of a bare industrial area contaminated by non-ferrous metals: in situ metal immobilization and revegetation. Environ. Pollut. **87:**51-9.

39. **Viprey, M., L. Guillou, M. Ferreol, and D. Vaulot.** 2008. Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. Environ. Microbiol. **10:**1804-22.

Site	BB	BS	LS	PZ	TV
vegetation type	*Fagus sylvatica*	*Picea abies*	*Pinus sylvestris*	*Pinus sylvestris*	*Pinus pinaster, Arbutus unedo,* grasses
soil texture	sandy clayey	sandy clayey	poor podzolized sandy soil	poor podzolized sandy soil	nutrient poor, non calcareous sand
soil pH	3.9 [b]	3.9 [b]	4.7	4.7	5.5
% of organic matter	13.4 [b]	22.2 [b]	1.58	0.49	0.5
C/N	19.2 [b]	20.2 [b]	20	24.3	ND
date of sampling	juil-07	juil-07	nov-06	nov-06	nov-04
soil moisture %	37.5 [b]	34 [b]	4	4.8	ND
No. of samples	16	14	20	20	27
volume of samples (cm^3)	750	750	2000	2000	2000
Soil temperature (°C)	ND	ND	11	11	11
18S rDNA accession Nos.	FN393222-3276	FN393277-3322	FN394732-4802	FN394803-73	AM409570-635[a]
18S rRNA accession Nos.	FN393180-3221	FN393323-3380	FN396422-6515	FN394874-4964	AM409518-569[a]
cDNA accession Nos.	FN432154-2328	FN431315-1660			AM409636-755[a] FN431234-1314

Table 1: Sites description, sampling characteristics and sequence accession numbers.

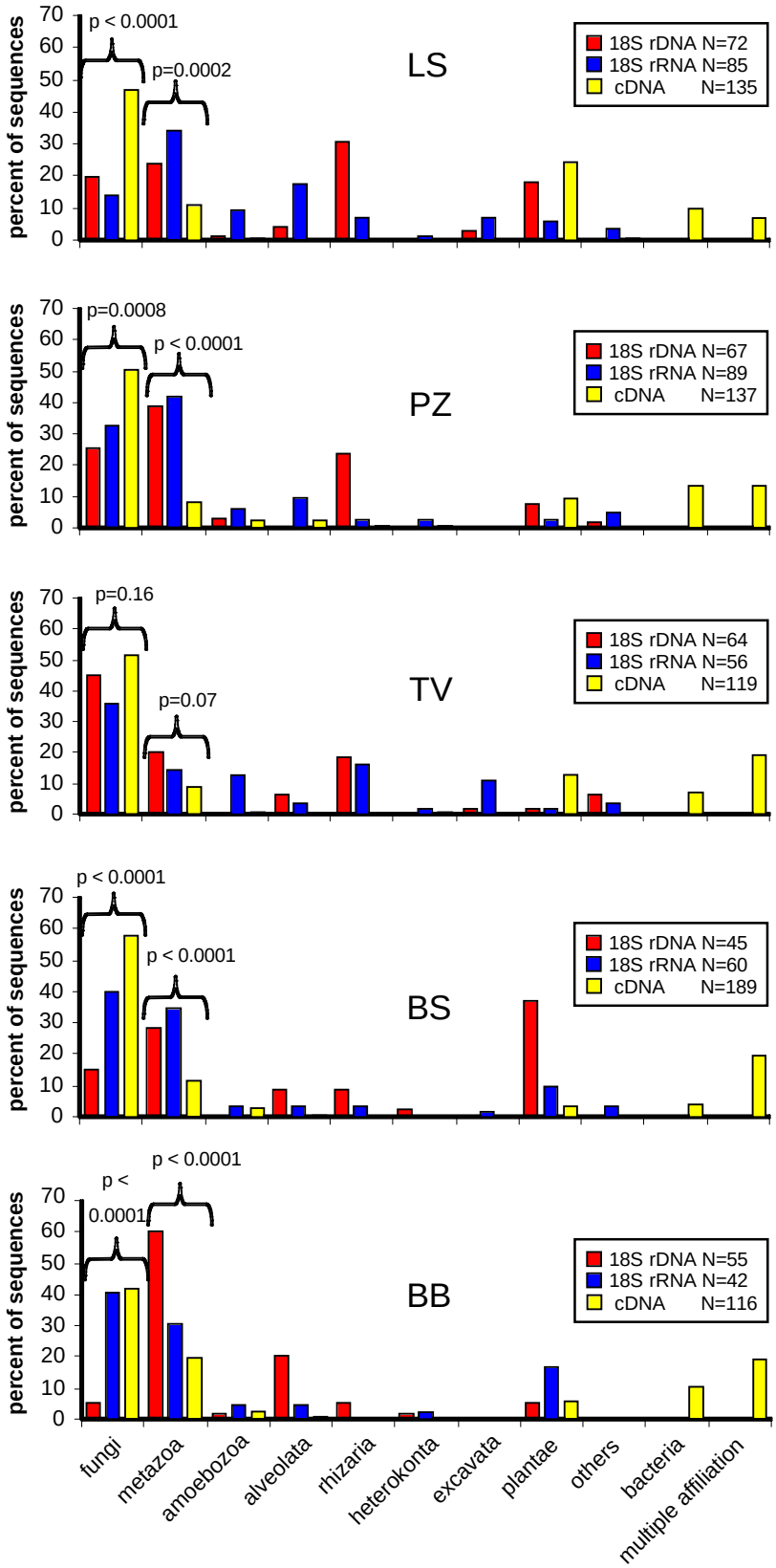
[a] From Bailly et al. Bailly *et al.*, 2007
[b] Data corresponding to the top organic layer of soil.
ND: not determined

Site	**LS**		**PZ**		**TV**		**BS**		**BB**	
% identity	98%	96%	98%	96%	98%	96%	98%	96%	98%	96%
Number of clones in library	157	157	156	156	121	121	106	106	94	94
Number of phylotypes observed	86	82	88	80	81	71	56	52	45	43
Predicted value of S_{ACE}	216	186	214	166	232	138	142	131	108	84
Predicted value of S_{Chao1}	179	164	176	158	159	122	130	141	84	69
Observed phylotypes / predicted S_{ACE}	0.4	0.4	0.4	0.5	0.3	0.5	0.4	0.4	0.4	0.5
Observed phylotypes / predicted S_{Chao1}	0.5	0.5	0.5	0.5	0.5	0.6	0.5	0.4	0.5	0.6

Table 2: Richness estimators S_{Chao1} and S_{ACE} from18S rDNA dataset, for each soil, with either 96 % and 98 % of identity cut-off value.

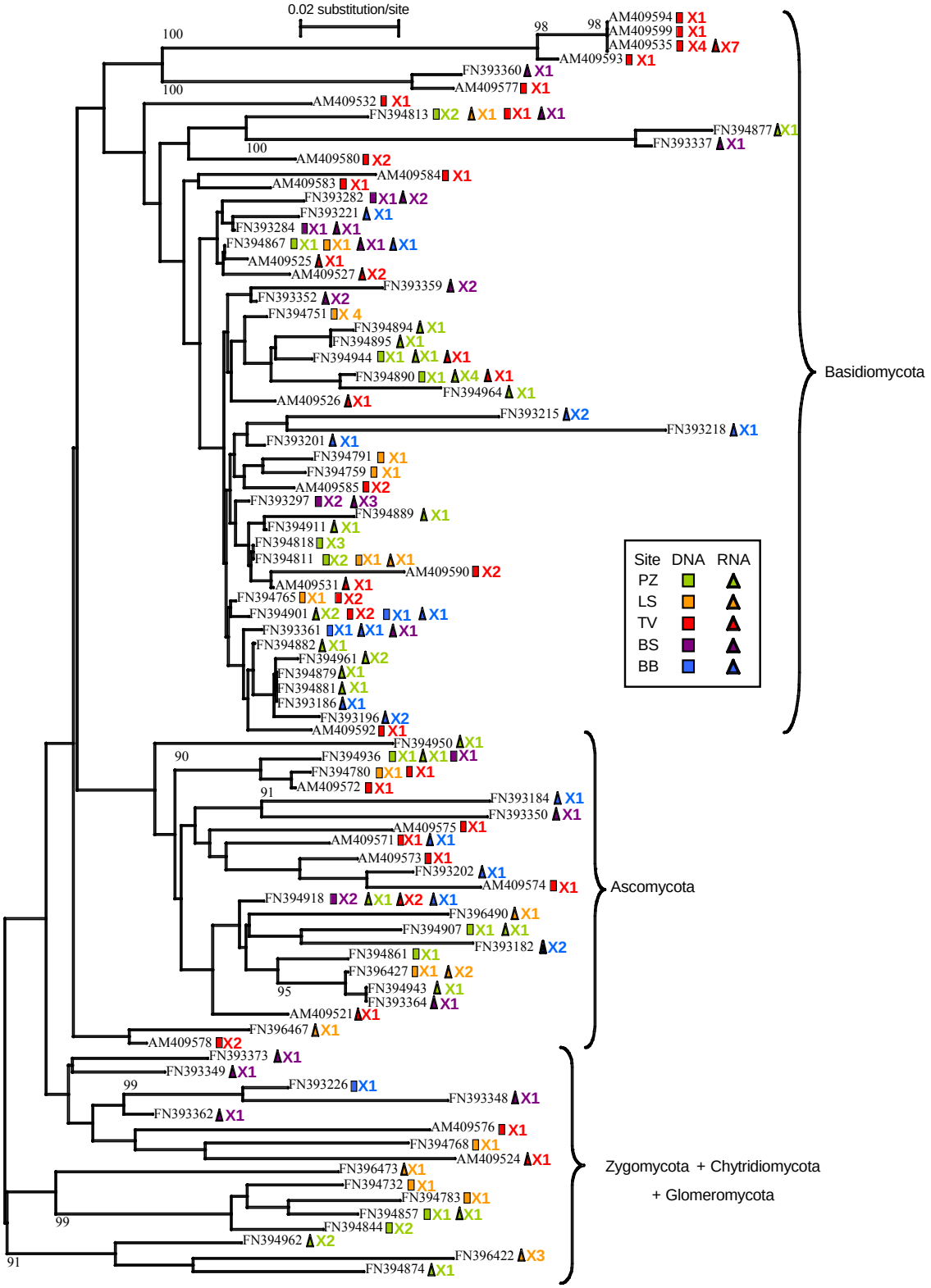
% Identity	100 - ≥98	97 - ≥95	94 - ≥90	<90
Fungi	51 %	25 %	24 %	0 %
Metazoa	46 %	23 %	20 %	11 %
Rhizaria	27 %	25 %	29 %	19 %
Alveolata	18 %	52 %	26 %	4 %
Amoebozoa	11 %	16 %	47 %	26 %
Excavata	0 %	0 %	67 %	33 %

Table 3: Distribution of the percentages of nucleotide sequence identity between the 18S sequences of the different phylotypes and their "best hit" in GenBank. The GenBank sequences can themselves be environmental ones. Analyses performed during the first semester of 2009.



Lehembre et al, Figure 1

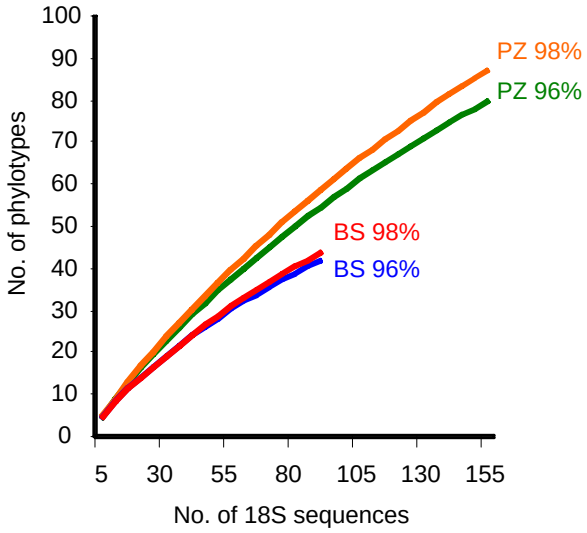
FIG. 1. Taxonomic assignment of the different categories of sequences (18S rDNA, 18S rRNA and cDNA). Pearson's Chi-square test was used to test differences between proportions of the different molecular markers (rDNA, rRNA and mRNA) within fungi and metazoa in each of the sites. The "other" category includes 18S sequences attributed to minor taxonomic groups such choanozoa or Ichthyosporea. cDNAs which displayed in Blastx searches similar e-values and percent of identity/similarity to sequences from several taxonomic groups were placed in the "multiple affiliation" category. For each site, the total No. of sequences used for the calculations is given.



Lehembre et al, Figure 2

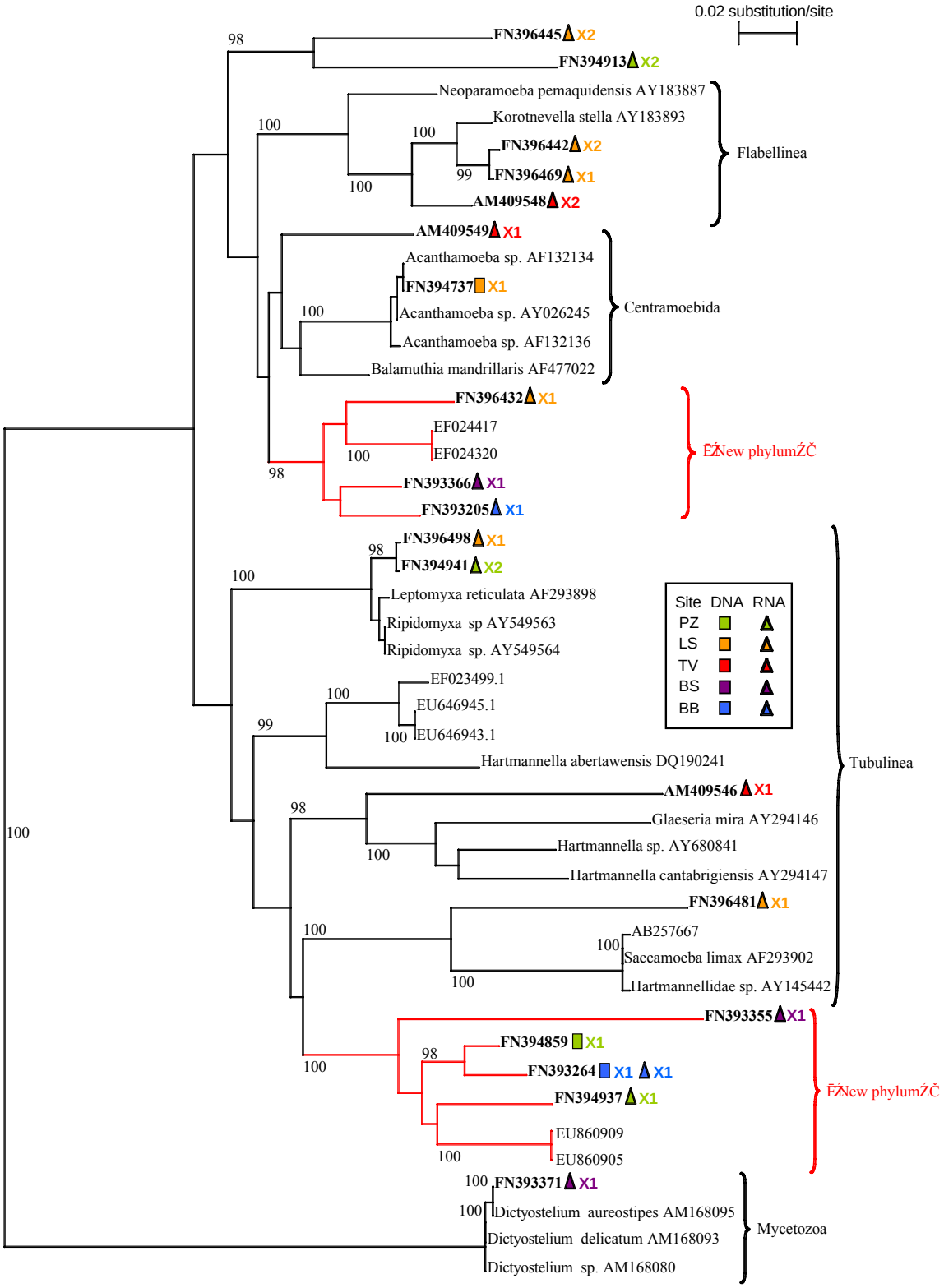
FIG. 2. Phylogenetic distribution of partial 18S fungal ribosomal sequences amplified from soil DNA (squares) or reverse-transcribed soil RNA (triangles). This tree illustrates that (i) a majority of the sequences were recovered only once and that (ii) there is little overlap between the rDNA and rRNA datasets and also between the different forest sites (a distinct colour has

been assigned to each site). The tree was constructed using the BioNJ distance method starting from an alignment which included the different environmental sequences plus their three most similar sequences retrieved from the ARB database using SINA. A first tree was computed with all sequences to define the different fungal phyla. Reference sequences were subsequently removed for legibility. Figures give the number of time a sequence was found in a particular dataset. Only bootstrap values (1000 replicates) above 90% are indicated. The use of PhyML gave an identical topology.



Lehembre et al, Figure 3

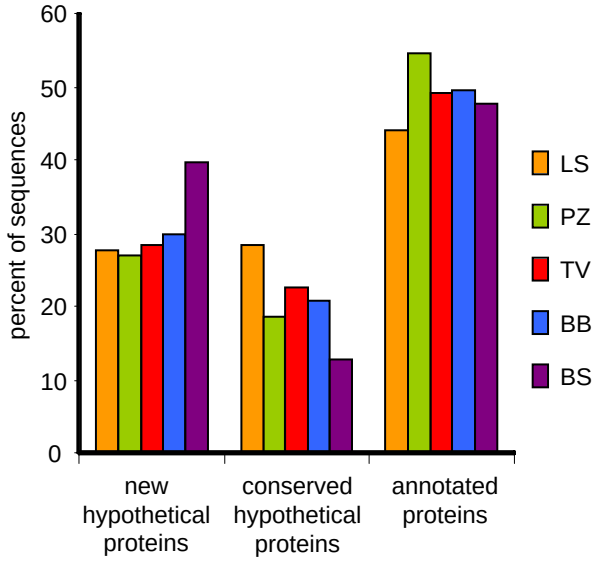
FIG. 3. Rarefaction curves determined for the 18S gene libraries generated for the PZ and BS forest sites. All 18S gene sequences were pooled irrespective of their origin (soil DNA or RNA) and taxonomic affiliation. For each library clustering of the sequences was performed using either a 98% or a 96% identity threshold.



Lehembre et al, Figure 4

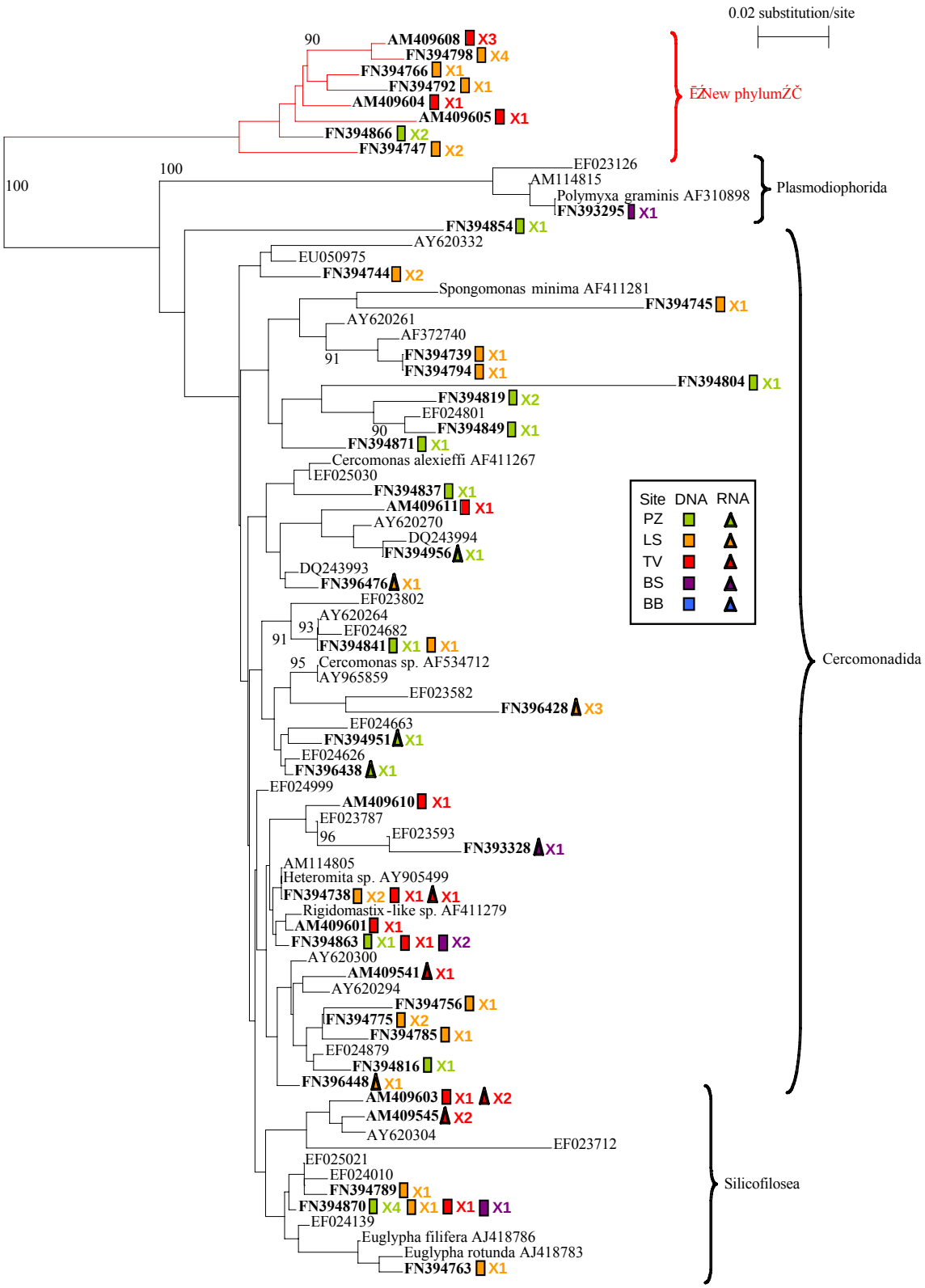
FIG. 4. Phylogenetic distribution of partial 18S amoebozoan ribosomal sequences amplified from soil DNA (squares) or reverse-transcribed soil RNA (triangles). The tree was constructed using the BioNJ distance method starting from an alignment which included the different environmental sequences plus their three most similar sequences retrieved from the

ARB or GenBank databases. Sequences from the present study are in bold. This tree illustrates that many soil amoebozoan sequences appear distantly related to sequences from characterised amoebae as well as other environmental sequences. Two well-supported clusters (in red) include only environmental sequences from different sites and from other studies; EF024320 and EF024417 are from forest soils Lesaulnier *et al.*, 2008; EU860909 and EU860905.1 are from fresh water Valster *et al.*, 2009. Figures give the number of time a sequence was found in a particular dataset. Only bootstrap values (1000 replicates) above 90% are indicated. The use of PhyML gave an identical topology.



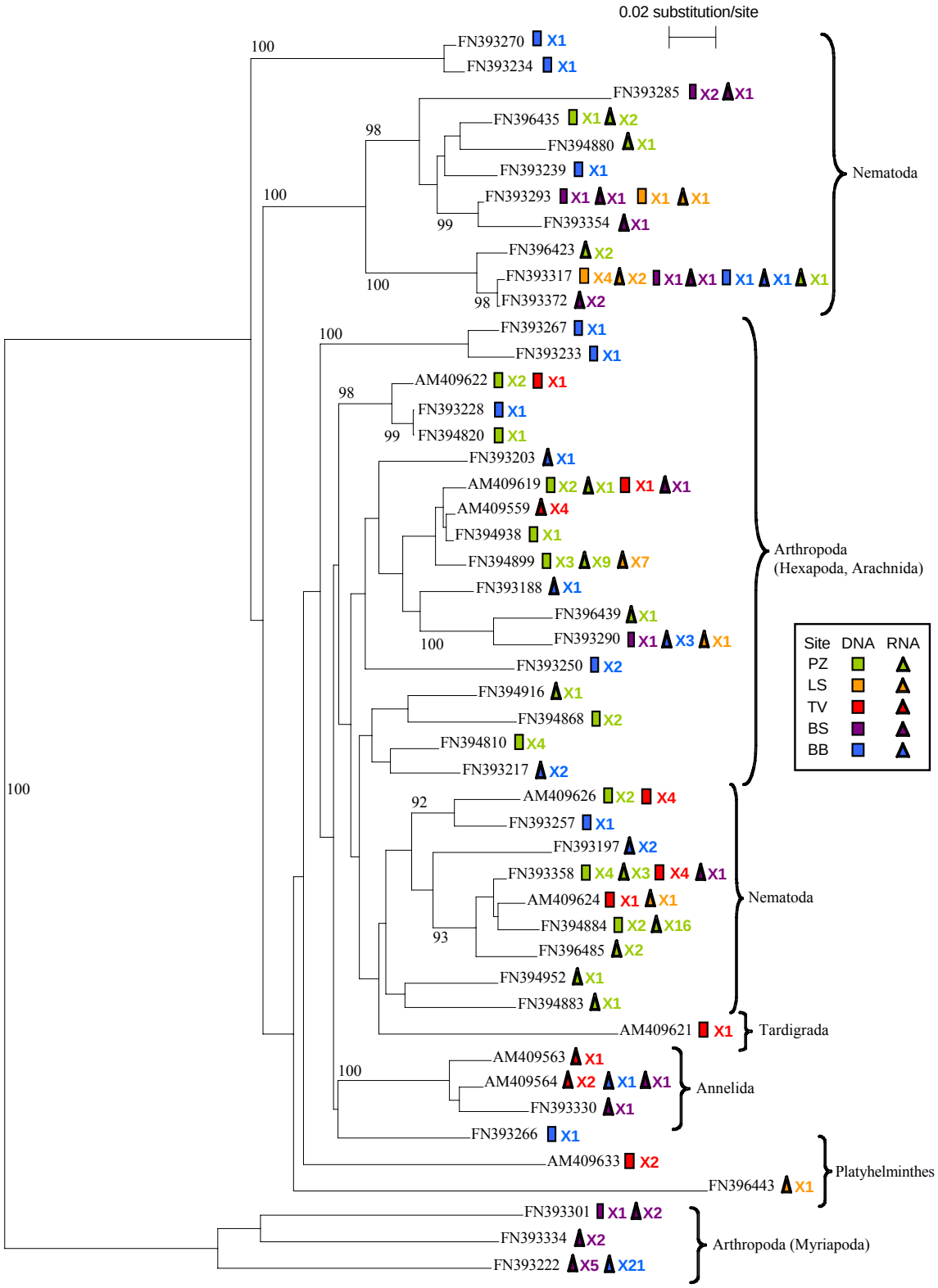
Lehembre et al, Figure 5

FIG. 5. Annotation of the cDNA sequences longer than 400bp from the five forest soils. Annotated proteins correspond to proteins with a known biological function. Conserved hypothetical proteins correspond to proteins of unknown function deposited in public sequence databases. New hypothetical proteins correspond to the fraction of the cDNA which, when searching public sequence databases, did not return any significant Blastx hit.



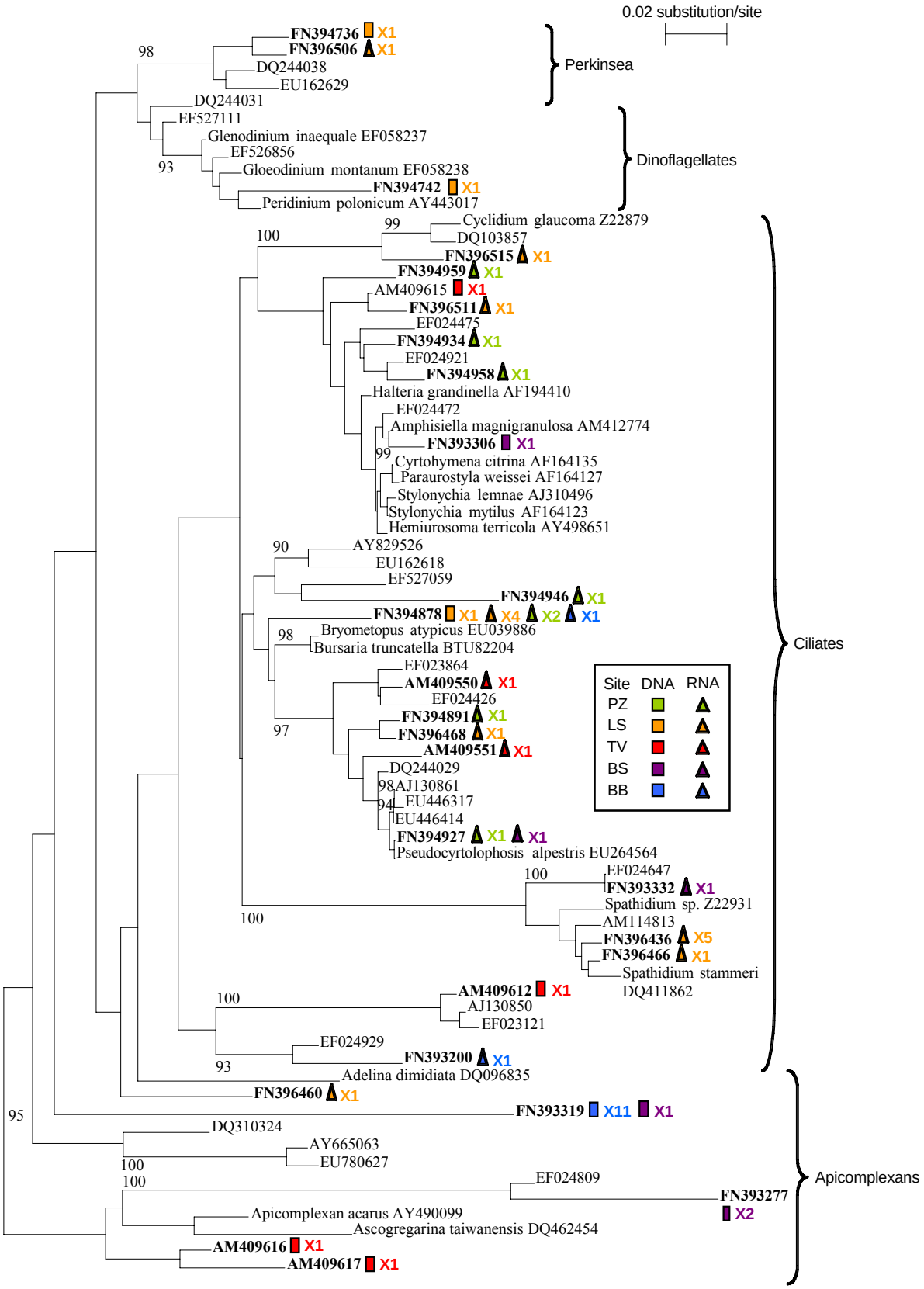
Lehembre et al, Figure S1

FIG. S1. Phylogenetic distribution of environmental partial 18S cercozoa (rhizaria) ribosomal sequences amplified from soil DNA (squares) or reverse-transcribed soil RNA (triangles). The tree was constructed using the BioNJ distance method starting from an alignment which included the different environmental sequences plus their three most similar sequences retrieved from ARB or GenBank. A well-supported cluster (in red) only includes environmental soil sequences from the TV, PZ and LS sites. For an explanation of the symbols, see legend of Fig. 4.



Lehembre et al, Figure S2

FIG. S2. Phylogenetic distribution of partial 18S metazoan ribosomal sequences amplified from soil DNA (squares) or reverse-transcribed soil RNA (triangles). The tree was constructed using the BioNJ distance method starting from an alignment which included the different environmental sequences plus their three most similar sequences retrieved from the ARB database using SINA. A first tree was computed with all sequences to define the different metazoan phyla. Reference sequences were subsequently removed for legibility. Figures give the number of time a sequence was found in a particular dataset. Only bootstrap values (1000 replicates) above 90% are indicated. Variable rates of molecular evolution within the arthropoda and the nematoda could explain sequence grouping in different clusters. The use of PhyML gave an identical topology.



Lehembre et al, Figure S3

FIG. S3. Phylogenetic distribution of environmental partial 18S alveolata ribosomal sequences amplified from soil DNA (squares) or reverse-transcribed soil RNA (triangles). The tree was constructed using the BioNJ distance method starting from an alignment which included the different environmental sequences plus their three most similar sequences retrieved from ARB or GenBank. For an explanation of the symbols, see legend of Fig. 4.

-Chapitre 2-

Identification d'un nouveau groupe de microorganismes eucaryotes

L'objectif de cette partie a été d'effectuer une analyse phylogénétique plus robuste de certaines séquences d'ADNr 18S détectées dans les banques ribosomiques de Truc Vert, Lommel et Paal (cf Chapitre 1) comme étant très divergentes de séquences de cercozoaires présentes dans les bases de données. Ces banques d'ADNr avaient été construites après amplification du premier tiers du gène codant la petite sous unité ribosomale 18S. Par conséquent, le faible nombre de sites informatifs analysés engendre un manque de résolution des analyses phylogénétiques et donc une affiliation incertaine de ces séquences.

Dans ce chapitre, je présente une analyse phylogénétique réalisée sur les résultats de phylogénie moléculaire obtenus avec la totalité du gène 18S qui a permis d'identifier un « nouveau » clade de microorganismes eucaryotes appartenant au groupe des Cercozoa (Rhizaria) ainsi que leur répartition dans différents environnements à travers le monde.

1. Analyse des banques ribosomiques 18S

L'analyse phylogénétique des banques ribosomiques 18S construites avec le premier tiers du gène 18S (cf Chapitre 1) conduit à deux cas de figures illustrés pour le phylum des Rhizaria, exclusivement représenté par des séquences appartenant au groupe des cercozoa (Fig. S1).

La plupart des séquences affiliées à ce phylum sont très similaires aux séquences présentes dans les bases de données et issues pour la plupart d'organismes décrits et répertoriés. Cependant, une minorité de séquences très divergentes des séquences déjà répertoriées ont été obtenues indépendamment et présentent une forte similarité entre elles. Ce résultat permet d'affirmer que ces séquences ne représentent pas des artéfacts de PCR. Pour cette raison nous avons décidé d'étudier de façon plus approfondi ces séquences et le nouveau clade de cercozoaires qu'elles identifient. Pour cela, une première étape a été d'obtenir la séquence 18S complète afin d'affiner nos analyses phylogénétiques.

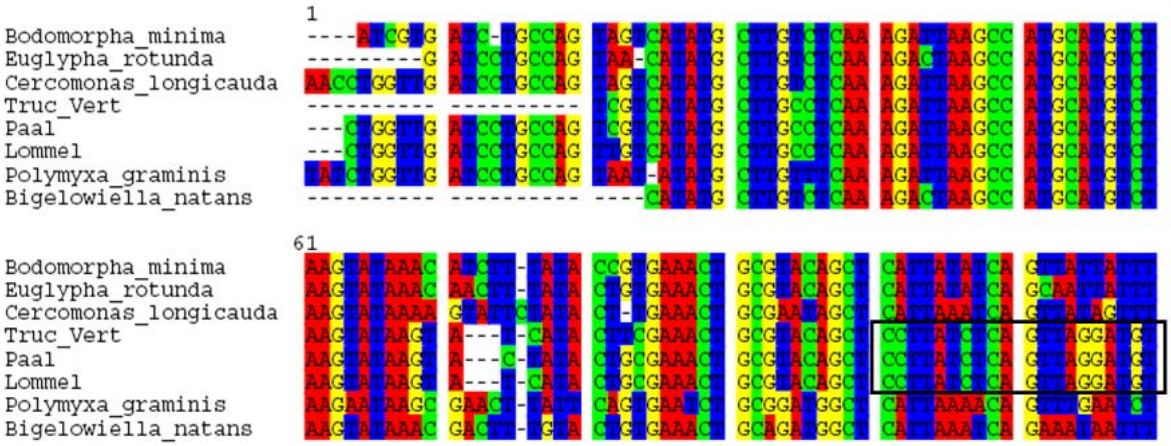


Figure 25: Alignement de séquences du premier tiers du gène 18S obtenues à partir des sols de Lommel, Paal et Truc Vert avec le couple d'amorces Euk1A/Euk516 avec des séquences de quelques cercozoaires répertoriées dans les bases de données. Ces séquences environnementales identifient le nouveau clade de cercozoaires. La séquence encadrée en noir représente l'amorce 5' (Cerco1F) dessinée pour identifier spécifiquement ce nouveau clade.

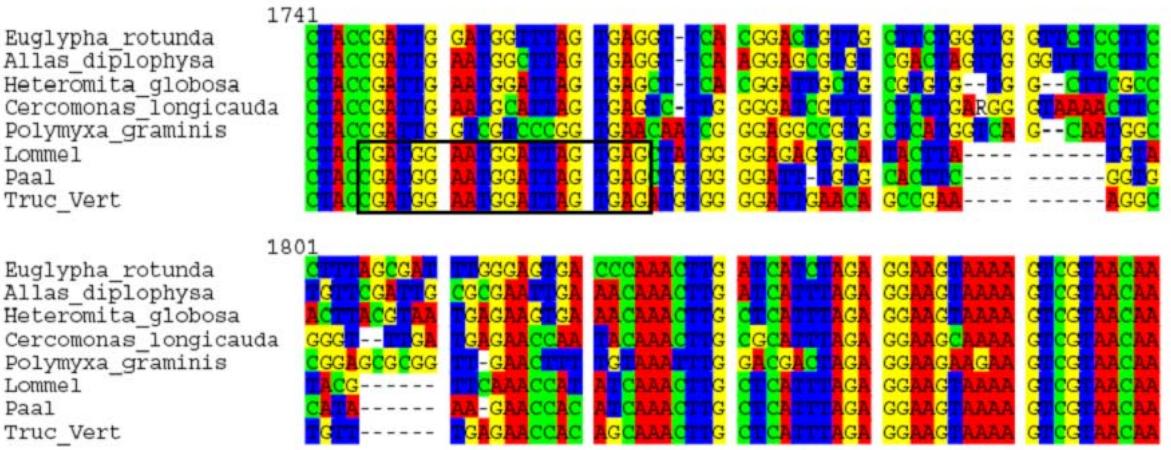


Figure 26: Alignement de séquences de la totalité du gène 18S obtenues à partir des sols de Lommel, Paal et Truc Vert avec le couple d'amorces Cerco1F/18R avec des séquences de quelques cercozoaires répertoriées dans les bases de données. Ces séquences environnementales identifient le nouveau clade de cercozoaires. La séquence encadrée en noir représente l'amorce 3' (Cerco1R) dessinée pour identifier spécifiquement ce nouveau clade.

2. Analyse spécifique de nouvelles séquences ribosomiques 18S

2.1 Dessin des amorces spécifiques

A partir de l'alignement des séquences de référence de cercozoa et des séquences du potentiel « nouveau clade » (Fig. 25), une amorce spécifique (Cerco1F) a pu être dessinée en 5' du gène codant la petite sous unité ribosomale 18S (Tableau 11).

Une amplification par PCR a ensuite été réalisée à partir de l'ADN métagénomique des sols de Lommel, Paal et Truc Vert (Bailly *et al.*, 2007) à l'aide de cette amorce spécifique et d'une amorce universelle (18R) s'hybridant à l'extrémité 3' de l'ADNr 18S (Tableau 11). Des fragments d'environ 1700 pb ont été obtenus pour chaque sol. A l'issue du clonage, 1 à 2 inserts de chaque sol ont été séquencés et, après une analyse phylogénétique, identifiés comme étant spécifiques de ce « nouveau clade ». A partir de l'alignement de ces nouvelles séquences (Fig. 26), une autre amorce spécifique (Cerco1R) a pu être dessinée en 3' du gène codant la petite sous unité ribosomale 18S (Tableau 11).

2.2 Vérification de la spécificité des amorces

Ce couple d'amorces spécifiques (Cerco1F et 1R) a tout d'abord été testé sur les clones contenant les séquences 18S générées avec le couple d'amorces Cerco1F/18R à partir de l'ADN métagénomique des sols de Lommel, Paal et Truc Vert. Chaque séquence amplifiée par PCR avec ce couple d'amorces a été séquencée. Après une analyse phylogénétique, 100% des séquences obtenues se sont bien révélées affiliées à ce « nouveau clade ». Certaines sont identiques entre elles et ont été regroupées, permettant aussi de définir des phylotypes.

Ensuite, une amplification par PCR à partir de l'ADN métagénomique des sols de Lommel, Paal et de Truc Vert (Bailly *et al.*, 2007) à l'aide de ce couple d'amorce spécifique (Cerco1F/Cerco1R) a été réalisée. Aucun signal n'a pu être détecté. Une étape préalable consistant à amplifier la totalité du gène 18S avec le couple d'amorces universelles (18F/18R) a été effectuée. Les séquences 18S obtenues ont ensuite été ré-amplifiées avec le couple d'amorces spécifique (Cerco1F/Cerco1R) par une PCR emboîtée. Un signal a ainsi pu être détecté et comme précédemment, à l'issue du clonage/séquençage des fragments

Séquences cibles	Amorces	Séquences nucléotidiques de l'amorce	Tm (°C)	Références
18S	18F 18R	5'-ACCTGGTTGATCCTGCCAG-3' 5'-TGATCCTTCYGCAGGTTCAC-3'	56	Moon-van der Staay *et al*., 2001
18S	Cerco1F Cerco1R	5'-CCTTATCTCAGTTAGGATGT-3' 5'-CTCACTAATCCATTCCATCG -3'	55	Lehembre

Tableau 11: Séquences des amorces PCR utilisées pour l'amplification universelle du gène 18S de tous les eucaryotes (18F/18R) et pour l'amplification spécifique du nouveau clade de cercozoaires (Cerco1F/Cerco1R).

d'amplification (1500-1600 pb), toutes les séquences amplifiée par PCR avec ce couple d'amorces spécifiques se sont bien révélées comme appartenant à ce « nouveau clade » (Fig. 27).

L'établissement d'une phylogénie moléculaire d'un groupe taxinomique nécessite une connaissance approfondie de ce groupe, notamment pour choisir un jeu de séquences représentatif des différents clades reconnus. Par ailleurs, dans le cas des eucaryotes unicellulaires il existe un grand nombre de séquences non encore publiées et dont la « non inclusion » dans une analyse phylogénétique peut affecter le résultat final. Pour ces raisons, pour la suite des analyses, nous avons pris contact avec deux spécialistes reconnus des Rhizaria (auxquels appartiennent les cercozoaires) : Jan Pawlowski (Université de Genève) et David Bass (Natural History Museum, Londres).

Ils ont confirmé que ces séquences environnementales identifient bien un nouveau clade chez les cercozoaires et constituent un groupe monophylétique (Fig. 28). Par ailleurs, au vu des longues branches qui soutiennent ces séquences, ils supposent que ce groupe d'organismes a évolué rapidement et avancent l'hypothèse d'un groupe d'organismes parasites.

2.3 Recherche dans d'autres environnements

Une fois la spécificité du couple d'amorces et l'authenticité des séquences qu'elles amplifient établie, une amplification par PCR emboîtée a été réalisée à partir des ADN métagénomiques des sols de Breuil (Morvan) prélevés sous hêtre et épicéa (présentés dans le chapitre 1) et à partir de l'ADN métagénomique du sol de Balen pollué aux métaux lourds (présenté dans le chapitre 3). Des séquences affiliées à ce « nouveau clade » ont été identifiées dans chacun de ces 3 sols (Fig. 29).

Afin de connaître plus précisément l'écologie de ce groupe d'organismes, nous avons entrepris une recherche dans des environnements aquatiques. Des ADN extraits d'échantillon d'eau douce et d'eau de mer nous ont été fournis par David Bass. L'amplification par PCR a été réalisée dans les mêmes conditions que précédemment. Des séquences affiliées à ce nouveau clade ont ainsi pu être identifiées dans des échantillons d'eau d'un étang (Priest Pot) et d'un lac (Cumbria) en Angleterre et d'un sol tropical du Pérou (Fig. 29). Aucune séquence appartenant à ce nouveau clade n'a pu être, jusqu'à présent, identifiée dans des échantillons d'eau de mer.

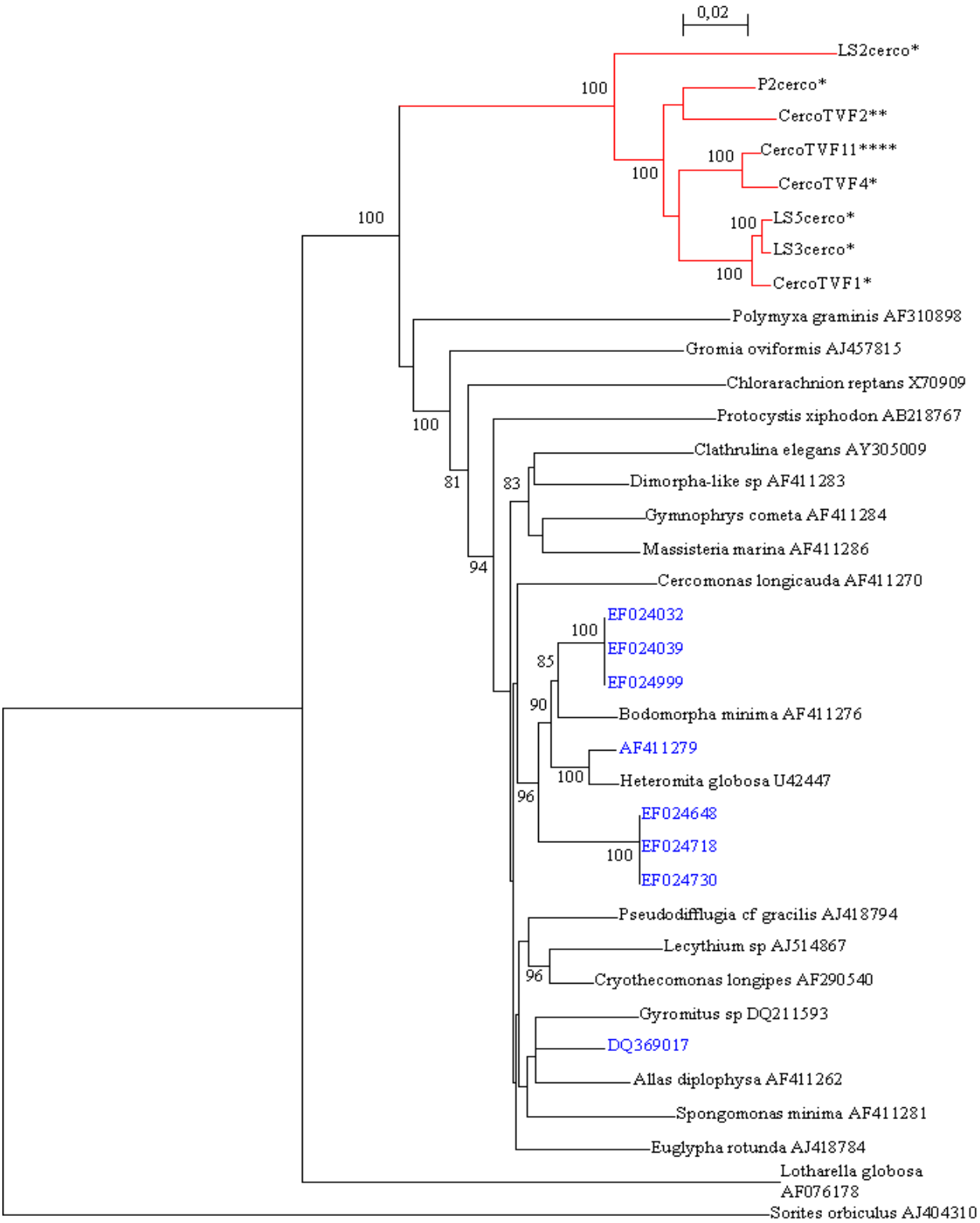


Figure 27: Distribution phylogénétique des séquences 18S totales du nouveau clade de cercozoa (en rouge) amplifiées à partir de l'ADN métagénomique du sol de Lommel (LScerco), Paal (Pcerco) et Truc Vert (CercoTVF). Les astérisques correspondent au nombre de séquences identiques identifiées dans chaque échantillon. L'arbre a été construit en utilisant la méthode du Neighbor Joining et des p-distances à partir d'un alignement qui inclut nos différentes séquences environnementales et leurs 3 séquences les plus similaires retrouvé dans GenBank (en bleu). Seules les valeurs de bootstrap (1000 répliques) supérieur à 80% sont indiquées. L'utilisation de la méthode du maximum de vraisemblance (PhyML) donne une topologie identique.

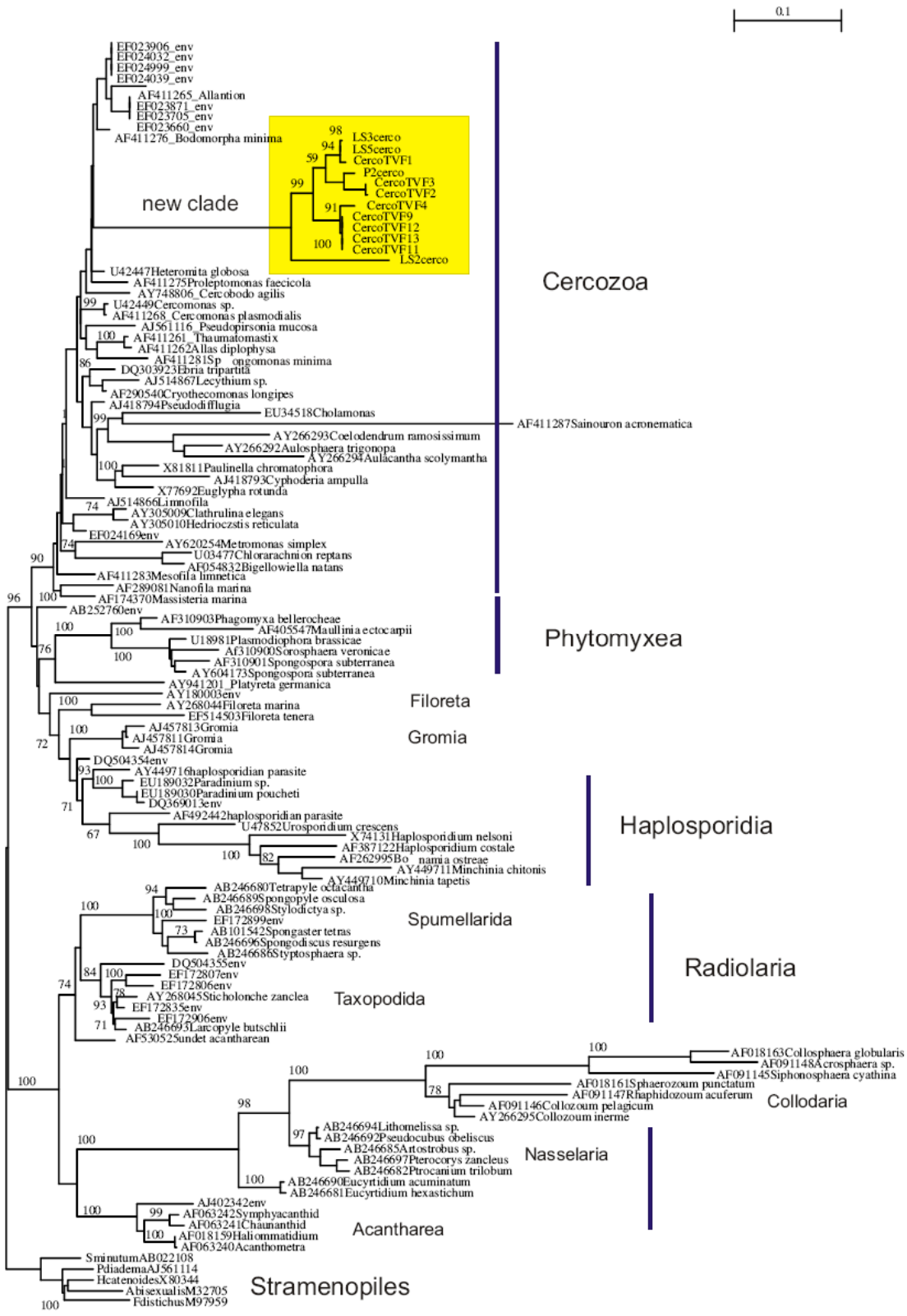


Figure 28: Distribution phylogénétique des séquences 18S totales du nouveau clade de cercozoa (en jaune) amplifiées à partir de l'ADN métagénomique du sol de Lommel (LScerco), Paal (Pcerco) et Truc Vert (CercoTVF) à l'aide du couple d'amorces Cerco1F/Cerco1R. L'arbre a été construit à partir d'un alignement qui inclut nos séquences environnementales, des séquences de cercozaoires référencés et des séquences non publiées répertoriées par Jan Pawlowski.

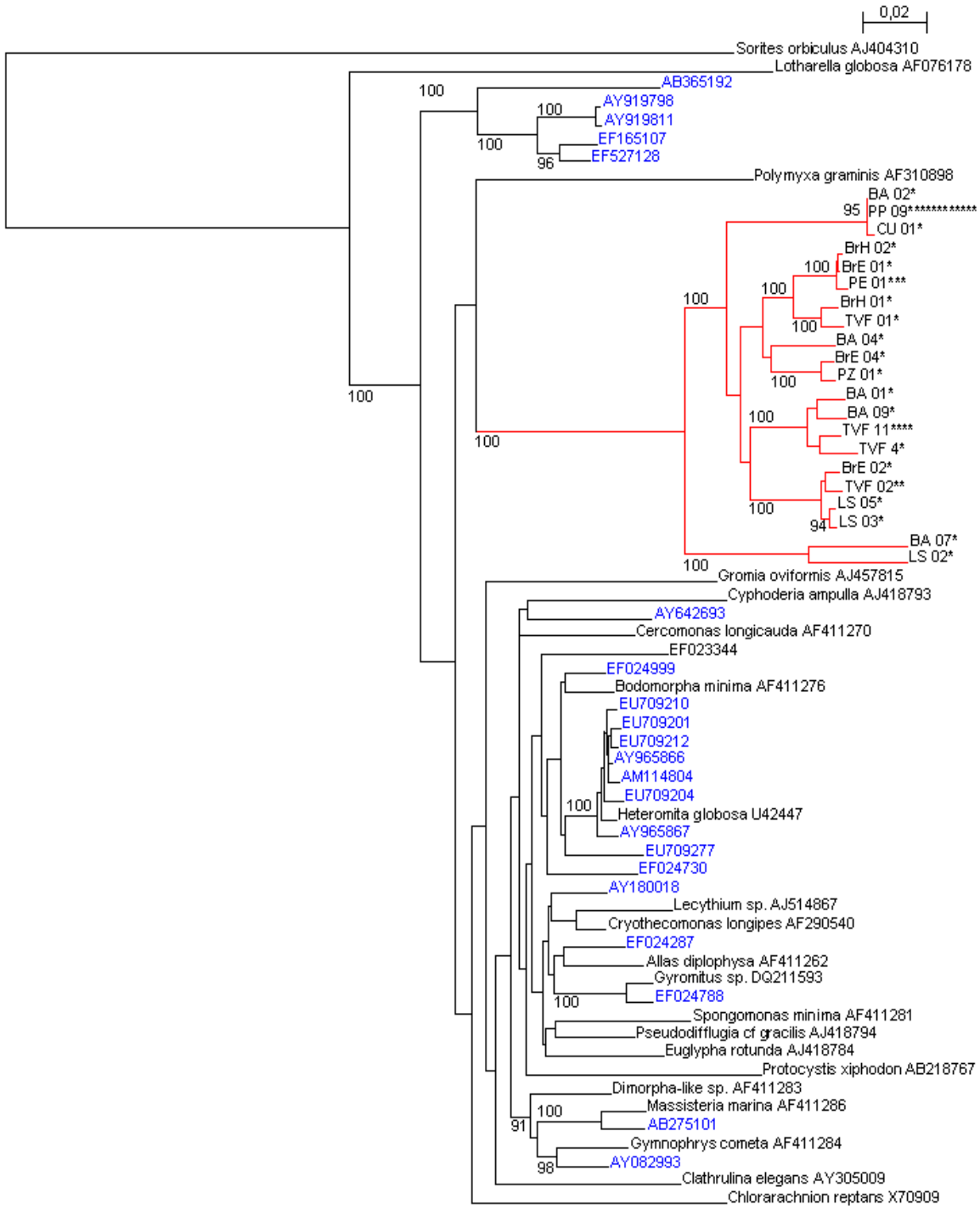


Figure 29: Distribution phylogénétique des séquences 18S totales du nouveau clade de cercozoa (en rouge) amplifiées à partir de l'ADN métagénomique du sol de Lommel (LS), Paal (PZ), Truc Vert (TVF), Breuil sous hêtre (BrH), Breuil sous épicéa (BrE), Balen (BA) et du Pérou (PE) et à partir d'ADN d'échantillon d'eau de Priest Pot (PP) et de Cumbria (CU). Les astérisques correspondent au nombre de séquences identiques identifiées dans chaque échantillon. L'arbre a été construit en utilisant la méthode du Neighbor Joining et des p-distances à partir d'un alignement qui inclut nos différentes séquences environnementales et leurs 3 séquences les plus similaires retrouvé dans GenBank (en bleu). Seules les valeurs de bootstrap (1000 répliques) supérieur à 90% sont indiquées. L'utilisation de la méthode du maximum de vraisemblance (PhyML) donne une topologie identique.

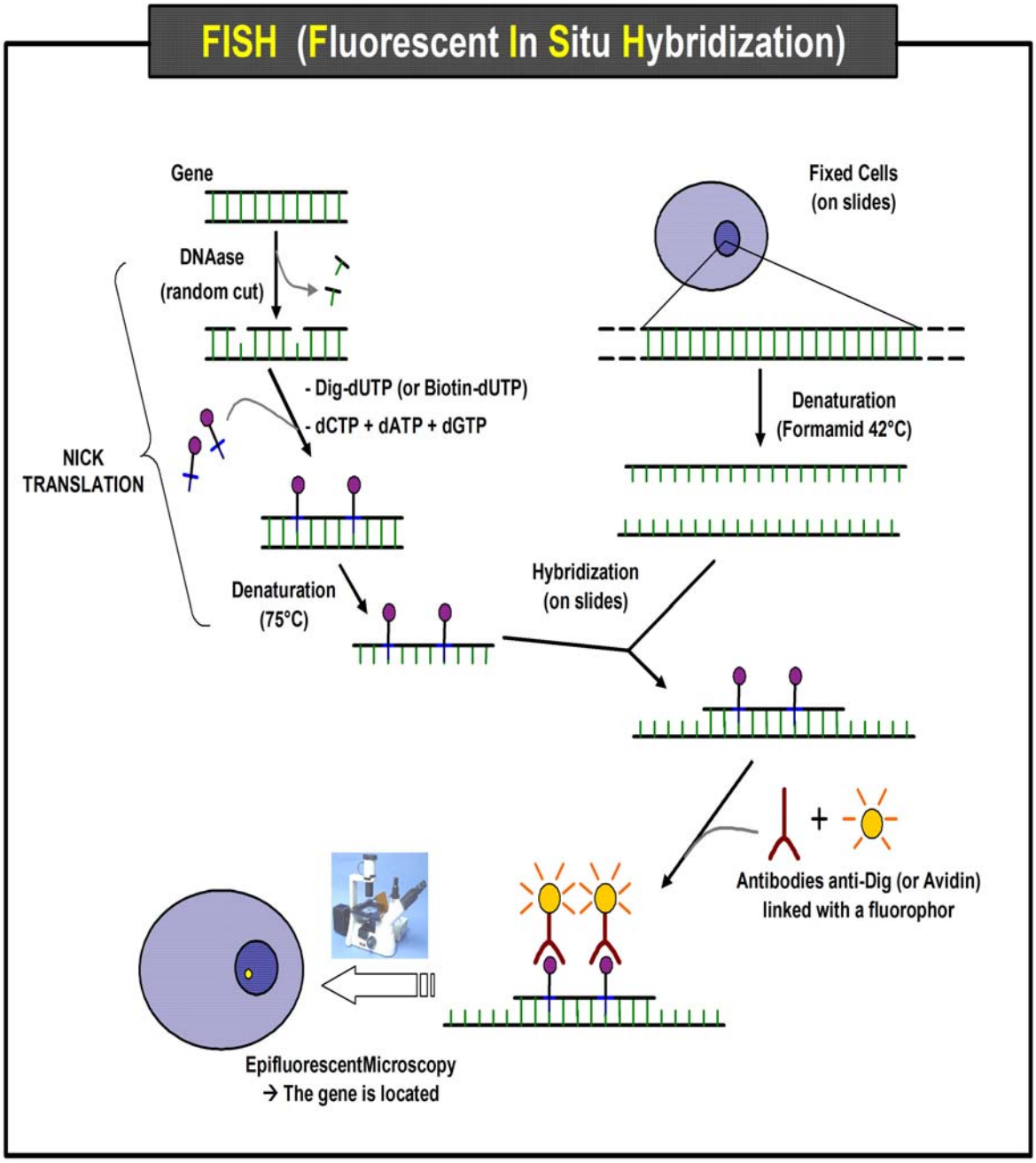


Figure 30: Schéma représentant les différentes étapes de l'hybridation fluorescente *in situ* ou FISH.

L'arbre phylogénétique obtenu révèle une grande diversité d'organismes au sein de ce nouveau clade. Au final, sur les 38 séquences obtenues, 21 phylotypes ont pu être définis sur une base d'identité de séquences de 100%. Le regroupement des séquences sur une base de 99% d'identité entre elles permet de définir 16 phylotypes distincts. Les deux séquences identifiées dans l'eau et une séquence identifiée dans le sol pollué de Balen se regroupent dans un même phylotype. Par ailleurs, deux phylotypes regroupent au moins deux séquences provenant de 2 sols distants géographiquement (Pérou et Breuil pour l'un, Breuil et Truc Vert pour l'autre). Tous les autres phylotypes sont constitués de séquences provenant d'un seul lieu.

Au vue des résultats obtenus, une étude supplémentaire d'identification morphologique de ce groupe d'organismes à été suggérée. Une technique permettant d'identifier un microorganisme non cultivable dans son environnement est l'hybridation fluorescente *in situ* (ou FISH: Fluorescent *In Situ* Hybridization) de sondes dessinées à partir de la séquence d'ARNr (Moter et Gobel, 2000) (Fig. 30). Cependant, cette technique est difficilement réalisable sur une matrice complexe telle que le sol. Les expériences de FISH sur des échantillons d'eau sont actuellement en cours dans le laboratoire de David Bass.

3. Discussion

Ces dernières années, le nombre d'études s'intéressant à la diversité microbienne eucaryote dans différents environnements aquatiques n'a cessé d'augmenter (par exemple, Moon-van der Staay *et al.*, 2001; Lopez-Garcia *et al.*, 2001; Moreira et Lopéz-Garcia, 2002; Stoeck *et al.*, 2003). Toutes ces études, basées sur la construction et l'analyse de banques de gènes 18S environnementaux, ont révélé une diversité de lignées de microorganismes eucaryotes jusqu'ici inconnues. Dans les sols, les études moléculaires de la diversité globale des microorganismes eucaryotes sont plus rares (Moon-van der Staay *et al.*, 2006 ; Bailly *et al.*, 2007 ; Lesaulnier *et al.*, 2008). Elles sont souvent restreintes à des groupes taxinomiques précis comme les champignons (O'Brien *et al.*, 2005; Fierer *et al.*, 2007; Yergeau *et al.*, 2007), les nématodes (Yergeau *et al.*, 2007) et les cercozoa (Bass et Cavalier-Smith, 2004). L'emploi d'amorces spécifiques à certains groupes a permis de mettre en évidence de nombreuses nouvelles espèces et montre que la diversité et l'abondance des microorganismes eucaryotes dans les sols est encore largement sous-estimée.

L'analyse globale de nos banques de gènes 18S (cf Chapitre 1 Résultats) a révélé une grande diversité eucaryote principalement représentée par des séquences de champignons et d'animaux (micro/méso-faune). Pour certains de nos sols, un nombre important de séquences d'ADNr 18S affiliées à certains groupes de protistes comme les rhizaria (exclusivement représenté par le groupe des cercozoa) et les alveolata (principalement représenté par le groupe des ciliata) ont également été retrouvées. Ces groupes de protistes sont connus pour être les plus abondants dans les sols (Ekelund et Patterson, 1997) et ont une importance écologique majeure. Ils sont notamment connus pour être des prédateurs de bactéries et de champignons (Ekelund, 1998) et pour être les proies des nématodes et des vers de terres (Ekelund et Patterson, 1997).

Les cercozoa ont été le premier groupe d'organismes majeurs à être identifié par un résultat de phylogénie moléculaire avant d'être caractérisé morphologiquement (Cavalier-Smith, 1996/7). Ils sont décrits comme des flagéllés hétérotrophes ou des amibes à tests présentant une grande diversité, estimée entre 1000 et 10.000 lignées distinctes (Bass et Cavalier-Smith, 2004) soit autant que certains groupes de protistes mieux caractérisés comme les ciliés et les foraminifères (Finlay, 2002). Les études visant à mieux caractériser leur diversité sont basées sur l'utilisation d'amorces spécifiques au groupe. Elles ont permis de construire des banques de gènes 18S spécifiques aux cercozoa à partir de différents environnements, principalement aquatiques, et d'identifier de nombreuses nouvelles lignées (par exemple, Bass et Cavalier-Smith, 2004; Bass *et al.*, 2009).

Notre étude de phylogénie moléculaire a également permis d'identifier de nouvelles lignées de cercozoa dans différents environnements, principalement des sols, à l'aide d'amorces spécifiques définies à partir de séquences divergentes obtenues par une analyse globale de la diversité eucaryotes dans les sols. La construction d'arbres phylogénétiques suivant une méthode phénétique (NeighBor Joining,) et suivant une méthode probabiliste (Maximum de vraisemblance) a démontré que ces séquences environnementales identifient bien un nouveau clade chez les cercozoa et constituent un groupe monophylétique. Par ailleurs, au vue des longues branches qui soutiennent ces séquences, il semble que ce groupe d'organisme ait évolué rapidement et l'hypothèse d'un groupe d'organismes parasites peut être avancée. En effet, les cercozoa sont également connus pour être des endoparasites d'invertébrés d'eau douce et marine, de plantes et d'algues (Cavalier-Smith et Chao, 2003).

Enfin, la répartition des séquences de ce nouveau clade, mais également d'autres clades (Bass et Cavalier-Smith, 2004), dans différents environnements et/ou dans des sites géographiquement distants ne permet pas d'établir une distribution écologique et

géographique pour ce groupe d'organismes. Des efforts d'échantillonnage et de séquençage sont donc à effectuer afin de mieux caractériser leur diversité et leur écologie, notamment dans les sols où une diversité cryptique semble exister. Dernièrement, une expérience d'hybridation fluorescente *in situ* (FISH), de sondes nucléotidiques définies à partir de séquences environnementales de cercozoa, a permis d'identifier des cercozoa capables de se nourrir de pico et de nano-plancton ou de parasiter des diatomées (Mangot *et al.*, 2009).

-Chapitre 3-

Le métatranscriptome eucaryote d'un sol pollué et non pollué aux métaux lourds

L'objectif de cette partie est d'analyser et de comparer la diversité taxinomique et fonctionnelle d'une communauté de microorganismes eucaryote présente dans des échantillons de sols prélevés sur 3 sites distincts. Il s'agit d'un sol contaminé par des métaux lourds (site de Balen), anciennement contaminé par ces mêmes métaux lourds (site de Lommel) et non contaminé (site de Paal). Quelques résultats obtenus à partir du sol anciennement contaminé (Lommel) et non contaminé (Paal) ont été décrits dans la partie « résultats chapitre 1 ».

Pour atteindre cet objectif, nous avons évalué la diversité taxinomique eucaryote par l'analyse des séquences d'ADNr et d'ARNr 18S suivant une approche de clonage et de séquençage. La diversité fonctionnelle a été estimée par l'analyse de banques d'ADNc construites à partir des ARNm polyadénylés directement extraits des échantillons de sols suivant une approche métatranscriptomique. Ces banques d'ADNc ont fait l'objet d'un séquençage aléatoire (approche basée sur la séquence) et d'un criblage par expression hétérologue dans la levure (approche basée sur la fonction).

1. Sites étudiés

Les sites de Balen, Paal et Lommel sont situés dans un périmètre d'environ 60 km dans la Province du Limbourg en Belgique. Les échantillons de sol prélevés sur ces sites ont une origine géologique commune et ont tous été récoltés à proximité de l'espèce végétale *Pinus sylvestris*, commune aux 3 sites. Il s'agit de sols sableux qui diffèrent par leur pH, leur capacité d'échange cationique, leur humidité et leur taux de matière organique. Ces paramètres physico-chimiques sont globalement plus élevés pour le site de Balen, sur lequel une végétation (herbacées, bryophytes) plus importante a été observée (cf Tableau 22 et 23 Matériels et méthodes). Ces sites diffèrent également par leur concentration en métaux lourds.

Le site de Balen, situé à proximité de la plus grande fonderie de zinc européenne toujours en activité, présente des concentrations totales, disponibles et biodisponibles en zinc, cadmium et plomb plus élevées que celles dosées sur les sites de Lommel et Paal. Ce site fait l'objet d'études de phytorémédiation depuis une dizaine d'années (van der Lelie *et al.*, 2001) et représente donc un site contaminé d'intérêt pour notre étude.

Le site de Lommel est situé à proximité d'une ancienne fonderie de zinc construite à la fin du 19ème siècle et démantelée dans les années 1960. Ce site, reboisé en 1975, est considéré comme anciennement contaminé car une contamination équivalente à celle de Balen était

encore observée en 1998 (Jan Colpaert, communication personnelle) mais n'est plus présente aujourd'hui sans doute par « lessivage » des métaux.

Le site de Paal est relativement éloigné de toute industrie métallurgique et son sol possède des caractéristiques physico-chimiques proches de celles de Lommel, excepté le taux de matière organique. Toutefois, la concentration biodisponible en cadmium dosée dans les aiguilles de pins à Paal est supérieure à celle communément retrouvée dans les aiguilles de pins de sites non contaminés (en général inférieure à 0,5 ppm) (Dmuchowski et Bytnerowicz, 1995 ; Yilmaz et Zengin, 2004) alors qu'aucune trace de cadmium n'a été détectée dans les échantillons de ce sol. Ce résultat peut s'expliquer par l'acidité de ce sol, qui facilite le passage des métaux sous formes ioniques, et par la capacité des pins à accumuler les métaux. Il y a une dizaine d'années, les sites de Lommel et de Paal (site témoin) ont fait l'objet de nombreuses études écotoxicologiques par le centre de sciences environnementales de l'université de Hasselt (Belgique) (Vangronsveld *et al.*, 1995 ; Vangronsveld *et al.*, 1996).

2. Etude de la diversité taxinomique

Après avoir vérifiée la qualité des différentes banques ribosomiques construites, une comparaison de la diversité taxinomique a été effectuée suivant différents paramètres : le sol étudié, la portion du gène 18S ciblée et la molécule (ADN ou ARN) analysée. La richesse taxinomique des eucaryotes a également était estimée et comparée pour ces 3 variables.

2.1 Construction des banques ribosomiques 18S eucaryotes

La diversité taxinomique des eucaryotes du sol contaminé aux métaux lourds (Balen) et du sol non contaminé (Paal) a été estimée par l'analyse du premier et du dernier tiers de la séquence du gène codant l'ARNr 18S. La diversité taxinomique du sol anciennement pollué (Lommel) a été évaluée uniquement par l'analyse du premier tiers du 18S (cf Chapitre 1). Pour chaque sol et chaque portion du gène 18S analysé, une banque d'ADNr et une banque d'ARNr rétrotranscrit ont été construites et analysées.

2.1.1 Les banques d'ADN ribosomiques

Lors de l'extraction des acides nucléiques de chacun des sols, une partie aliquote de 100µL a été prélevée avant l'étape de digestion à la DNaseI. Les ADN contenus dans ces fractions ont été purifiées sur colonne Sephadex G50 puis re-précipitées. A partir de l'ADN ainsi obtenu, une à deux portions (suivant le sol) du gène codant l'ARNr 18S des Eucaryotes a été amplifiée par 25 cycles de PCR à l'aide des amorces Euk1A-Euk516 (premier tiers du gène) et S12.2-SB (dernier tiers du gène) (cf Matériels et méthodes, Tableau 24, Fig. 51). Les fragments de 500 pb (premier tiers) et de 700 pb (dernier tiers) obtenus et purifiés à partir de 8 répétitions de la réaction de PCR ont été utilisés pour construire les banques d'ADN ribosomiques dans le vecteur plasmidique pCR4Blunt-TOPO (Invitrogen).

2.1.2 Les banques d'ARN ribosomiques

Les banques d'ARN ribosomiques ont été construites selon le même principe. Les ARNr recueillis lors de l'étape de purification des ARNm sur billes magnétiques ont été soumis à une étape de transcription inverse. Trois répétitions de cette réaction ont été réalisées en présence de 500 ng d'ARNr avec l'amorce Euk516, ou l'amorce SB ou un mélange d'hexanucléotides aléatoires comme oligonucléotide d'amorçage. Pour chaque condition, les produits des 3 répétitions ont été rassemblés et 2% de la quantité finale ont été utilisés pour chaque répétition d'amplification par PCR. Une à deux portions (suivant le sol) de la molécule d'ARNr rétrotranscrite a été amplifiée par 25 cycles de PCR avec les amorces Euk1A-Euk516 (premier tiers) et S12.2-SB (dernier tiers). Les fragments de 500 pb (premier tiers) et de 700 pb (dernier tiers) obtenus et purifiés à partir de 8 répétitions de la réaction de PCR ont été utilisés pour construire les banques d'ADN ribosomiques dans le vecteur plasmidique.

2.2 Vérification de la qualité des banques

2.2.1 Vérification des inserts

A l'issue du clonage, les séquences d'ADNr et d'ARNr obtenues ont été comparées aux séquences de la base de données GenBank à l'aide du programme BLASTn

Site		ADNr 5’	ARNr 5’	ADNr 3’	ARNr 3’
Balen	Nombre de séquences initiales	**274**	**275**	**246**	**244**
	Nombre d’artéfacts	3	1	1	2
	Nombre de chimères	5	1	3	3
	Nombre de séquences finales	266	273	242	239
Paal	Nombre de séquences initiales	**348**	**371**	**273**	**271**
	Nombre d’artéfacts	14	7	0	14
	Nombre de chimères	1	6	4	3
	Nombre de séquences finales	333	358	269	254
Lommel	Nombre de séquences initiales	**93**	**95**	**nd**	**nd**
	Nombre d’artéfacts	22	2	nd	nd
	Nombre de chimères	0	1	nd	nd
	Nombre de séquences finales	71	92	nd	nd

Tableau 12: Analyse des séquences 18S: nombre de séquences obtenues, nombre d’artéfacts et de chimères, et nombre de séquences finales utilisées pour les analyses de diversité. La taille moyenne de chaque jeu de séquences est indiquée en paires de bases (pb). Le terme 5’ désigne le premier tiers de l’ARN ribosomique 18S et 3’ le dernier tiers.
nd : non déterminé

(http://www.ncbi.gov/BLAST) et aux séquences de la base de données Silva SSU par l'intermédiaire du serveur MG-RAST (http://metagenomics.nmpdr.org/). Les séquences ne présentant pas d'identité avec des séquences ribosomiques 18S ont été considérées comme des artéfacts et éliminées des jeux de séquences.

2.2.2 Détection des séquences chimériques

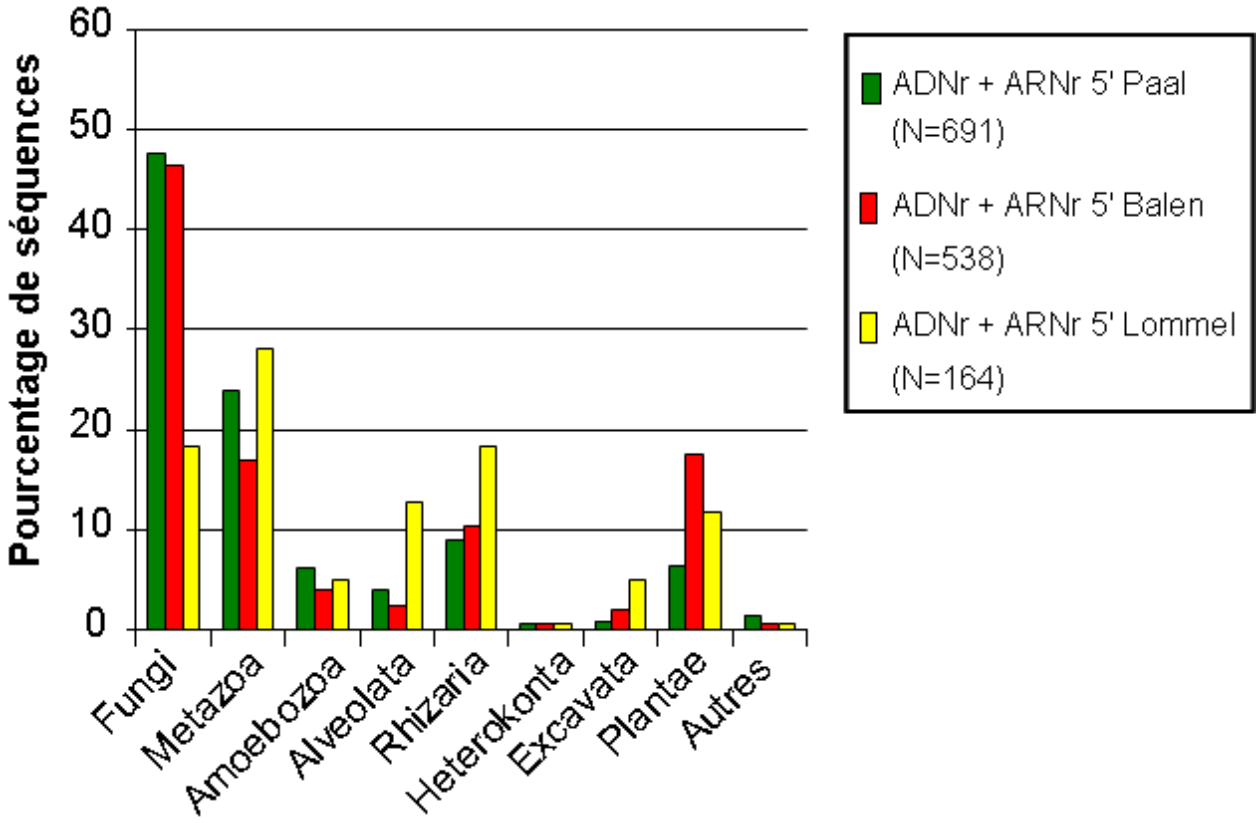
Les séquences chimériques sont des artéfacts de la PCR qui sont constituées en général par l'association d'au moins deux séquences provenant d'organismes différents. Il est nécessaire d'identifier ces chimères qui contribuent à augmenter artificiellement la richesse phylogénétique et spécifique des communautés microbiennes étudiées. Compte tenu du fait que les séquences ribosomiques 18S sont fortement conservées et que nos banques ont été construites à partir de produits de PCR, nous avons vérifié l'éventuelle présence de telles séquences au sein de chaque banques à l'aide du logiciel Check Chimera (http://rdp8.cme.msu.edu/cgis/chimera.cgi?su=SSU) du « ribosomal database project II ». Les séquences chimériques identifiées ont été éliminées des jeux de séquences pour la suite des analyses (tableau 12).

2.3 Répartition au sein des phyla eucaryotes

Chaque séquence a été affiliée à un des 8 grands phyla eucaryotes définis par Lopez-Garcia et Moreira (2008). Il s'agit des fungi, des metazoa, des plantae, des amoebozoa, des alveolata, des heterokonta, des rhizaria et des excavata (les 5 derniers groupes peuvent être rassemblés sous le terme « protistes »).

A l'issue du BLAST contre la base de données GenBank, les séquences présentant plus de 95% d'identité avec une séquence appartenant à un organisme isolé ont été directement classées dans le phylum auquel appartient cet organisme.

Les séquences présentant moins de 95% d'identité avec un organisme isolé ou présentant une plus ou moins forte identité avec une séquence environnementale ont fait l'objet d'analyses phylogénétiques. Pour cela nous avons incorporé des séquences provenant d'organismes représentatifs des grands phyla eucaryotes. Ces séquences proviennent généralement d'études de phylogénie moléculaire faisant référence dans leur domaine, ceci



	Fungi	Metazoa	Amoebozoa	Alveolata	Rhizaria	Heterokonta	Excavata	Plantae	Autres
CHI-2 Balen/Paal	**b**	**a**	**b**	**b**	**b**	**b**	**b**	**a**	**b**
CHI-2 Balen/Lommel	**a**	**a**	**b**	**a**	**a**	**b**	**a**	**b**	**b**
CHI-2 Paal/Lommel	**a**	**b**	**b**	**a**	**a**	**b**	**a**	**a**	**b**

Figure 31: Affiliation taxinomique des séquences représentant le premier tiers (5') de l'ADNr et de l'ARNr ribosomiques 18S obtenues à partir des sols de Paal, Balen et Lommel. Le nombre de séquences (N) correspond à la somme des ADNr et des ARNr 18S. La catégorie « autres » inclut des séquences 18S appartenant à des groupes taxinomiques mineurs tels ques les choanozoa, les ichtyosporeae ou les nucleariidae. Un test de Chi-2 comparant l'abondance de chacun des phylums entre chaque site d'étude a été effectué ; les valeurs numériques exactes sont indiquées en annexe.
La lettre (a) signifie que la différence entre les 2 jeux de données est significative ($p<0,05$). La lettre (b) signifie que la différence entre les 2 jeux de données n'est pas significative ($p\geq0,05$).

pour minimiser les risques d'erreur d'identification de phyla. A partir de ce nouveau jeu de séquences, des arbres phylogénétiques ont été construits. L'alignement des séquences a été réalisé avec le logiciel ClustalW (http://www.ebi.ac.uk/) et les arbres ont été construits pour chaque phylum avec le logiciel PhyloWin en utilisant le modèle du Neighbour Joining et les « p-distances ».

Malgré ces différentes précautions, un certain nombre de séquences n'ont pas pu être affiliées à un phylum avec certitude. Il est probable que ces problèmes résultent du fait que nous analysons des séquences courtes (500 et 700 pb) qui ne contiennent pas toute l'information nécessaire à la séparation claire de certains phyla eucaryotes. L'attribution de certaines séquences à un phylum donné doit donc parfois être considérée comme conditionnelle.

2.3.1 Analyse globale du premier tiers (en 5') de l'ARN ribosomique 18S

Les séquences d'ADNr et d'ARNr disponibles pour les 3 sols étudiés ont tout d'abord été étudiées dans leur ensemble (Fig. 31).

Tous les phyla eucaryotes sont représentés dans les jeux de séquences des 3 sites.

A l'exception des séquences affiliées aux Metazoa et aux Plantae, les profils d'affiliations taxinomiques du site de Balen, contaminé par des métaux lourds, et du site de Paal, non contaminé, ne présentent pas de différences significatives. Pour ces 2 sites, les jeux de séquences sont dominés par les séquences de champignons et d'animaux (Fig. 31). Les groupes de protistes les mieux représentés sont les rhizaria (exclusivement représentés par les cercozoa), les amoebozoa et les alveolata.

Le profil d'affiliation taxinomique des séquences ribosomiques de Lommel présente plus de différences significatives avec les profils de Paal et de Balen (Fig. 31). Le jeu de séquences du site de Lommel est légèrement dominé par les séquences d'animaux. La principale différence par rapport aux 2 autres sites concerne le faible nombre de séquences appartenant aux champignons équivalent au nombre de séquences appartenant aux phyla des rhizaria et des alveolata, apparemment mieux représentées dans ce site.

Les séquences de plantes obtenues appartiennent majoritairement au genre *Pinus* présent sur les 3 sites d'études. Elles reflètent la quantité des racines fines qui sont passées à travers le tamis lors de l'échantillonnage.

Après avoir exclu les séquences de plantes des jeux de séquences, un regroupement des séquences présentant 97% d'identité entre elles, considérées comme appartenant à un même

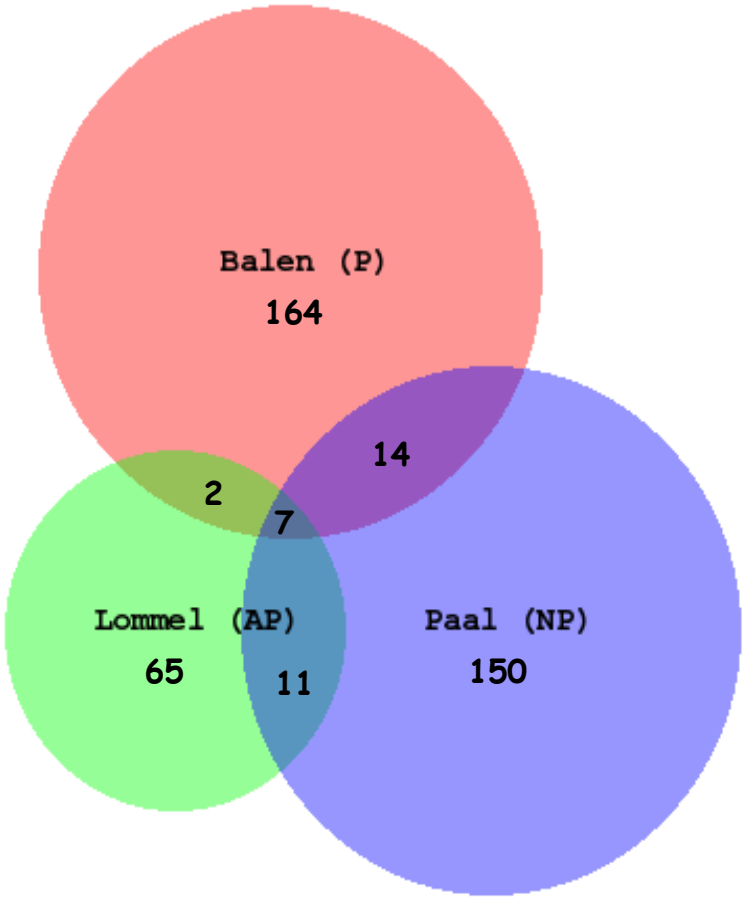


Figure 32: Nombres de phylotypes définis par les séquences d'ADNr et d'ARNr ribosomiques 18S 5' présentant 97% d'identité entre elles, spécifiques de chacun des sites et communs à 2 ou 3 sites. Le site de pollué (P) de Balen est représenté en rose, le site anciennement pollué (AP) de Lommel en vert et le site non pollué (NP) de Paal en bleu.

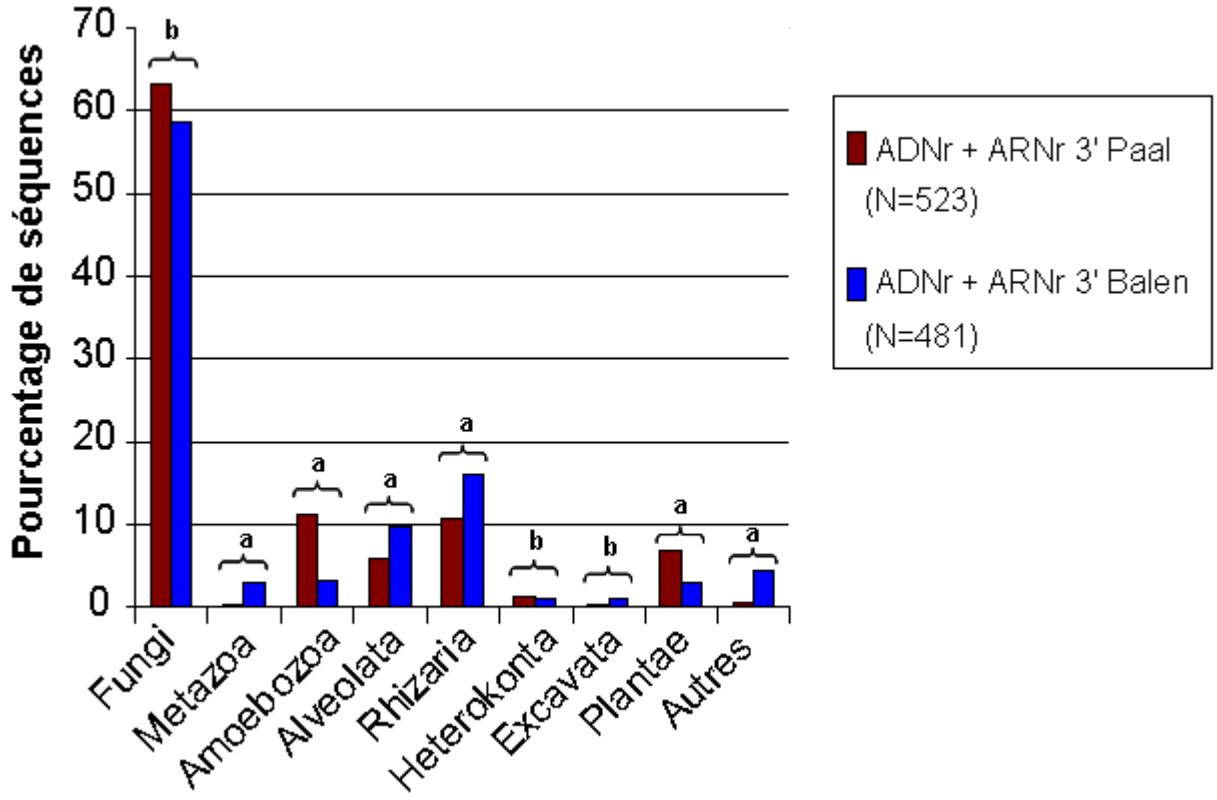


Figure 33: Affiliation taxinomique des séquences représentant le dernier tiers (3') de l'ADNr et de l'ARNr ribosomiques 18S obtenues à partir des sols de Paal et Balen. Le nombre de séquences (N) correspond à la somme des ADNr et des ARNr 18S. La catégorie « autres » inclut des séquences 18S appartenant à des groupes taxinomiques mineurs tels que les choanozoa, les ichtyosporeae ou les nucleariidae. Un test de Chi-2 comparant l'abondance de chacun des phylums entre chaque site d'étude a été effectué ; les valeurs numériques exactes sont indiquées en annexe.
La lettre (a) signifie que la différence entre les 2 jeux de données est significative ($p<0,05$). La lettre (b) signifie que la différence entre les 2 jeux de données n'est pas significative ($p\geq0,05$).

phylotype, a été effectué pour chaque jeu de séquences à l'aide du logiciel RapidOTU (http://genome.jouy.inra.fr/rapidotu/). Nous avons ensuite déterminé le nombre de phylotypes communs entre les différents sites en réalisant un diagramme de Venn à l'aide du logiciel BioVenn (http://www.cmbi.ru.nl/cdd/biovenn/) (Fig. 32). La majorité des phylotypes (92%) est spécifique de chaque site. Seul 1,7% des phylotypes sont communs aux 3 sites. Ces derniers appartiennent aux phylums des champignons (3 phylotypes) et des metazoa (4 phylotypes).

2.3.2 Analyse globale du dernier tiers (en 3') de l'ARN ribosomiques 18S

Pour cette portion, nous disposons seulement des séquences d'ADNr et d'ARNr pour les sols de Balen et de Paal (Fig. 33).

Comme précédemment, tous les phyla eucaryotes sont représentés dans les jeux de séquences des deux sites. Les profils d'affiliations taxinomiques présentent plus de différences significatives que précédemment (Fig. 33). Les jeux de séquences sont très nettement dominés par les séquences de champignons. Les groupes de protistes les mieux représentés sont à nouveau les rhizaria (exclusivement représenté par les cercozoa), les amoebozoa et les alveolata. Ces deux derniers phyla semblent être mieux représentés que dans les jeux de séquences du premier tiers de l'ARN ribosomique. Cependant, la principale différence avec l'analyse précédente concerne la quasi disparition des séquences affiliées au phylum des metazoa.

Les séquences de plantes obtenues appartiennent aussi majoritairement au genre *Pinus* présent sur les sites d'études.

Après avoir exclu les séquences de plantes et regroupé les séquences présentant 97% d'identité entre elles en phylotype, la construction d'un diagramme de Venn a permis de mettre en évidence que 7,5% des phylotypes était communs au deux sites contre 6% pour les phylotypes « 5' » (Fig. 34). Parmi ces phylotypes, 61,5% sont d'origine fongique (contre 62% en 5'), 23% appartiennent au phylum des rhizaria (contre 9,5% précédemment) et 15,5% sont affiliés aux alveolata (contre 0% en 5').

2.3.3 ADN versus ARN

La comparaison des jeux de séquences d'ADNr et d'ARNr 18S pour chaque phylum met en évidence des différences significatives et des similitudes suivant le site et la portion du

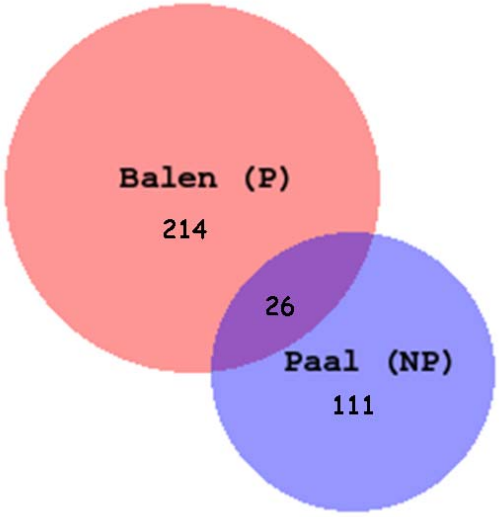


Figure 34: Nombres de phylotypes, définis par les séquences d'ADNr et d'ARNr ribosomiques 18S 3' présentant 97% d'identité entre elles, spécifique de chacun des sites et communs aux 2 sites étudiés. Le site de pollué (P) de Balen est représenté en rose et le site non pollué (NP) de Paal en bleu.

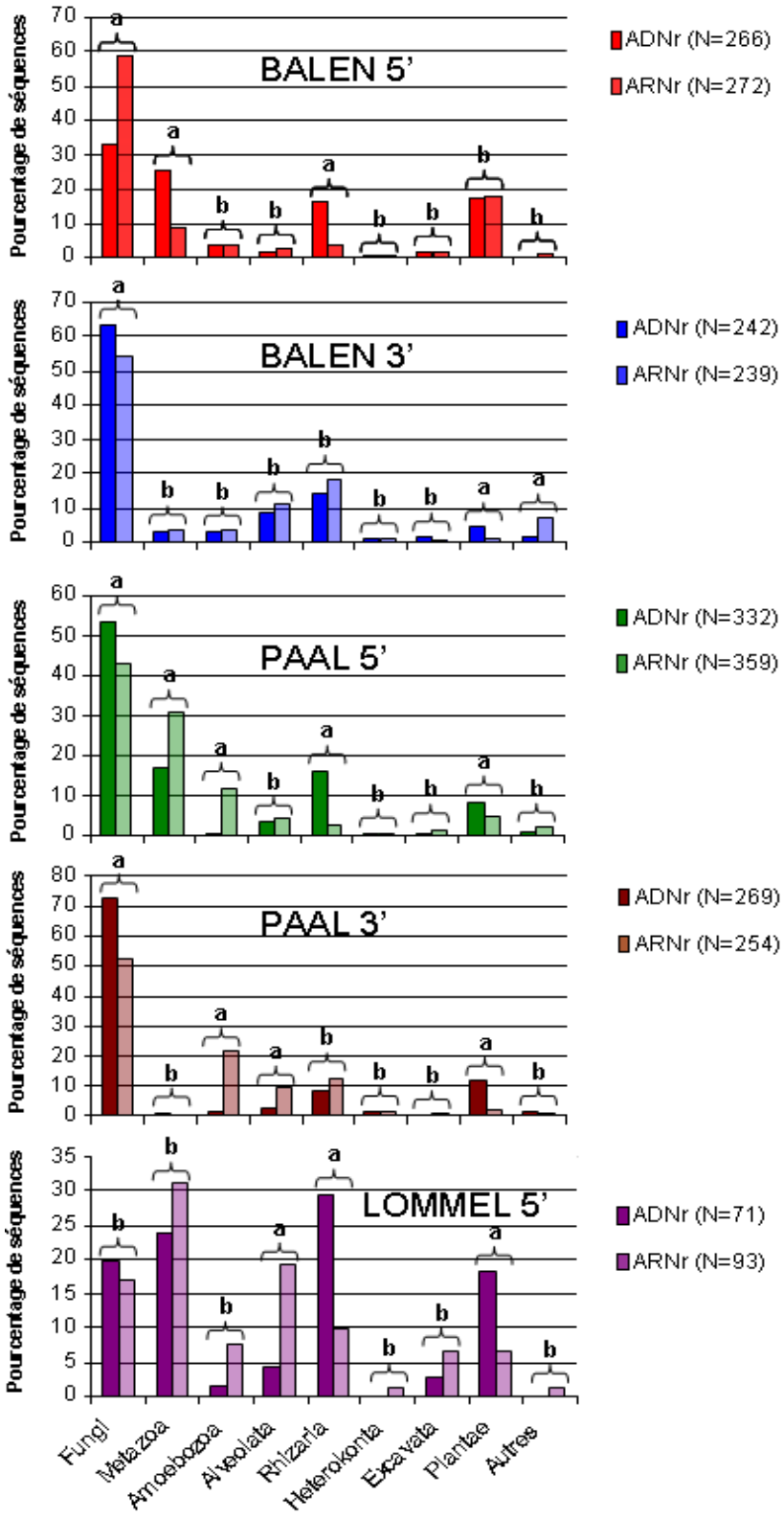


Figure 35: Affiliation taxinomique des séquences représentant le premier (5') et le dernier tiers (3') de l'ADN et de l'ARN ribosomiques 18S obtenues à partir des sols de Paal, Lommel et Balen. La catégorie « autres » inclut des séquences 18S appartenant à des groupes taxinomiques mineurs tels que les choanozoa, les ichtyosporeae ou les nucleariidae. Un test de Chi-2 comparant l'abondance de chacun des phylums entre chaque site d'étude a été effectué ; les valeurs numériques exactes sont indiquées en annexe.
La lettre (a) signifie que la différence entre les 2 jeux de données est significative ($p<0,05$). La lettre (b) signifie que la différence entre les 2 jeux de données n'est pas significative ($p\geq0,05$).

gène 18S analysés (Fig. 35). Par ailleurs, la répartition séparée ou cumulée de ces séquences dans chaque phylum donne des profils d'affiliations relativement semblables.

Les champignons, phylum le mieux représenté à Balen et à Paal, présentent en général un ratio ADNr/ARNr supérieur à un (sauf en 5' à Balen) (Fig. 35). Les metazoa, quasi absents de l'analyse de la portion 3', ont un ratio ADNr/ARNr inférieur à un (sauf en 5' à Balen) (Fig. 5). Concernant les groupes de protistes majoritaires, on s'aperçoit que le ratio ADNr/ARNr des rhizaria est toujours supérieur à un pour la portion 5' et proche de un en ce qui concerne la portion 3' indépendamment du site étudié. Les amoebozoa sont quant à eux mieux représentés par leurs séquences ARN quel que soit la portion du gène 18S analysée, à l'exception du site de Balen où le ratio ADNr/ARNr est toujours égal à un. Les alveolata sont autant représentés par leurs séquences d'ADNr que d'ARNr sauf à Lommel (Fig. 35).

En général, les différents phyla sont représentés aussi bien par des séquences d'ADNr que d'ARNr quel que soit la portion du gène 18S analysée (sauf pour les héterokonta de Lommel et les metazoa et les excavata de Paal). Cependant, après avoir défini des phylotypes avec les séquences d'ADNr et d'ARNr de chaque portion et de chaque site, on s'aperçoit que le nombre de phylotypes définis par les ARN est toujours supérieur à celui des ADN et qu'il y a très peu de redondance entre les banques d'ADNr et d'ARNr construites. En effet, seulement 10 à 20% des phylotypes ont été identifiés à la fois *via* les ADN et les ARN quel que soit la portion du gène analysée et le site étudié (tableau 13).

2.4 Estimation de la richesse taxinomique des Eucaryotes

L'objectif est d'estimer le nombre total de phylotypes eucaryotes présents dans ces sols afin de déterminer le nombre d'espèces dont les gènes exprimés contribuent aux banques d'ADNc environnementales analysée dans les paragraphes suivants.

Différents outils analytiques ont été utilisés. Les courbes de raréfaction indiquent si le panel de séquences obtenu couvre la diversité réelle des communautés. Différents estimateurs de richesse spécifique ont aussi été calculés.

2.4.1 Courbes de raréfaction

Les courbes de raréfactions ont été construites à l'aide du logiciel Analytical rarefaction 1.3 de S. Holland (http://www.uga.edu/strata/software/). Quel que soit le site étudié et la

	Balen 5’	Balen 3’	Paal 5’	Paal 3’	Lommel 5’
Nombre de phylotype ADNr	73	94	66	45	27
Nombre de phylotype ARNr	92	123	96	66	49
Nombre de phylotype commun	17	21	21	25	9
Pourcentage de phylotype commun	9,3%	8,8%	11,5%	18,4%	10,5%

Tableau 13: Nombres de phylotypes identifiés sur chacun des 3 sites étudiés et pour chacune des portions du gène 18S analysées. Il s’agit des phylotypes représentés soit par des séquences d’ADNr, soit par des séquences d’ARNr et soit par les deux.

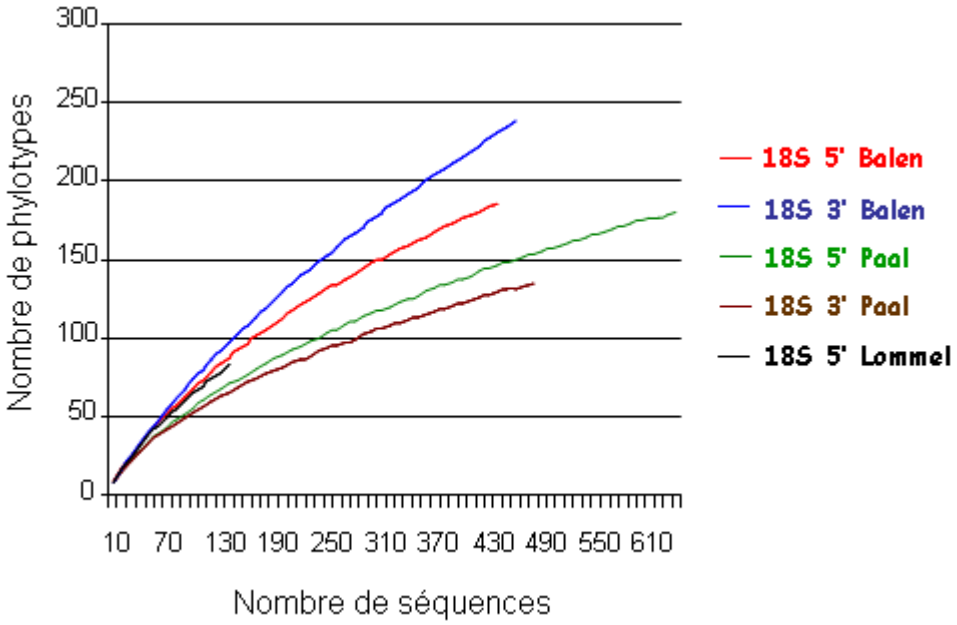


Figure 36: Courbes de raréfaction des phylotypes 18S de champignons, d’animaux et de protistes obtenus pour les 3 sites étudiés à partir des séquences environnementales d’ADNr et d’ARNr 18S cumulées représentant le premier (5’) et le dernier (3’) tiers du gène 18S.

Balen						
Molécules analysées	**ADN 5'**	**ARN 5'**	**Total 5'**	**ADN 3'**	**ARN 3'**	**Total 3'**
Nombre de séquences	220	225	444	231	236	467
Phylotypes observés	90	109	187	115	144	241
S_{ACE}	165	241	369	320	421	613
S_{Chao1}	159	244	363	289	424	606
Phylotypes observés / S_{ACE}	0,57	0,46	0,51	0,37	0,36	0,41
Phylotypes observés / S_{Chao1}	0,59	0,46	0,52	0,40	0,37	0,41

Tableau 14: Estimation de la richesse en microorganismes eucaryotes (champignons, protistes et animaux) présents dans les échantillons de sols prélevés à Balen par calcul des indices S_{Chao1} et S_{ACE} à partir des séquences environnementales d'ADNr, d'ARNr et d'ADNr + ARNr représentant le premier (5') et le dernier (3') tiers du gène 18S.

Paal						
Molécules analysées	**ADN 5'**	**ARN 5'**	**Total 5'**	**ADN 3'**	**ARN 3'**	**Total 3'**
Nombre de séquences	307	344	649	238	249	487
Phylotypes observés	87	118	182	70	91	137
S_{ACE}	150	277	366	126	174	264
S_{Chao1}	143	227	327	117	168	261
Phylotypes observés / S_{ACE}	0,59	0,44	0,52	0,55	0,54	0,55
Phylotypes observés / S_{Chao1}	0,59	0,53	0,56	0,59	0,55	0,56

Tableau 15: Estimation de la richesse en microorganismes eucaryotes (champignons, protistes et animaux) présents dans les échantillons de sols prélevés à Paal par calcul des indices S_{Chao1} et S_{ACE} à partir des séquences environnementales d'ADNr, d'ARNr et d'ADNr + ARNr représentant le premier (5') et le dernier (3') tiers du gène 18S.

portion du gène 18S analysée, les courbes n'atteignent pas une valeur asymptotique (Fig.36) ce qui signifie que les jeux de séquences analysés ne décrivent pas totalement les richesses spécifiques des communautés de microorganismes eucaryotes étudiées. Ces courbes permettent, cependant, de prédire que la diversité de microorganismes eucaryotes présente à Balen est plus importante qu'à Lommel et à Paal, quelle que soit la portion du gène 18S analysée.

2.4.2 Estimation des indices de richesse des communautés

Différentes approches statistiques permettent d'estimer la richesse spécifique d'une communauté à partir d'un échantillon de celle-ci. Nous avons choisi deux estimateurs S_{chao1} et S_{ACE} tous deux basés sur l'abondance des taxons. Ils ont été calculés grâce au logiciel développé par Kemp et Aller (2004) (http://www.also.org/lomethods/free/2004/0114a.html).

Pour chaque site et chaque portion du gène 18S analysé, la valeur des estimateurs est toujours plus importante pour les molécules d'ARNr par rapport à celle des molécules d'ADNr (tableau 14, 15 et 16). Cependant, les valeurs obtenues avec les données cumulées (ADNr + ARNr) sont nettement supérieures et montrent que le nombre de séquences obtenues a permis de recouvrir entre 40 et 55% de la diversité taxinomique quel que soit le site et la portion du gène analysés (tableau 14, 15 et 16). Enfin ces résultats corroborent ceux obtenus avec la construction des diagrammes de Venn et des courbes de raréfaction montrant que le site de Balen contaminé par des métaux lourds contient une plus grande diversité taxinomique de microorganismes eucaryotes que le site de Paal non pollué et le site de Lommel anciennement pollué.

2.5 Conclusions

L'analyse bioinformatique de centaines de séquences 18S de banques d'ADNr et d'ARNr construites à partir des différents sols a permis de montrer :

- que ces banques sont de bonne qualité puisqu'elles contiennent très peu d'artéfacts (chimères, autres séquences que 18S, séquences 18S tronquées).
- que les sols étudiés contiennent une grande diversité d'organismes eucaryotes puisque tous les phyla eucaryotes y sont représentés et que la redondance entre ARNr 18S et ADNr 18S est très faible, quel que soit la portion du gène 18S analysée.

Lommel						
Molécules analysées	**ADN 5'**	**ARN 5'**	**Total 5'**	**ADN 3'**	**ARN 3'**	**Total 3'**
Nombre de séquences	58	86	144	n.d	n.d	n.d
Phylotypes observés	36	58	85	n.d	n.d	n.d
S_{ACE}	105	172	246	n.d	n.d	n.d
S_{Chao1}	71	157	205	n.d	n.d	n.d
Phylotypes observés / S_{ACE}	0,32	0,37	0,36	n.d	n.d	n.d
Phylotypes observés / S_{Chao1}	0,47	0,40	0,44	n.d	n.d	n.d

Tableau 16: Estimation de la richesse en microorganismes eucaryotes (champignons, protistes et animaux) présents dans les échantillons de sols prélevés à Lommel par calcul des indices S_{Chao1} et S_{ACE} à partir des séquences environnementales d'ADNr, d'ARNr et d'ADNr + ARNr représentant le premier (5') tiers du gène 18S.

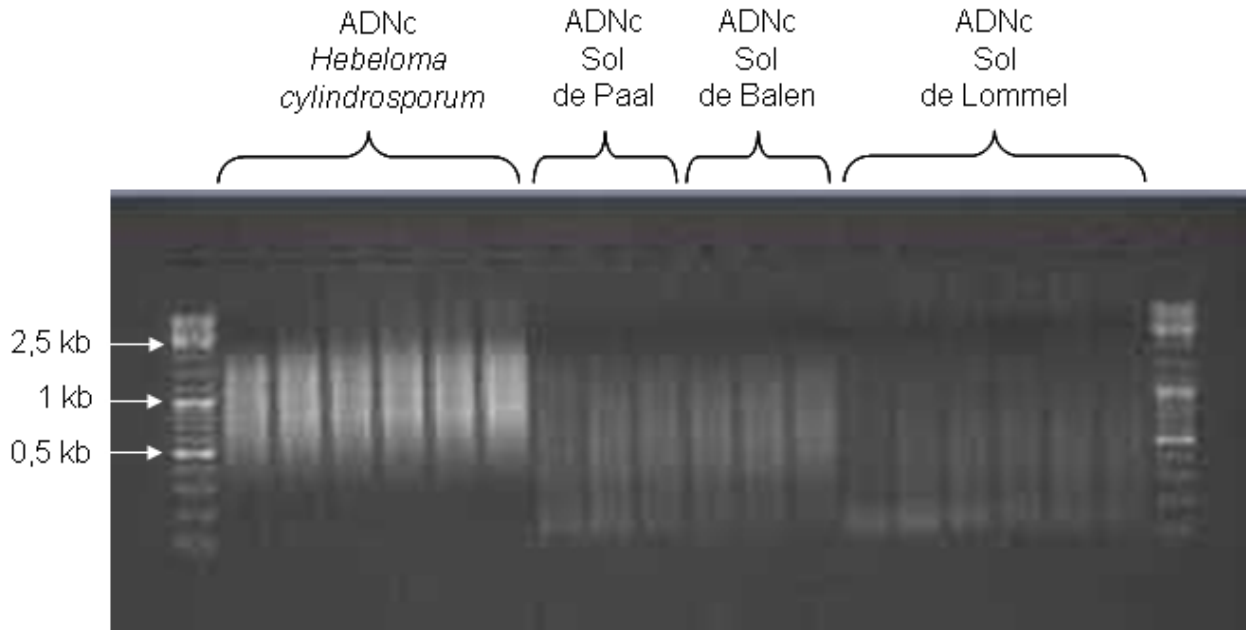


Figure 37: Profil d'électrophorèse des ADNc générés avec le kit SMART cDNA Library Construction (Clontech) à partir des ARNm extraits du champignon *Hebeloma cylindrosporum* (utilisés comme témoins) et des différents sols étudiés.

- que chaque sol présente une diversité taxinomique spécifique. Peu de phylotypes communs à 2 ou 3 sols ont été détectés.

- que la richesse spécifique de la communauté de microorganismes eucaryotes du sol contaminé aux métaux lourds semble équivalente (portion 5') voire plus importante (portion 3') que celle d'un sol témoin non contaminé ou anciennement contaminé par des métaux lourds.

3. Etude de la diversité fonctionnelle

La diversité fonctionnelle des sols de Balen et Paal a été estimée de façon globale par l'analyse bioinformatique de plusieurs milliers de séquences d'ADNc séquencées par le Genoscope (Evry, France). Les séquences présentant des homologues de fonction connue dans les bases de données ont été réparties en groupes fonctionnels.

Les banques d'ADNc de Balen, Paal et Lommel ont aussi été criblées par expression hétérologue dans des mutants de levures sensibles au cadmium en vue d'estimer la diversité des mécanismes eucaryotes de résistance aux métaux au sein d'une communauté présente dans un sol contaminé, anciennement contaminé et non contaminé aux métaux lourds.

3.1 Construction des banques d'ADNc

3.1.1 Purification des ARNm et synthèse des ADNc

Les ARNm polyadénylés eucaryotes ont été purifiés par affinité sur billes magnétiques recouvertes d'oligo-$(dT)_{25}$ à partir des ARN totaux issus des sols de Paal, Lommel et Balen.

La transcription inverse des ARNm en ADNc a été effectuée avec le kit SMART cDNA Library Construction (Clontech) (cf Matériels et méthodes, Fig. 52). L'amplification des ARNm rétrotranscrits par PCR avec des amorces fournies par le kit a généré des ADNc de tailles comprises entre 0,1 et 2 kb pour les 3 sols (Fig. 37). Ces ADNc sont bordés par les sites de restriction *Sfi* IA et *Sfi* IB permettant leur clonage directionnel dans un vecteur possédant ces 2 sites.

3.1.2 Les vecteurs de clonage

Deux vecteurs plasmidiques ont été utilisés.

Le plasmide pDNR-LIB, fourni par le kit SMART cDNA Library Construction, a été utilisé pour la construction des banques qui ont été séquencées de façon systématique par le Génoscope.

Le plasmide pFL61 (Minet *et al.*, 1992), modifié au laboratoire par insertion des sites *Sfi* IA et *Sfi* IB (cf Matériels et méthodes, Fig. 53), a été utilisé pour construire les banques criblées par expression hétérologue dans des mutants de levure sensibles au cadmium. Il s'agit d'un plasmide navette possédant un gène de résistance à l'ampicilline permettant sa sélection dans *E. coli* et le gène URA-3 pour sa sélection dans *S. cerevisiae*. Il possède aussi un promoteur fort constitutif (PGK) permettant l'expression dans la levure des ADNc clonés. Pour permettre le clonage directionnel d'ADNc générés par le kit SMART cDNA Library Construction (Clontech) il a été nécessaire d'introduire les deux sites de restrictions *Sfi* IA et *Sfi* IB en aval du promoteur PGK au niveau des sites *Not* I déjà présents dans pFL61. Le séquençage de la région modifiée a permis de vérifier la bonne insertion des sites *Sfi* IA et *Sfi* IB.

Afin de vérifier la fonctionnalité de ce nouveau vecteur de clonage pour l'expression hétérologue d'ADNc dans *S. cerevisiae*, un gène HIS3 codant une imidazole glycérol phosphate déshydratase et bordés par les sites *Sfi* IA et *Sfi* IB (Bailly *et al*., 2007) a été inséré par ligation dans les sites *Sfi* IA et *Sfi* IB du plasmide pFL61 modifié. Le plasmide pFL61-HIS3 ainsi obtenu a été introduit par transformation dans la souche de *S. cerevisiae* BY4741 (ura^-, his^-). Soixante transformants ont été isolés sur milieu sélectif YNB + glucose dépourvu d'uracile et d'histidine et le gène HIS3 a pu être amplifié par PCR et séquencé à partir des ADN extraits de levures recombinantes. La croissance de ces levures transformées sur un milieu de culture dépourvu d'histidine montre que le gène HIS3 s'exprime correctement et que le vecteur d'expression pFL61 modifié est donc fonctionnel.

3.1.3 Clonage des ADNc

Les ADNc générés (Fig. 37) ont été digérés par l'enzyme de restriction *Sfi I* et fractionnés sur une colonne d'exclusion. Les ADNc de tailles inférieures à 400pb ont été éliminés. Les autres ont été clonés dans les deux vecteurs de clonage pDNR-LIB et pFL61 modifié.

	Titre de la banque dans :	
Plasmides	**pDNR**	**pFL61**
Paal	13.10^6	3.10^5
Lommel	15.10^6	5.10^6
Balen	$1,5.10^6$	$3,5.10^6$

Tableau 17: Titres (nombre total d'ADNc clonés) des banques d'ADNc construites à partir des différents sols dans les 2 vecteurs de clonage utilisés.

	Balen	**Paal**
Nombre de séquences analysées	7101	8372
Nombre de séquences uniques	5405	8075
Nombre de contigs	592	87

Tableau 18: Nombre de contigs et de séquences uniques détectés par le logiciel Basic CAP3 dans les jeux de séquences d'ADNc de Balen et de Paal.

Les ADNc ont été introduits par transformation dans la souche électrocompétente DH10B d'*E. coli*. Les cellules transformées ont été sélectionnées sur un milieu sélectif et un dénombrement des clones a été effectué afin de déterminer le titre des banques d'ADNc ainsi générées pour chaque sol dans les 2 vecteurs de clonage (Tableau 17).

3.2 Analyse des banques d'ADNc par séquençage massif

Les banques d'ADNc construites dans le vecteur pDNR (Clontech) à partir du sol de Balen ($1{,}5.10^6$ clones) et du sol de Paal (13.10^6 clones) ont fait l'objet d'un séquençage massif par le Génoscope (Evry, France) suivant la méthode Sanger (ABI 3730). Un programme de séquençage de 30.000 inserts d'ADNc de chaque banque est en cours de réalisation. Actuellement, nous disposons de 7101 séquences d'ADNc pour le sol de Balen et de 8372 séquences d'ADNc pour le sol de Paal. Ces séquences nous étant parvenues tardivement en cours de rédaction de ce manuscrit, les analyses ci-dessous restent préliminaires et devront être complétées et affinées.

3.2.1 Détection des séquences redondantes ou chevauchantes

Les séquences redondantes ou chevauchantes ont été détectées à l'aide du logiciel Basic CAP3 (https://www.genome.clemson.edu/cgi-bin/cugi_cap3?advanced=1). Les séquences présentant 98% d'identité sur une longueur minimale de 100 pb ont été regroupées en contigs (Tableau 18). La plupart des contigs identifiés sont composés de séquences identiques. Seuls quelques uns sont constitués de séquences qui se chevauchent sur un minimum de 100 pb avec 98% d'identité et permettent de former une séquence consensus.

Les séquences définissant des contigs représentent 3,5% du jeu de séquences d'ADNc de Paal et 24% du jeu de séquences de Balen (tableau 18). Plus de 90% des contigs des 2 jeux de séquences sont constitués au maximum de 5 séquences (Fig. 38 et 39). Seuls quelques rares contigs contiennent plus de 10 séquences. Deux contigs de Balen contiennent 33 et 265 séquences d'ADNc codant respectivement une saccharopine deshydrogénase et une protéine hypothétique. Deux contigs de Paal contiennent 24 et 51 séquences d'ADNc qui codant respectivement une sous unité de l'ATP synthase et un ARNr 18S.

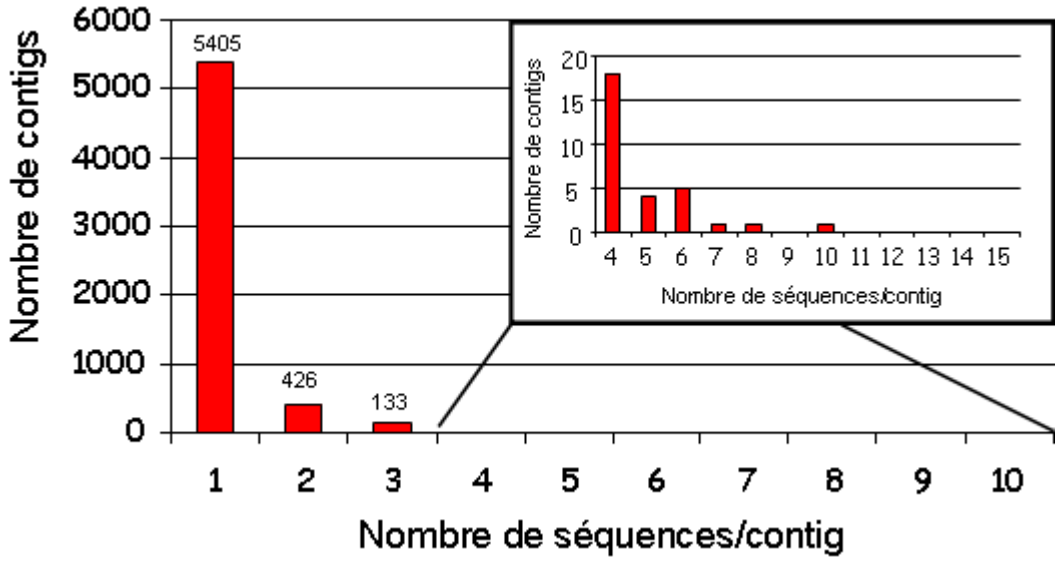


Figure 38: Nombre de contigs en fonction du nombre de séquences d'ADNc par contigs identifiés pour le sol de Balen. Deux contig, non représentés, contenant respectivement 33 et 265 séquences d'ADNc ont également été identifiés.

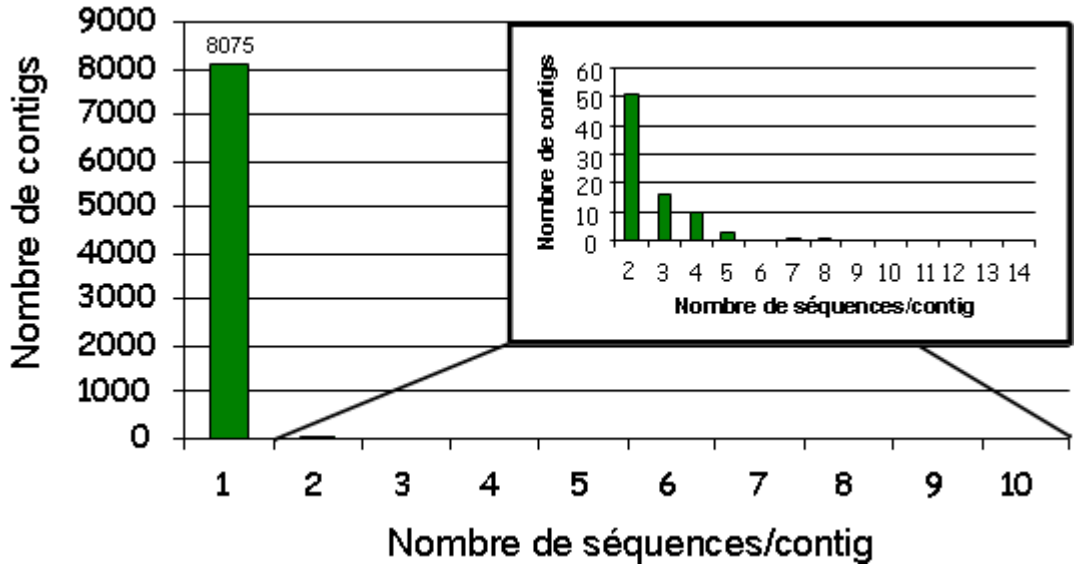


Figure 39: Nombre de contigs en fonction du nombre de séquences d'ADNc par contigs identifiés pour le sol de Paal. Deux contig, non représentés, contenant respectivement 24 et 51 séquences d'ADNc ont également été identifiés.

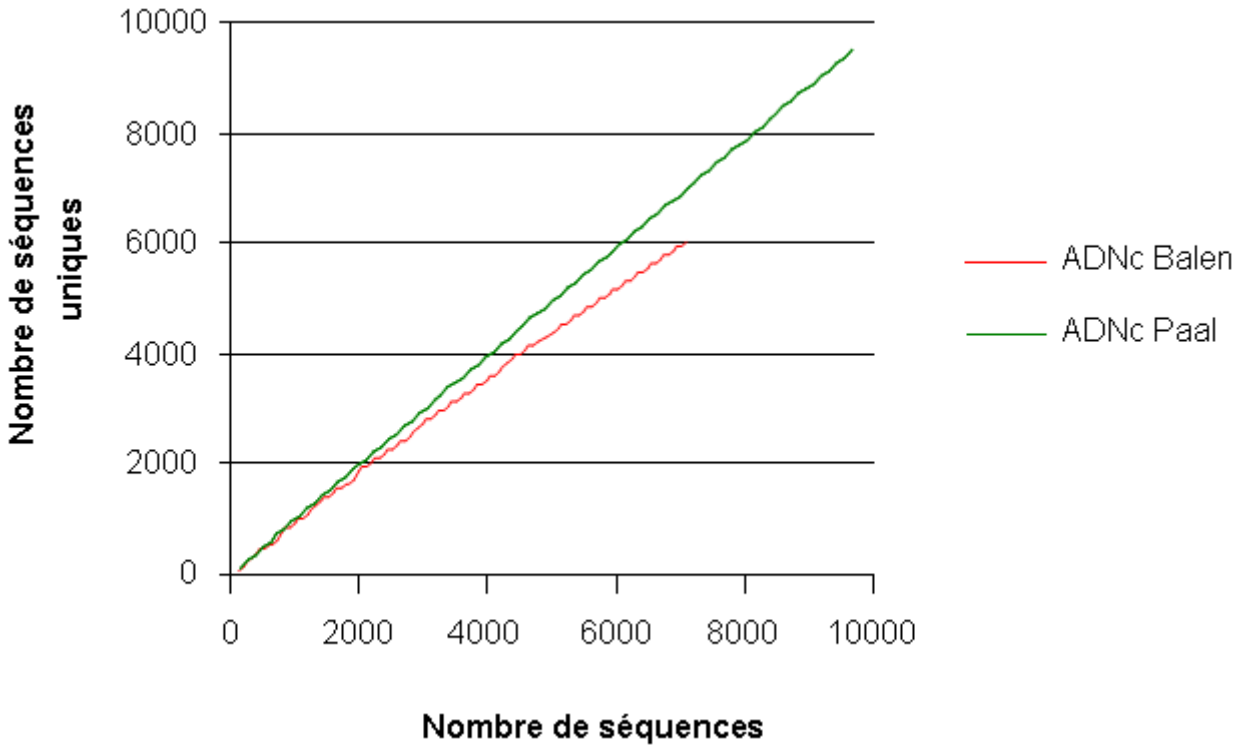


Figure 40: Courbes de raréfactions des séquences d'ADNc eucaryotes uniques obtenues pour le site de Balen et le site de Paal à partir des ARNm polyadénylés.

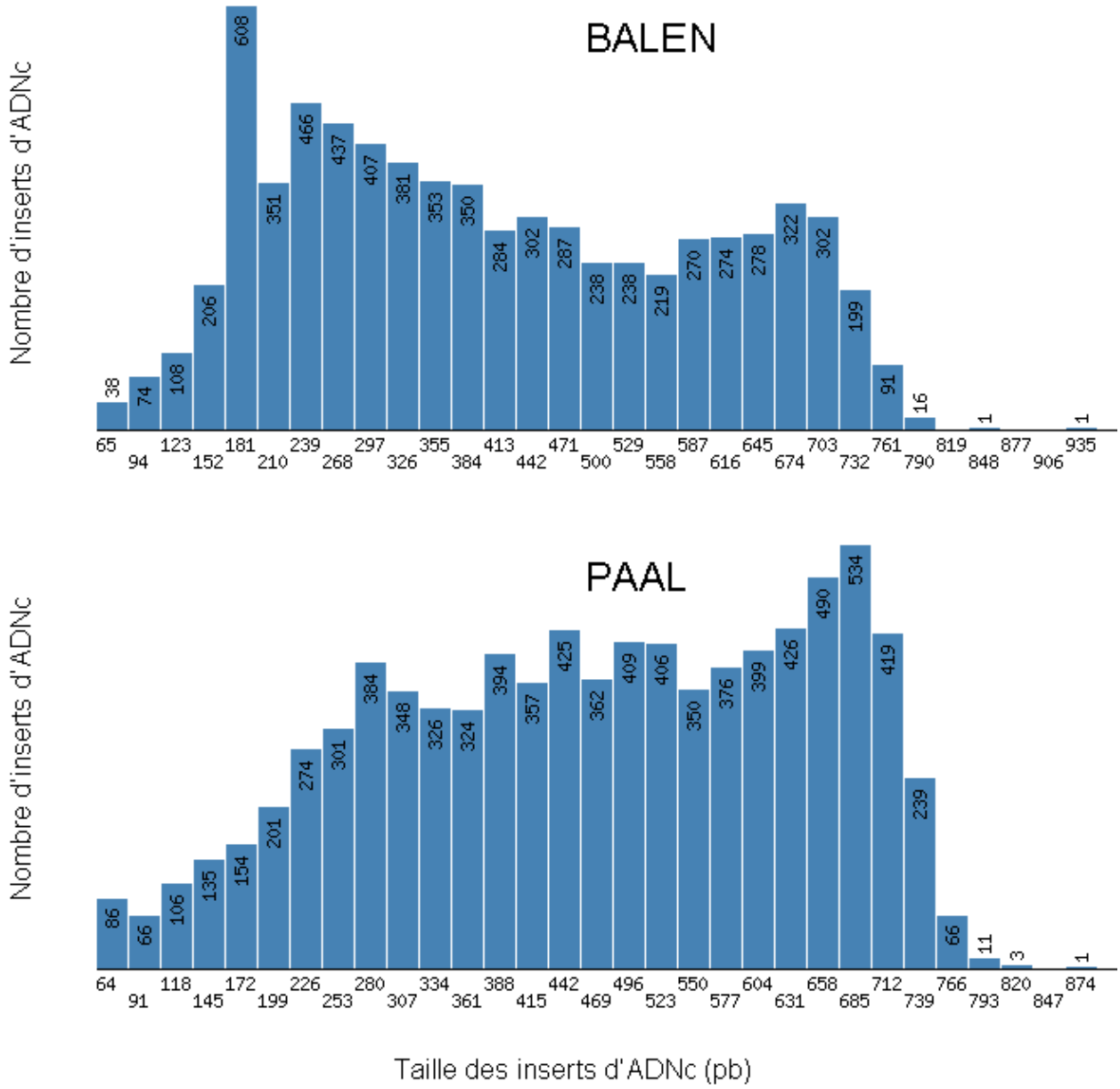


Figure 41: Distribution des séquences d'ADNc de Balen et de Paal selon leurs tailles.

3.2.2 Estimation de la diversité moléculaire

La complexité moléculaire de ces jeux de séquences d'ADNc se traduit par des courbes de raréfaction qui non seulement n'atteignent pas de valeurs asymptotiques et mais qui aussi ne présentent qu'une très faible inflexion visible pour le jeu de données de Balen (Fig. 40).

3.2.3 Analyse bioinformatique des séquences d'ADNc et répartition en groupes fonctionnels

Toutes les séquences d'ADNc de chaque site ont été comparées, à l'aide du serveur d'analyse métagénomique MG-RAST (http://metagenomics.nmpdr.org/), à la base de données SEED (http:/www.theSEED.org) via un BLASTX. Cette analyse permet de distribuer les séquences dans différentes catégories fonctionnelles définies par ce serveur.

Tout d'abord, une répartition des séquences d'ADNc suivant leur taille a été réalisée (Fig. 10). Le jeu de séquences de Balen est constitué de séquences de plus petites tailles (420 pb en moyenne) que celui de Paal (480 pb en moyenne) (Fig. 41).

Ensuite, les séquences ont été classées soit en séquences codant des protéines annotées (de fonctions connues) dans les bases de données, soit en séquences codant des protéines hypothétiques conservées, soit en séquences sans homologues dans les bases de données (e-value > $1e^{-5}$). Cette dernière catégorie représente les nouvelles protéines hypothétiques potentielles identifiées par l'approche métatranscriptomique (Fig.42). Pour chacune de ces 3 catégories, des différences significatives ont été observées entre les deux jeux de données. Un nombre plus important de nouvelles protéines hypothétiques est observé à Balen. Cependant, au vu du plus grand nombre de séquences de petites tailles contenues dans cette banque, il se peut que des séquences tronquées, trop courtes pour être identifiées, soient comptabilisées comme de nouvelles protéines hypothétiques.

Enfin, les 1080 séquences d'ADNc de Balen et les 1569 de Paal codant des protéines annotées dans les bases de données ont été réparties dans 15 groupes fonctionnels à l'aide du serveur MG-RAST (Fig. 43). Le nombre de séquences analysées jusqu'à présent a permis de mettre en évidence des différences significatives entre sites concernant les pourcentages de protéines impliquées dans la respiration, le métabolisme des acides animés, des lipides et le métabolisme secondaire (Fig. 43). Aucune différence significative concernant le nombre de protéines impliquées dans la réponse au stress n'a été observée. Toutefois, un plus grand

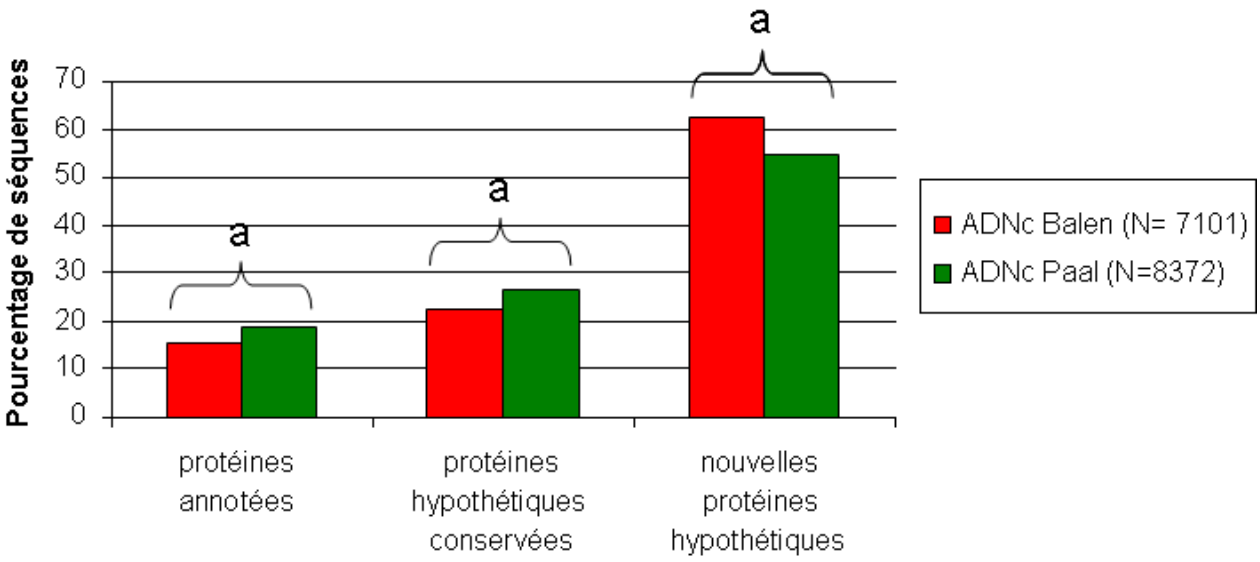


Figure 42: Distribution des ADNc des sites de Balen et Paal selon la nature des protéines potentiellement codées. Protéines annotées: protéines présentant une fonction biologique connue ; protéines hypothétiques conservées: protéines de fonctions inconnues déposées dans les bases de données ; nouvelles protéines hypothétiques: ADNc sans homologues (Blastx) dans les bases de données publiques (e-value > $1e^{-5}$).
Un test de Chi-2 comparant les proportions de chacune de ces catégories entre les 2 sols étudiés a été effectué ; les valeurs numériques sont indiquées en annexe.
La lettre (a) signifie que la différence entre les 2 jeux de données est significative ($p<0,05$).

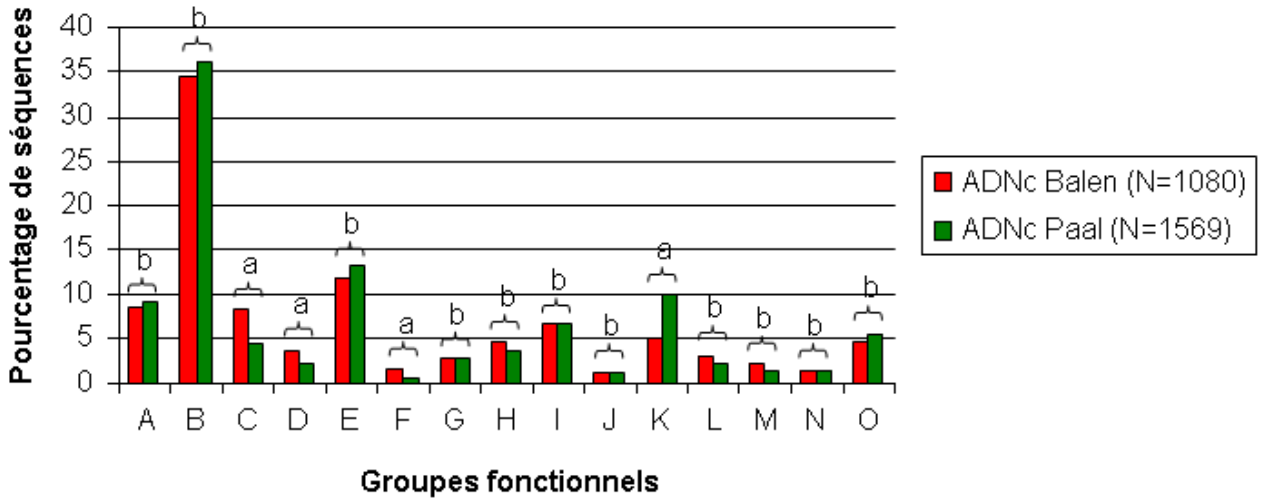


Figure 43: Répartition des séquences protéiques déduites des ADNc annotés au sein de groupes fonctionnels définis par le serveur MG-RAST.
A, métabolisme des acides nucléiques ; B, métabolisme des protéines ; C, métabolisme des acides aminés et dérivés ; D, métabolisme des acides gras et des lipides ; E, métabolisme des glucides ; F, métabolisme secondaire ; G, autres métabolismes (K, N, S…) ; H, cofacteurs, vitamines, groupes prosthétiques, pigments ; I, trafic cellulaire ; J, contrôle du cycle cellulaire ; K, respiration ; L, réponses au stress ; M, virulence ; N, paroi cellulaire ; O, autres fonctions. Un test de Chi-2 comparant les proportions de chacune de ces catégories entre les 2 sols étudiés a été effectué ; les valeurs numériques sont indiquées en annexe.
La lettre (a) signifie que la différence entre les 2 jeux de données est significative ($p<0,05$). La lettre (b) signifie que la différence entre les 2 jeux de données n'est pas significative ($p\geq0,05$).

	Balen	Paal
Choc froid	-Cold shock protein CspD -Cold shock protein CspC	-Cold shock protein CspE -Cold shock protein CspE -Cold shock protein CspG -Cold shock protein CspA
Stress osmotique		-Aquaporin Z
Stress oxydatif	-Glutathione synthetase -Glutathione S-transferase -Glutathione S-transferase -Glutathione S-transferase -Glutathione S-transferase -Glutathione S-transferase -Peroxidase -Peroxidase -Peroxidase -Peroxidase -Peroxidase -Peroxidase -Peroxidase -Peroxidase -Peroxidase -Peroxidase -Alkyl hydroperoxide reductase subunit C-like protein -Cytochrome c551 peroxidase -Cytochrome c551 peroxidase -Poly [ADP-ribose] polymerase-1 -Poly [ADP-ribose] polymerase-1 -Poly [ADP-ribose] polymerase-1 -NAD-dependent deacetylase	-Superoxide dismutase [Cu-Zn] precursor -Superoxide dismutase [Cu-Zn] precursor -Superoxide dismutase [Fe] -Superoxide dismutase [Mn] -Superoxide dismutase [Mn] -Superoxide dismutase [Mn] -Catalase -Rubrerythrin -Alkyl hydroperoxide reductase subunit C-like protein -Alkyl hydroperoxide reductase subunit C-like protein -Glutathione peroxidase -5-oxoprolinase -Poly[ADP-ribose] polymerase-1
Autres	-Carbon storage regulator -Multicopper oxidase -Putative diheme cytochrome c-553 -Cytochrome c-type biogenesis protein CcmG/DsbE -Copper-translocating P-type ATPase -Ornithine aminotransferase -NG,NG-dimethylarginine dimethylaminohydrolase 1	-Arginine decarboxylase -Carbon storage regulator -Multicopper oxidase -Protein arginine N-methyltransferase 1 -FKBP-type peptidyl-prolyl cis-trans isomerase -FKBP-type peptidyl-prolyl cis-trans isomerase -FKBP-type peptidyl-prolyl cis-trans isomerase -Heat shock protein GrpE -RNA polymerase sigma factor RpoH -RNA polymerase sigma factor RpoH -Signal peptidase-like protein -RNA-binding protein Hfq -Universal stress protein family

Tableau 19: Annotation des protéines codées par chacun des ADNc placés dans le groupe fonctionnel « réponses au stress ». Les protéines de même annotation retrouvées à Balen et à Paal sont en rouge (Classification selon MG-RAST).

nombre de protéines impliquées dans le stress oxydant est observé à Balen et la diversité de gènes exprimés en réponse à un stress semble différente au sein de chacun des sols (Tableau 19).

3.3 Criblage des banques d'ADNc par expression hétérologue dans la levure

Les banques métatranscriptomiques de Balen, Paal et Lommel construites dans le plasmide pFL61 ont été criblées par complémentation fonctionnelle de deux mutants de levures sensibles au cadmium. Ces mêmes banques ont fait l'objet d'un criblage par complémentation fonctionnelle d'un mutant de levure sensible au zinc par nos partenaires nancéens (Didier Doillon, Damien Blaudez et Michel Chalot ; Unité Interactions Arbres-Microorganismes, INRA Université Nancy1).

3.3.1 Complémentation fonctionnelle du mutant ΔYcf1

Le mutant ΔYcf1 (Mat a ; his3Δ1 ; leu2Δ0 ; met15Δ0 ; ura3Δ0; Ycf1::kanMX4) de la souche BY4741 de *Saccharomyces cerevisiae* est délété du gène *ycf1* codant pour un transporteur vacuolaire du glutathion de type ABC (ATP-binding cassette). Par conséquent, le mutant obtenu n'est plus capable de transporter le complexe glutathion-Cd du cytoplasme vers la vacuole et accumule des concentrations toxiques de Cd dans le cytoplasme.

La complémentation fonctionnelle de ce mutant par la banque de Balen ($1{,}33.10^8$ transformants criblés, soit 38 fois la taille initiale de la banque) a conduit à isoler 567 transformants sur milieu W0 + Cd (40µM). Après re-isolement de ces transformants par stries d'épuisement sur le même milieu puis élimination du plasmide par l'acide 5-fluoro-orotique et test de la sensibilité au Cd des souches sur milieu W0 + Cd (40µM) + uracile, 130 « vrais » transformants résistants au Cd ont été conservés. Par la suite, sur la base de tests en gouttes, seulement 16 transformants se sont révélés capables de croître au moins autant que la souche sauvage et mieux que le mutant sur le milieu sélectif (Fig. 44). Enfin, après extraction du plasmide contenant l'insert d'ADNc responsable de la complémentation fonctionnelle et re-tranformation du mutant ΔYcf1, le test en goutte de confirmation a aboutit à la sélection de 9 gènes conférant une résistance au Cd.

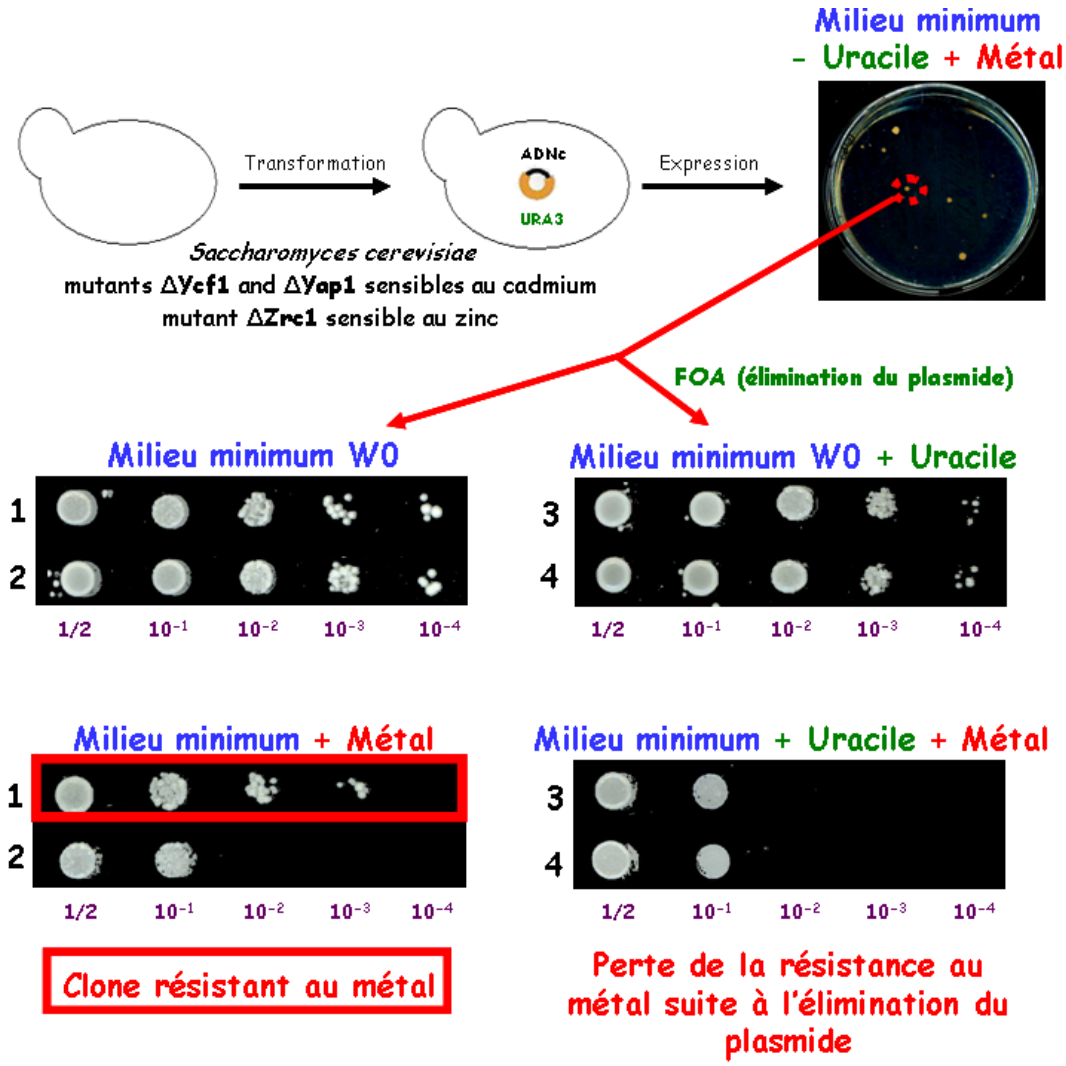


Figure 44: Les différentes étapes de sélection d'un clone résistant au métal par complémentation fonctionnelle d'un mutant de levure sensible au métal avec une banque métatranscriptomique eucaryote.
Le transformant primaire résistant au métal est réisolé par stries d'épuisement sur milieu W0 + métal ansi que sur milieu W0 + Uracile + 5FOA afin d'éliminer le plasmide.
1, gène de résistance ; 2, vecteur vide ; 3, gène de résistance éliminé ; 4, souche originelle de levure sensible au métal.

	ADNc ΔYcf1	ADNc ΔYcf1 vs ADNc Génoscope Balen	ADNc ΔYcf1 vs ADNc Génoscope Paal
Balen	Protéine hypothétique 1 Protéine hypothétique 1 Protéine hypothétique 1 Protéine hypothétique 1	265	0
	Protéine hypothétique 2	0	1
	Protéine hypothétique 3	0	0
	Protéine hypothétique 4	0	0
	Zn-binding alcohol dehydrogenase (bactérie)	0	1
	Métallothionéine (mollusque)	0	0
Lommel	Nouvelle MT 1 Nouvelle MT 1 Nouvelle MT 1 Nouvelle MT 1 Nouvelle MT 2 Nouvelle MT 2 Nouvelle MT 2 Nouvelle MT 3 Nouvelle MT 3 Nouvelle MT 3 Nouvelle MT 4 Nouvelle MT 4 Nouvelle MT 4 Nouvelle MT 4 Nouvelle MT 4 Nouvelle MT 5	0	7
	MT (champignon)	2	20
Paal	Nouvelle MT 4	0	7

Tableau 20: Annotation des séquences d'ADNc des banques métatranscriptomiques de Balen, Lommel et Paal complémentant le phénotype du mutant de levure ΔYcf1 sensible au cadmium. Une recherche de ces séquences dans les jeux de séquences d'ADNc de Balen et de Paal fournis par le Génoscope a été effectuée à l'aide du programme TBLASTN. MT, métallothionéine.

La même démarche expérimentale a été suivie pour les banques de Paal et de Lommel. Dans le cas de Paal, sur $8{,}1.10^5$ transformants obtenus (2,7 fois la taille de la banque), 100 poussaient sur milieu sélectif. A la suite des différents tests, seul 1 gène conférant une résistance avérée au Cd a été retenu.

Pour Lommel, sur $6{,}1.10^6$ transformants obtenus (1,2 fois la taille de la banque), 124 poussaient sur milieu sélectif. A la suite des différents tests, 17 gènes conférant une résistance avérée au Cd ont été sélectionnés.

3.3.2 Analyse des séquences d'ADNc restaurants le phénotype ΔYcf1

Les séquences d'ADNc capables de restaurer le phénotype de sensibilité au cadmium du mutant ΔYcf1 ont été comparées à l'aide du programme BLASTX à la base de données GenBank nr qui ne contient pas les séquences d'EST. Certaines séquences, notamment celles ne montrant aucune homologie avec des séquences connues ont ensuite été analysées à l'aide du programme TBLASTX permettant la comparaison avec les données issues d'EST (http://www.ncbi.nlm.nih.gov/BLAST). Une fois annotées, ces séquences ont été recherchées dans les jeux de séquences d'ADNc de Balen et de Paal fournis par le Génoscope (transformés en bases de données à l'aide du programme formatdb) à l'aide du programme TBLASTN.

L'analyse bioinformatique a montré que 7 des 9 séquences d'ADNc obtenue à partir de la banque de Balen correspondent à des protéines hypothétiques. Parmi ces 7 séquences, 4 sont identiques entre elles et aux 265 séquences formant un même contig retrouvées dans le jeu de séquences d'ADNc de Balen fourni par le Génoscope (cf 3.2.1). Un ADNc code la portion C-terminale d'une « Zn-binding alcohol dehydrogenase », retrouvée une fois dans le jeu de séquence d'ADNc de Paal. Cette portion de la protéine correspond au domaine de liaison au NADP (domaine dit de NADB-Rossmann). Un autre ADNc code une protéine riche en cystéine (25%) pouvant correspondre à une métallothionéine. Les deux séquences les plus similaires correspondent à une protéine hypothétique de Trichomonas et une métallothionéine de mollusque (Tableau 20).

Le criblage de la banque de Lommel a permis d'identifier 16 séquences homologues entre elles codant des protéines hypothétiques riches en cystéines, supposées être de nouvelles métallothionéines (Tableau 20). Il s'agit de protéines de 110 à 132 acides aminés dont 24 à 27 résidus cystéines. L'alignement des séquences protéiques a permis de répartir ces protéines en 5 sous-familles suivant le motif dessiné par l'enchainement de leurs résidus cystéines. Elles

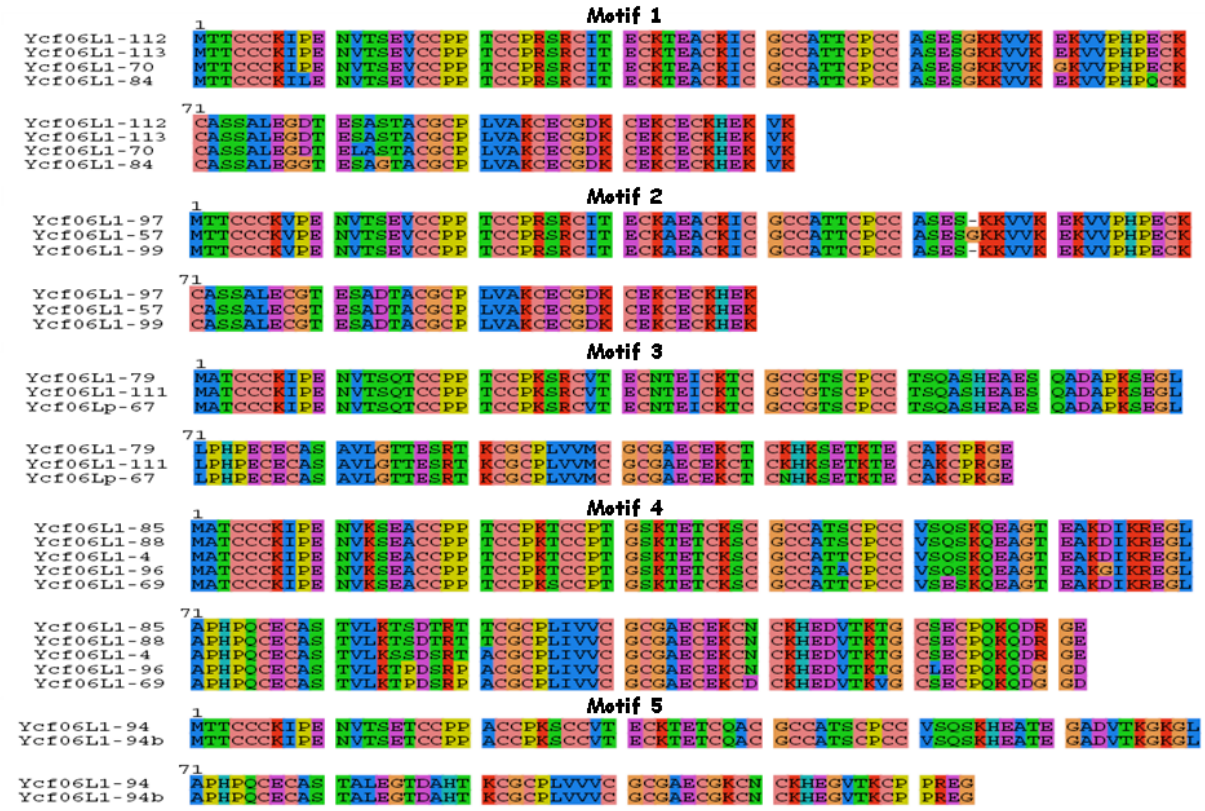


Figure 45: Alignement des séquences protéiques hypothétiques riches en cystéine obtenues lors du criblage de la banque métatranscriptomique de Lommel par complémentation fonctionnelle du mutant de levure ΔYcf1 sensible au cadmium. Les protéines ont été réparties en 5 familles suivant le motif dessiné par l'enchainement de leurs résidus cystéines Une séquence protéique similaire à celles du motif 4 a été obtenue à partir de la banque métatranscriptomique de Paal.

présentent toutes un triplet de cystéine à l'extrémité N-terminale (Fig. 45) et l'amplification de leurs séquences génomiques, à partir du métagénome de Lommel, n'a pas révélé la présence d'intron (Franck Lejzerowicz, communication personnelle). Toutes ces nouvelles métallothionéines ont été retrouvées 7 fois dans le jeu de séquences d'ADNc de Paal fourni par le Génoscope. La dernière protéine correspond à une métallothionéine de champignon basidiomycète retrouvée 2 fois dans le jeu de séquences de Balen et 20 fois dans le jeu de séquence de Paal (Tableau 20).

Enfin, la seule séquence d'ADNc obtenue à partir de la banque de Paal correspond à une des nouvelles métallothionéines (motif 4) (Fig. 45) trouvées à Lommel.

3.3.3 Complémentation fonctionnelle du mutant ΔYap1

Le mutant ΔYap1 (Mat a ; his3Δ1 ; leu2Δ0 ; met15Δ0 ; ura3Δ0 ; Yap1::kanMX4) de la souche BY4741 de *Saccharomyces cerevisiae* est délété du gène *yap1* codant un facteur de transcription de type bZIP (basic leucine zipper) impliqué dans la tolérance au stress oxydant. Le mutant n'est donc plus capable d'activer la transcription de nombreux gènes impliqués dans la résistance au stress oxydant tel que les thiorédoxines et les peroxyrédoxines.

Pour la sélection de gènes de résistance, la même démarche expérimentale que celle exposée dans le cas du mutant ΔYcf1 a été suivie sauf que la concentration en cadmium utilisée était de 20 µM.

Dans le cas de Balen, sur 8.10^7 transformants obtenus (23 fois la taille de la banque), 899 poussaient sur milieu sélectif. A la suite des différents tests, 9 gènes conférant une résistance avérée au Cd ont été retenus.

Pour Paal, sur $1{,}4.10^6$ transformants obtenus (4,6 fois la taille de la banque), 162 poussaient sur milieu sélectif. A la suite des différents tests, 3 gènes conférant une résistance avérée au Cd ont été sélectionnés.

Pour Lommel, sur 2.10^6 transformants obtenus (0,4 fois la taille de la banque), 158 poussaient sur milieu sélectif. A la suite des différents tests, aucun gène conférant une résistance avérée au Cd n'a été sélectionné.

	ADNc ΔYap1	ADNc ΔYap1 vs ADNc Génoscope Balen	ADNc ΔYap1 vs ADNc Génoscope Paal
Balen	Protéine hypothétique 1	0	0
	Protéine hypothétique 2	0	0
	Protéine hypothétique 3	0	0
	Protéine hypothétique 4	42	10
	Protéine hypothétique 5	0	0
	HSP 12 (champignon) HSP 12 (champignon)	1	2
	Saccharopine deshydrogenase (animal) Saccharopine deshydrogenase (animal)	33	1
Paal	Proteine hypothétique 6	0	0
	Protéine ribosomale 40S S15 (champignon)	8	5
	Protéine ribosomale 60S L3 (plante)	6	3

Tableau 21: Annotation des séquences d'ADNc des banques métatranscriptomiques de Balen, et Paal complémentant le phénotype du mutant de levure ΔYap1 sensible au cadmium. Une recherche de ces séquences dans les jeux de séquences d'ADNc de Balen et de Paal fournis par le Génoscope a été effectuée à l'aide du programme TBLASTN. HSP, Heat Shock Protein.

3.3.4 Analyse des séquences d'ADNc restaurants le phénotype ΔYap1

Comme précédemment, les séquences d'ADNc capables de restaurer le phénotype de sensibilité au cadmium du mutant ΔYap1 ont été annotées et recherchées dans les jeux de séquences d'ADNc de Balen et de Paal fournis par le Génoscope.

Les séquences obtenues par criblage de la banque de Balen correspondent à 5 protéines hypothétiques différentes, à 2 saccharopine déshydrogenase d'origine animale identiques et à 2 HSP12 (Heat Shock Protein) fongiques identiques. Pour les HSP12 et les saccharopine déshydrogenase les ADNc apparaissent pleine longueur. Certaines de ces séquences contiennent des homologues dans les jeux de séquences d'ADNc de Balen et de Paal (Tableau 21). C'est le cas notamment de la protéine hypothétique 4 présente 42 fois dans le jeu de séquence de Balen et 10 fois dans celui de Paal. Cependant, ce groupe de séquences identiques ne défini pas un contig avec les paramètres d'assemblage utilisés. En revanche, la saccharopine déshydrogenase est identique aux 33 séquences formant un même contig retrouvées dans le jeu de séquences d'ADNc de Balen (cf 2.2.1).

Le criblage de la banque de Paal a permis d'identifier une protéine hypothétique, une protéine ribosomale de champignon et une protéine ribosomale de plante (Tableau 21). Ces deux dernières protéines ont été retrouvées plusieurs fois dans les 2 jeux de séquences d'ADNc fournis par le Génoscope mais ces groupes de séquences ne définissent pas de contigs avec les paramètres utilisés.

3.4 Conclusions

La construction de banques métatranscriptomiques et l'analyse de plusieurs milliers de séquences d'ADNc par séquençage massif a permis de montrer:

- que 50 à 60% des protéines codées par ces séquences ne présentent aucuns homologues dans les bases de données publiques
- que les profils de répartition des protéines annotées en groupes fonctionnels sont relativement semblables entre ces sols.

Le criblage fonctionnel de banques métatranscriptomiques par expression hétérologue dans des mutants de levure sensibles aux métaux lourds a permis de montrer :

- que cette approche permet de découvrir de nouvelles protéines impliquées dans la résistance aux métaux
- que ces nouvelles protéines peuvent avoir une fonction connue (MTs, HSP) ou inconnue (saccharopine deshydrogenase, protéines hypothétiques) dans la résistance aux métaux
- que les protéines conférant une résistance aux métaux peuvent être présentent en plusieurs exemplaires dans les banques métatranscriptomiques
- que l'approche fonctionnelle permet de sélectionner des gènes conférant une résistance au Cd aussi bien de sols contaminés que non contaminés par ce métal.

4. Discussion

L'étude du métatranscriptome ou l'analyse des gènes exprimés par une communauté d'organismes permet de se rendre compte de l'état physiologique global dans lequel se trouve cette communauté à un instant précis.

Sur la base de ce postulat, nous avons entrepris de révéler et d'analyser l'ensemble des gènes exprimés par une communauté de microorganismes eucaryotes colonisant un sol contaminé aux métaux lourds (site de Balen) et de les comparer à ceux exprimés par une communauté de microorganismes eucaryotes colonisant un sol non contaminé (site de Paal) ou anciennement contaminé (site de Lommel).

La construction de banques de gènes environnementaux est soumise à différents biais liés aux étapes d'échantillonnage et d'extraction des acides nucléiques. Ces biais, propres aux études *in situ*, affectent la représentativité des organismes réellement présents dans un échantillon environnemental.

L'échantillonnage du sol sur les sites de Balen, Paal et Lommel a été effectué en 20 points de prélèvement sur une surface d'environ $40m^2$. Afin d'être représentatif de la diversité biologique présente et pour détecter d'éventuelles espèces rares, une grande quantité de sol a été récupérée en chaque point d'échantillonnage. Dans le cas d'étude moléculaire des microorganismes du sol, il est nécessaire de limiter l'échantillonnage des macroorganismes (animaux et végétaux), susceptibles de contaminer les banques de gènes, par un tamisage du sol à 2 mm. Toutefois, la contamination de nos jeux de séquences par des séquences de plantes « supérieures » et d'animaux prouve que de fines racines de plantes (*Pinus sylvestris*) et des représentants de la micro-mésofaune des sols (nématodes, annélides, collemboles,

acariens…) sont tout de même capables de passer à travers le tamis. Un tamisage plus fin serait nécessaire, mais il engendrerait une plus forte perturbation du sol capable de modifier la composition et le profil d'expression de gènes de la communauté de microorganismes. Enfin, un moyen de limiter la modification du profil d'expression de gènes de la communauté de microorganismes entre l'échantillonnage et l'extraction des acides nucléiques a été d'effectuer rapidement une congélation du sol dans de l'azote liquide et un stockage à -70°C. D'autres méthodes comme la lyophilisation du sol et un stockage à -20°C ou une congélation à -80°C dans du glycérol sont également appropriées (Sessitsch *et al.*, 2002). Tous les sols ont été prélevés, tamisés, congelés et stockés de la même façon et le plus rapidement possible afin de limiter les biais liés à l'échantillonnage.

Les échantillons de sols congelés ont ensuite été rapidement utilisés pour l'extraction des acides nucléiques des organismes présents. De nombreuses études ont mis en évidence une corrélation entre la méthode d'extraction utilisée et la composition de la communauté microbienne révélée, indépendamment des propriétés (physico-chimiques, couvert végétal) du sol étudié et du rendement d'extraction obtenu (Martin-Laurent *et al.*, 2001; Sessitsch *et al.*, 2002; Sagova-Mareckova *et al.*, 2008). Ces observations concernent uniquement les bactéries et on peut supposer qu'elles ne s'appliquent pas aux microorganismes eucaryotes de plus grande taille et donc plus facilement exposés aux méthodes de lyse cellulaire. De plus, le protocole d'extraction utilisé dans notre étude (Bailly *et al.*, 2007) fait intervenir deux étapes de broyage dont une à sec dans de l'azote liquide reconnue pour casser plus efficacement les membranes cellulaires des organismes présents (Sessitsch *et al.*, 2002). Enfin, cette méthode d'extraction a été employée sur plusieurs centaines de grammes de sols (400g pour Paal et Lommel et 250g pour Balen) afin de limiter les biais liés à la distribution hétérogène des microorganismes dans les sols et d'inclure un plus grand nombre d'organismes présents en faible densité dans les sols, notamment la mésofaune.

Cette méthode d'extraction reproductible et adaptable à différents types de sols (cf Chapitre 1) a permis la co-extraction des ADN et des ARN à partir de nos 3 sols et la construction de banques ribosomiques d'ADNr et d'ARNr 18S et de banques d'ADNc environnementales.

L'analyse de la petite sous unité ribosomique 18S permet de décrire la diversité des microorganismes eucaryotes dans différents environnements. Ces dernières années, cette approche a notamment été utilisée pour révéler la diversité des protistes dans les eaux douces et marines ainsi que dans des sédiments (par exemple, Lopez-Garcia *et al.*, 2001; Dawson et

Pace, 2002; Stoeck *et al.*, 2003). Cependant, peu d'études ont été menées afin d'estimer la diversité taxinomique de l'ensemble des microorganismes eucaryotes dans les sols (Lawley *et al.*, 2004; Tringe *et al.*, 2005 ;Moon-van der Staay *et al.*, 2006; Lesaulnier *et al.*, 2008). Ces rares études ont montré que l'emploi d'amorces universelles eucaryotes permet d'identifier tous les groupes, y compris les plantes et les métazoaires et pas seulement les microorganismes. Des profils d'affiliation taxinomiques variables ont ainsi pu être obtenus pour chaque sol.

La diversité taxinomique eucaryote présente dans les échantillons de sols prélevés sur les sites de Balen, Paal et Lommel a été évaluée par l'analyse d'une ou deux portions du gène codant la petite sous unité ribosomique 18S amplifiées par PCR à partir de l'ADN et de l'ARN environnemental.

Le faible pourcentage de séquences chimériques (compris entre 0,25 et 1%) détectées dans ces différentes banques assure de leur bonne qualité. La génération de tels artéfacts, souvent inférieur à 10% (Richards *et al.*, 2005; Nikolaev *et al.*, 2006), a certainement été limitée en utilisant un faible nombre de cycles d'amplification et en réalisant plusieurs PCR indépendantes pour la construction de chacune de ces banques. La détection de ces séquences chimériques à l'aide du logiciel Check chimera reste néanmoins très délicate et parfois subjective. Il n'est pas à exclure que dans notre étude, ainsi que dans d'autres, certaines chimères soient passées inaperçues.

A l'image des résultats du chapitre 1, les banques ribosomiques du premier tiers du 18S comprennent des séquences de tous les phylums eucaryotes avec une nette dominance des opisthokontes (champignons et animaux). Toutes les séquences animales sont affiliées à des groupes taxinomiques typiques de la micro et de la méso-faune (nématodes, acariens, collemboles…) alors que les séquences de basidiomycètes et d'ascomycètes dominent les séquences fongiques comme cela a déjà été observé dans les forêts tempérées (O'Brien *et al.*, 2005). Pour les groupes de « protistes », les séquences d'amoebozoa, d'alveolata et de rhizaria (exclusivement représenté par les cercozoa) sont mieux représentées que les séquences d'heterokonta et d'excavata. Concernant les séquences du derniers tiers du 18S, une répartition taxinomique relativement similaire est observée sauf pour le phylum des metazoa. La quasi disparition des séquences animales dans ce jeu de données tend à vérifier l'hypothèse émise dans d'autres études quant à l'utilisation de plusieurs jeu d'amorces PCR pour recouvrir complètement la diversité des communautés eucaryotes (Stoeck *et al.*, 2006; Jeon *et al.*, 2008).

Par ailleurs, mêmes si les profils taxinomiques sont relativement semblables entre les différents sols quelle que soit la portion du gène 18S analysée (Fig. 31 et 33), des différences significatives d'abondance de certains phylums entre sites d'étude existent. D'autres paramètres, tels que la faible redondance des phylotypes entre sol (Fig 32 et 34) et les valeurs des estimateurs de richesse spécifique (Tableau 14, 15 et 16), indiquent que ces sols abritent des communautés d'organismes eucaryotes différentes et que celle du sol contaminé aux métaux lourds (site de Balen) est relativement plus diversifiée. Les rares études s'intéressant aux communautés de microorganismes eucaryotes dans des environnements pollués (Amaral Zettler *et al.*, 2002; Aguilera *et al.*, 2006) font état d'une diversité inattendue. D'autres études, s'intéressant à des organismes eucaryotes, ont mis en évidence une augmentation de la diversité suite à une pollution. Une hypothèse plausible serait que la disparition d'organismes majoritaires et sensibles à la pollution favorise la colonisation du milieu par une diversité plus importante d'organismes résistants. Notre étude représente la première analyse comparative de communauté de microorganismes eucaryotes présente dans un sol pollué et dans un sol non pollué.

Enfin, des différences significatives existent entre l'abondance des séquences d'ADNr et d'ARNr réparties dans un même phylum (Fig.35). En écologie microbienne, il est admis que les ARNr décrivent mieux la diversité des organismes métaboliquement actifs dans une communauté (Muttray et Mohn, 1999; Mills *et al.*, 2005). Cependant, chez les eucaryotes, le nombre de copies du gène codant l'ARNr 18S est très variable entre taxons, tout comme le nombre de noyaux par unité de cytoplasme. Par conséquent, afin de mieux décrire l'activité physiologique des microorganismes eucaryotes, une comparaison des proportions relatives en ADNr 18S, en ARNr 18S et en ADNc de chaque phylum sera effectuée une fois le séquençage aléatoire des banques métatranscriptomiques terminé (cf chapitre 1). Par ailleurs, la faible redondance entre les jeux de séquences d'ADNr et d'ARNr (Tableau 13) ainsi que la forme des courbes de raréfactions qui n'atteignent pas de plateaux (Fig. 36) indique que des efforts de séquençage supplémentaires combinés à l'utilisation d'amorces PCR spécifiques à certains groupes taxinomiques sont nécessaires pour décrire complètement la diversité des communautés de microorganismes eucaryotes présente dans ces 3 sols.

Une autre façon d'étudier des communautés de microorganismes dans leur environnement est d'utiliser l'approche métatranscriptomique. Cette approche, basée sur l'analyse des gènes exprimés par l'ensemble d'une communauté d'organismes, a récemment été développée pour l'étude de la structure et de la fonction de communautés bactériennes

dans des eaux marines (Poretsky *et al.*, 2005; Frias-Lopez *et al.*, 2008; Gilbert *et al.*, 2008; Poretsky *et al.*, 2009) et dans les sols (Leininger *et al.*, 2006; Urich *et al.*, 2008). Les résultats obtenus dans ces études ont permis de mettre en évidence les fonctions effectuées par les organismes actifs d'une communauté et de détecter une diversité de nouveaux gènes susceptibles de coder de nouvelles fonctions. Cette technologie a également été développer afin de décrire spécifiquement des communautés de microorganismes eucaryotes dans des boues activées (Grant *et al.*, 2006) et dans un sol forestier (Bailly *et al.*, 2007)..

Les banques métatranscriptomiques construites à partir de nos 3 sols ont été analysées suivant deux approches. D'une part, par une approche basée sur le séquençage aléatoire des banques du sol contaminé aux métaux lourds (site de Balen) et du sol non contaminé (site de Paal) et d'autre part, par une approche basée sur le criblage fonctionnel de nos 3 banques dans un hôte hétérologue eucaryote.

Les données partielles du séquençage des banques de Balen et de Paal ont été obtenues au cours de la rédaction de ce manuscrit. L'analyse de ces résultats est incomplète et nos observations sont à prendre au conditionnel.

Dans un premier temps, la plus ou moins forte redondance contenue dans chaque jeu de séquences (3,5% pour Paal et 24% pour Balen) et la forme des courbes de raréfactions indiquent que le sol non contaminé (site de Paal) abriterait une plus grande « diversité moléculaire » que le sol contaminé aux métaux lourds (site de Balen). Toutefois, la taille variable des séquences contenues dans chaque banque, en moyenne plus petite pour le sol de Balen (420 pb contre 480 pb pour Paal) peut être à l'origine de cette différence. En effet, la méthode de construction de nos banques, faisant intervenir une étape de clonage, est soumise à un biais lié à la dégradation des ARNm et à un clonage préférentiel des petits inserts d'ADNc. Récemment, de nouvelles technologies de séquençage, dîtes de pyroséquençage (par exemple, 454 de Roche, SOLEXA de Illumina et SOLiD3 de ABI) ont vu le jour et permettent le séquençage direct d'ADN ou d'ADNc sans aucune étape de clonage (Medini *et al.*, 2008). Ces technologies permettent aussi d'augmenter considérablement le nombre de paires de base séquencées par « run » et de diminuer le coût par base séquencée

Ensuite, le fort pourcentage de séquences d'ADNc sans homologues dans les bases de données, avoisinant les 60% dans les deux banques, montre que cette approche est susceptible d'identifier de nouvelles protéines hypothétiques (Fig. 42). Là encore, il se peut que des séquences tronquées, trop courtes pour être identifiées, soient comptabilisées comme de nouvelles protéines hypothétiques. En effet, les études métatranscriptomiques menées chez les procaryotes et les eucaryotes dans les sols et les boues activées révèlent un pourcentage de

séquences sans homologues d'environ 20% (Grant *et al.*, 2006; Bailly *et al.*, 2007; Urich *et al.*, 2008). De même, notre annotation manuelle de jeux de séquences plus limités (cf Chapitre 1) conduisait à un pourcentage de nouvelles séquences bien inférieur (de l'ordre de 30%). A l'opposé, les banques métatranscriptomiques construites à partir d'eau marine montrent contiennent entre 50 et 90% de nouvelles séquences hypothétiques (Frias-Lopez *et al.*, 2008; Gilbert *et al.*, 2008; Poretsky *et al.*, 2009).

Enfin, la répartition des séquences codant des protéines de fonctions connues en catégories fonctionnelles indiquent que les communautés présentent dans les 2 sols ont un profil d'expression génique relativement identique (Fig. 43). On s'aperçoit cependant que des différences existent quant à la diversité et l'abondance de fonctions exprimées à l'intérieur d'une même catégorie fonctionnelle. Par exemple, au sein de la catégorie fonctionnelle « réponses aux stress » une abondance et une diversité plus importante de gènes impliqués dans la réponse au stress oxydant est observée pour le sol contaminé aux métaux lourds (Tableau 19) alors qu'aucune différence significative n'est observée entre le jeu de données de Balen et de Paal lorsque l'on considère cette catégorie de façon globale. Le faible nombre de séquences analysé, pour l'instant, ne permet pas d'émettre d'hypothèses quant aux différences de fonctions exprimées entre une communauté de microorganismes eucaryotes colonisant un sol contaminé aux métaux lourds et celle d'un sol non contaminé. L'analyse de 20.000 inserts d'ADNc supplémentaires pour chaque sol est à venir.

Ces deux banques métatranscriptomiques ainsi que celle du sol anciennement contaminé (site de Lommel) ont également été exploitées par criblage fonctionnel dans la levure *Saccharomyces cerevisiae*. Cette approche originale, développée dans l'équipe, avait permis la complémentation fonctionnelle d'un mutant de levure auxotrophe pour l'histidine par une banque d'EST environnementales (Bailly *et al.*, 2007). Cette technique avait déjà été employée chez les procaryotes par exemple pour la complémentation d'une souche d'*E. coli* déficiente pour un transporteur Na^+/H^+ avec une banque métagénomique de sol. Cela avait conduit à identifier deux nouveaux gènes codant des transporteurs Na^+/H^+ à partir d'une banque contenant 1.480.000 clones (Majernik *et al.*, 2001).

Le criblage fonctionnel de nos banques a permis de sélectionner des gènes capables de complémenter le phénotype de sensibilité au cadmium de deux mutants de levures (ΔYcf1 et ΔYap1). Les gènes de résistance au cadmium ont été obtenus à des fréquences variables suivant la banque criblée et le mutant utilisé. Pour aucun des deux mutants utilisés nous n'avons isolé de gènes homologues à *Ycf1* ou *Yap1*.

La complémentation du phénotype sensible du mutant ΔYcf1 a conduit à isoler majoritairement des protéines riches en résidus cystéine nommées métallothionéines. Les métallothionéines sont des peptides de faible masse moléculaire (25 à 60 acides aminés) issus de la traduction d'un ARN messager (Cobbett et Goldsbrough, 2002). Elles sont riches en cystéine et chélatent les ions métalliques par coordination avec les thiolates. Par conséquent, l'accumulation des ions cadmium dans le cytoplasme du mutant ΔYcf1, suite à l'absence de transport du complexe glutathion-Cd dans la vacuole, est neutralisée par l'expression constitutive de ces protéines via le vecteur dans lequel la banque d'ADNc a été construite. L'approche utilisée a permis d'isoler une métallothionéine fongique (Lommel), une métallothionéine potentiellement apparentée à celle de mollusques (Balen) et une nouvelle potentielle famille de métallothionéines décrites par 17 nouvelles séquences se répartissant en 5 sous-familles (Lommel et Paal). En effet, ces nouvelles métallothionéines, classées en différentes sous-familles sur la base de leur séquence protéique et plus particulièrement suivant le motif dessiné par l'arrangement de leurs résidus cystéine, ne s'apparentent à aucune famille de métallothionéines décrites dans les bases de données. Toutefois, des caractéristiques communes avec la seule famille de métallothionéines de protistes (décrite exclusivement chez des Ciliés du genre *Tetrahymena*) comme leurs grandes tailles (entre 110 et 132 acides aminés dont 24 à 27 résidus cystéines), la présence d'un triplet de cystéines à leur extrémité N-terminale et l'absence d'introns dans leurs séquences génomiques (Franck Lejzerowicz, communication personnelle) permettent d'avancer l'hypothèse d'une affiliation à un groupe de Ciliés de ces protéines.

Deux autres catégories de protéines ont pu être identifiées suite au criblage de la banque de Balen. Il s'agit d'une « Zn-binding alcohol dehydrogenase » et plus particulièrement de sa portion C-terminale correspondant au domaine de liaison au NADP (domaine dit de NADB-Rossmann). Le NADP est un coenzyme d'oxydoréduction impliqué dans la régulation du stress oxydant. Il est notamment connu pour être impliqué dans le fonctionnement de la glutathion réductase responsable de la réduction du glutathion oxydé en glutathion réduit via l'oxydation du NADPH en NADP. On peut donc supposer que l'expression de cette protéine tronquée et sa fixation au NADP entraine une diminution de la concentration cytoplasmique en NADP et déplace l'équilibre NADP/NADPH. Ce déplacement d'équilibre entraine alors une augmentation de l'oxydation du NADPH en NADP via l'action, notamment, de la glutathion réductase et la réduction du glutathion impliqué dans la chélation du Cd.

La dernière catégorie de protéines identifiées par le criblage de la banque de Balen dans le mutant ΔYcf1 est constituée de 7 protéines hypothétiques. Parmi ces 7 séquences, 4 sont

identiques entre elles et aux 265 séquences formant un même contig retrouvées dans le jeu de séquences d'ADNc de Balen fourni par le Génoscope. Notre approche a permis d'isoler de nouvelles protéines impliquées dans la résistance aux métaux et qui, pour certaine, semble être surexprimées par la communauté de microorganismes eucaryotes colonisant un sol contaminé par des métaux lourds.

Le criblage des banques par complémentation fonctionnelle du mutant ΔYap1 a permis d'identifier un ensemble de protéines plus divers. Il s'agit d'une part de protéines HSP12 (Balen) d'origine fongique connues pour jouer un rôle dans le maintien de l'intégrité de la membrane plasmique de *S. cerevisiae* lors d'un stress oxydant (Shamrock et Lindsey, 2008) et d'autre part, de protéines ribosomales (Paal) dont la surexpression chez le champignon basidiomycète *Ganoderma lucidum* lors d'un traitement au Cd a récemment été démontrée (Chuang *et al.*, 2009). La surexpression de la protéine ribosomale L10a de *Chlamydomonas* dans la souche ΔYap1 de *S. cerevisiae* conduit aussi à une meilleure résistance au stress oxydant par stimulation de la synthèse de pigments caroténoïdes (Mendez-Alvarez *et al.*, 2000).

Ce crible a également permis d'identifier 2 saccharopine déshydrogénase d'origine potentiellement animale impliquées dans la synthèse et/ou la dégradation de la lysine via l'oxydation du NADH en NAD et/ou la réduction du NAD en NADH. Des analyses supplémentaires permettront de mieux caractériser le rôle de cette protéine dans la réponse au stress oxydant. Par ailleurs, les inserts d'ADNc codant ces saccharopine déshydrogénase sont homologues entre eux et aux 33 séquences formant un même contig retrouvées dans le jeu de séquences d'ADNc de Balen. Il semblerait donc que ces protéines soient surexprimées par la communauté de microorganismes eucaryotes colonisant un sol contaminé par des métaux lourds.

Enfin, comme pour ΔYcf1, le criblage des banques de Paal et de Balen a permis d'isoler 6 protéines hypothétiques différentes susceptibles de coder de nouvelles protéines impliquées dans la résistance aux métaux. Des analyses supplémentaires seront réalisées afin de mieux caractériser le rôle de ces protéines.

L'identification de différentes protéines de résistance au cadmium par complémentation fonctionnelle de mutants de levures sensibles à ce métal avec des banques métatranscriptomiques de sol contaminé, anciennement contaminé et non contaminé au cadmium montre qu'une diversité de mécanisme de résistance existe au sein des communautés qui colonisent ces sols et qu'une capacité à résister aux métaux préexiste dans un sol non contaminé. Des analyses supplémentaires permettront de vérifier la spécificité de

ces mécanismes vis-à-vis du cadmium. Enfin, cette approche semble ouvrir des perspectives intéressantes quant à son utilisation pour la détection de gènes d'intérêt en biotechnologie.

Conclusions et Perspectives

Le sol est un des environnements sur Terre abritant la plus grande diversité de microorganismes (Torsvik et Ovreas, 2002). Depuis une trentaine d'année, de nombreuses études d'écologie microbienne se sont intéressées à vouloir identifier les communautés bactériennes du sol et à comprendre leurs interactions avec leur environnement sans toutefois prendre en compte toute la diversité de microorganismes eucaryotes présente. Depuis quelques années seulement, des études de phylogénie moléculaire ont montré qu'une énorme diversité d'eucaryotes existe sur Terre et estiment que le sol, encore très peu étudié par rapport aux environnements aquatiques, recèle une diversité cryptique considérable.

Notre étude a eu pour objectifs de révéler et d'analyser *in situ* la diversité et les fonctions exprimées par l'ensemble d'une communauté de microorganismes eucaryotes colonisant un sol contaminé aux métaux lourds et des sols témoins anciennement contaminé et non contaminé. Ce travail ne s'est pas limité pas à la seule étude d'une ou plusieurs espèces modèles mais a permis de mettre en évidence une grande partie de la diversité des microorganismes eucaryotes d'un sol, connus ou inconnus, cultivables ou non cultivables.

L'étude de la diversité taxinomique via l'analyse du gène codant la petite sous-unité ribosomique 18S montre qu'un sol contaminé semble abriter une diversité de microorganismes eucaryotes plus importante qu'un sol non contaminé de même origine géologique ou qu'un sol forestier plus riche en matière organique. Par ailleurs, même si la diversité eucaryote présente dans ces différents sols n'a pas été complètement décrite, tous les phylums d'eucaryotes connus y ont été identifiés et de potentiels nouveaux groupes taxinomiques y ont été révélé. Cela a été confirmé par une analyse phylogénétique d'un ensemble de séquences issues de différents sols et d'eaux douces. Celles-ci décrivent un nouveau clade de microorganismes eucaryotes appartenant au groupe des Cercozoa (Rhizaria) (cf Chapitre 2). Enfin, il semble que notre perception de la diversité des microorganismes eucaryotes dans l'environnement soit biaisée par la nature de la séquence nucléique analysée (cf Chapitre 1) voire même de la portion du gène cible (cf Chapitre 3).

La diversité fonctionnelle de ces mêmes communautés de microorganismes eucaryotes à été entrevue par une approche originale basée sur l'étude de leur métatranscriptome (cf Chapitre 3). Une analyse comparative partielle et encore incomplète du séquençage aléatoire de banques d'ADNc d'un sol contaminé et non contaminé aux métaux lourds révèle l'existence d'une grande diversité moléculaire. La majorité des gènes exprimés par les microorganismes présents dans ces sols semble avoir une fonction inconnue et la classification des gènes annotés en catégories fonctionnelles ne permet pas, pour le moment, de constater un effet significatif de la pollution métallique sur la diversité et l'abondance des

fonctions exprimées par ces communautés. Le criblage de ces banques par complémentation fonctionnelle de mutants de levures sensibles au cadmium a permis de découvrir de nouvelles protéines impliquées dans la résistance aux métaux. Aussi bien des protéines hypothétiques de fonctions inconnues que des protéines de fonctions connues pour être directement (métallothionéines, HSP) ou indirectement (saccharopine déshydrogenase, protéines ribosomales) impliquées dans la résistance au cadmium ont été identifiées par cette méthode à partir des sols contaminé et non contaminé aux métaux lourds..

Tous ces résultats démontrent l'intérêt de l'approche métatranscriptomique tant pour l'étude des écosystèmes et la compréhension des interactions entre les microorganismes et leur environnement que pour la compréhension de processus fondamentaux tels que les mécanismes cellulaires permettant une résistance aux métaux lourds. Une autre voie d'intérêt est l'étude du métatranscriptome pour la recherche de nouveaux biocatalyseurs et molécules actives. De nombreuses perspectives immédiates et à long terme sont donc envisageables.

En termes de perspectives, dans un premier temps, le données partielles de séquençage seront complétés rapidement par plusieurs milliers de séquences supplémentaires et seront analysées avec des outils bioinformatiques et statistiques plus adaptés que ceux utilisés jusqu'à présent. Ceci permettra d'affiner la classification fonctionnelle des protéines identifiées et surtout d'effectuer une affiliation taxinomique à chacune d'entre elles. La structure et la fonction d'une communauté de microorganismes eucaryotes colonisant un sol contaminé et non contaminé aux métaux lourds pourront alors être comparées. Nos analyses préliminaires suggèrent toutefois que le jeu de données de séquences d'ADNc puisse être trop limité pour aboutir à une analyse fonctionnelle des communautés. Des méthodes de séquençage à très haut débit devraient être utilisées pour répondre à ce problème. Dans cette optique, des analyses complémentaires de diversité et de phylogénie moléculaire seront effectuées sur les jeux de données d'ADN et d'ARN ribosomiques présentées dans le chapitre 3. La construction d'arbres phylogénétiques sera réalisée pour chaque phylum à partir des séquences des différentes portions du gène 18S. De plus, la comparaison d'autres paramètres, comme par exemple les indices de Shannon ou le nombre de phylotypes majoritaires, permettront de formuler de nouvelles hypothèses quant à l'impact des métaux lourds sur la structure d'une communauté d'organismes eucaryotes.

Les gènes de résistance isolés par le criblage fonctionnel des banques métatranscriptomique dans des mutants de levures sensibles au cadmium seront comparés à

ceux obtenus, par nos partenaires de Nancy, par le criblage fonctionnel de ces mêmes banques dans un mutant de levure sensible au zinc. Leur capacité de résistance sera testée avec des concentrations plus importantes en métal. Des tests croisés, avec différents métaux (Cd, Zn, Cu et Co) ainsi qu'avec un agent oxydant (ménadione), seront réalisés afin de vérifier l'éventuelle spécificité des mécanismes de résistance vis-à-vis d'un métal.

Dans un second temps, il sera intéressant de mieux caractériser la fonction de ces gènes de résistance par des tests biochimiques plus poussés afin de comprendre le mode d'action des différentes protéines codées. Par ailleurs, le dessin d'amorces spécifiques à partir des séquences de ces gènes permettra d'effectuer une recherche des copies génomiques de ces gènes et de leurs séquences promotrices par marche chromosomique à partir de l'ADN métagénomique extrait des sols. L'obtention de telles séquences permettra d'exprimer ces gènes dans d'autres hôtes eucaryotes hétérologues et de révéler leur mode de régulation transcriptionnel en réponse à un stress métallique ou oxydant. Nos banques de gènes pourront également être criblées par complémentation fonctionnelle de levures mutées pour d'autres phénotypes comme par exemple la résistance à la dessiccation ou par expression hétérologue chez la levure dans le but d'identifier de nouvelles enzymes d'intérêts.

Les jeux de données fournis par le Génoscope pourront aussi faire l'objet d'une recherche de nouveaux membres de familles de gènes connues pour participer à la détoxication des métaux par une approche bioinformatique ou par PCR à l'aide d'amorces appropriées obtenues à partir de régions conservées de gènes connus.

Enfin, la diversité taxinomique des microorganismes eucaryotes présents dans ces sols pourra être mieux évaluée avec l'utilisation de plusieurs jeux d'amorces universelles ou groupes spécifiques, comme cela a été réalisé avec succès au sein du laboratoire pour le groupe des foraminifères (Franck Lejzerowicz, communication personnelle). A l'image de l'étude de phylogénie réalisée dans le chapitre 2, des amorces spécifiques pourront également être dessinées à partir de séquences divergentes affiliées à certains phylums, comme les Amoebozoa.

Matériels et Méthodes

Site	pH dans l'eau	CEC (méq/kg)	% Humidité	Carbone (C) Organique (g/ kg)	Azote (N) total (g/kg)	C/N	MO (g/kg)
Paal	4,68	26	4,8	1,75	0,0719	24,3	4,9
Lommel	4,66	25	4	9,2	0,4	20	15,8
Balen	5,79	3,45	8,7	17,9	0,932	19,2	30,9

Tableau 22: Analyses physico-chimiques des sols prélevés en novembre 2006 et juin 2008. (CEC : Capacité d'Echange Cationique ; MO : Matière Organique)

	Végétaux et Lichens	Champignons
Paal	*Agrostis capillaris* *Betula sp.* *Campylopus introflexus* *Campylopus sp.* *Cladonia sp.* *Quercus petraea* *Pinus sylvestris* *Polytrichum piliferum* *Populus nigra*	*Amanita muscaria* *Laccaria sp* *Lactarius mommensus* *Lactarius rufus* *Leccinum sp.* *Paxillus involutus* *Scleroderma sp.* *Xerocomus badius*
Lommel	*Agrostis capillaris* *Algues vertes* *Betula sp.* *Cladonia sp.* *Deschampsia flexuosa* *Pinus sylvestris* *Polytrichum piliferum*	*Amanita muscaria* *Amanita rubescens* *Boletus edulis* *Laccaria bicolor* *Laccaria sp.* *Lactarius rufus* *Paxillus involutus* *Scleroderma sp.* *Suillus bovinus* *Suillus luteus*
Balen	*Betula sp.* *Bromus mollis* *Bryum sp.* *Cerastium sp.* *Ceratodon sp.* *Holcus lanatus* *Pinus sylvestris* *Pleuridiuma acuminatum* *Rumex acetosella* *Rumex crispus* *Rumex scutatus* *Senecio jacoboea*	*Hebeloma sp.*

Tableau 23: Inventaire des espèces végétales et fongiques identifiés visuellement sur les sites de Paal, Lommel et Balen.

1. Sols étudiés

Les échantillons de sols ont été prélevés dans le nord-est de la Belgique (Province du Limbourg) sur les sites de Lommel Sahara (LS, 51° 14' N, 5° 15' E), de Paal (PZ, 51° 04' N, 5° 10' E) et de Balen (BA, 51° 10' N, 5° 10' E). Il s'agit de sols sableux, pauvres en matière organique sur lesquels se développent des pins sylvestres (*Pinus sylvestris*) et quelques bouleaux (*Betula sp*.). Le site de Lommel Sahara, reboisé en 1975, a été contaminé par des métaux lourds dispersés par une ancienne fonderie de zinc (Vangronsveld *et al.*, 1995). Le site de Balen est contaminé par des métaux lourds dispersés par une fonderie de zinc, toujours en activité, et est utilisé comme site pilote pour des études de phytorémédiation (van der Lelie *et al.*, 2001).

Les prélèvements ont été effectués, en novembre 2006 pour les sites LS et PZ et en juin 2008 pour le site BA, en 20 points différents de chaque site sur une profondeur de 20 cm et sur 5 cm de diamètre. Les échantillons de sols d'un même site ont été réunis et tamisés à 2 mm, puis congelés dans de l'azote liquide et stockés à -70°C. Les dosages du Zn, du Cd et du Pb ont été réalisés par ICP-MS (Inductively Coupled Plasma-Mass Spectrometry) (Trassy et Mermet, 1984) pour chaque point d'échantillonnage afin de déterminer leur concentration totale (Fig. 46, 47 et 48) et leur concentration disponible environnementale (soluble dans l'eau) (Fig. 46, 47 et 48). Leur dosage dans les aiguilles de pins présents sur ces sites et situés à proximité des points d'échantillonnage permet de déterminer leur biodisponibilité (Fig. 46, 47 et 48). Le pool des échantillons de sols de chaque site a fait l'objet d'une analyse physico-chimique (Tableau 22). De plus, tous les carpophores de champignons ainsi que les végétaux présents ont été ramassés pour une identification visuelle (Tableau 23) et stockés dans un tampon d'extraction pour des analyses moléculaires ultérieures.

2. Extraction des acides nucléiques

Les extractions ont été réalisées sur un minimum de 200g de sol pour chacun des sites (450g pour LS, 450g pour PZ et 200g pour BA).

Les sols ont été broyés 3 min dans un broyeur de roche en agate préalablement congelé à -70°C. Un gramme de sol est alors placé dans un tube Eppendorf de 2 mL contenant 0,5 g de billes de verre (diamètre 0,1 mm), x µL de solution dénaturante (guanidine isothiocyanate 4 M, Tris-HCl 10 mM, Na2EDTA 1mM, pH 8,0), y µL de tampon de lyse (Tris-HCl 100 mM,

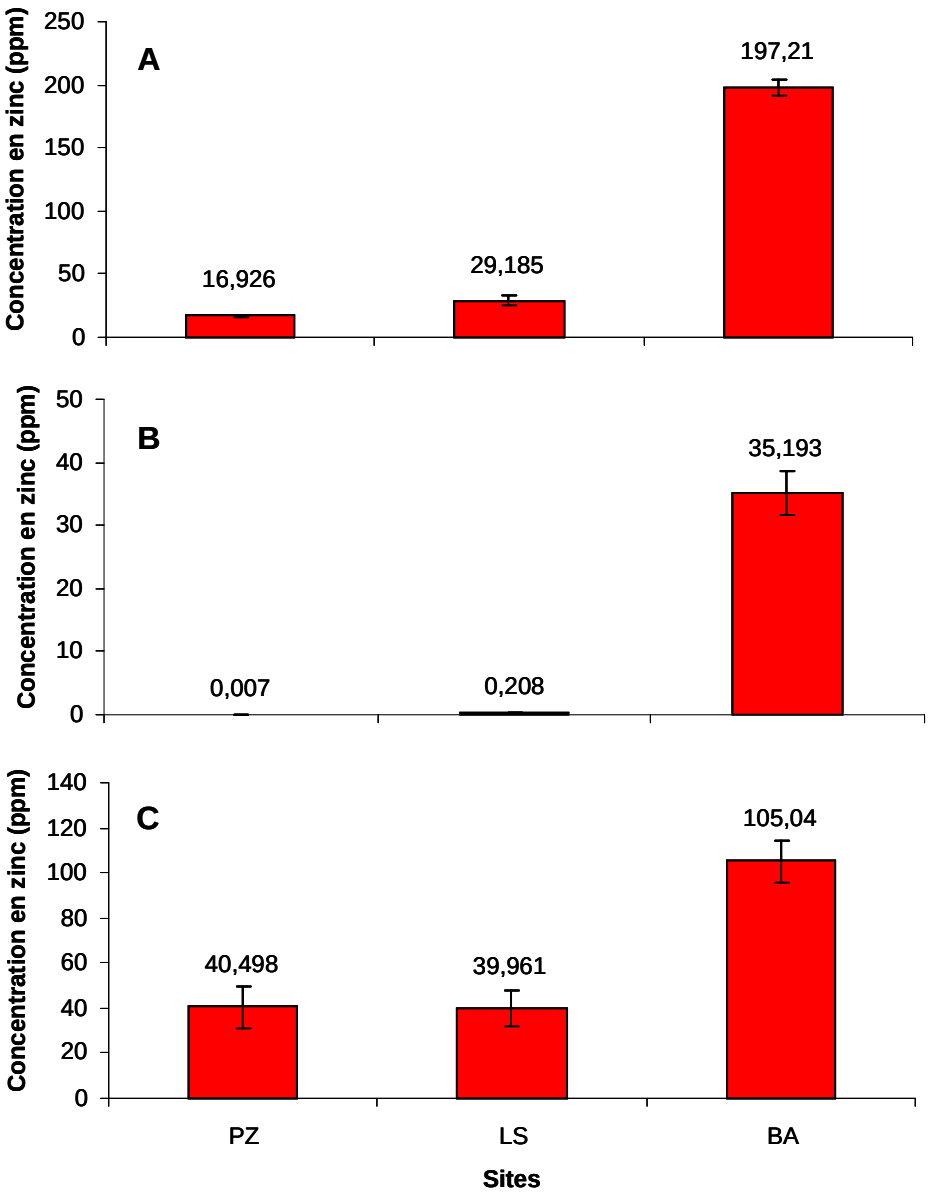


Figure 46: Dosage du zinc total (A), disponible (B) et biodisponible (C) présent dans les échantillons de sols (A et B) et d'aiguilles de pins (C) prélevés en novembre 2006 et juin 2008 sur les sites de Paal (PZ), Lommel (LS) et Balen (BA).

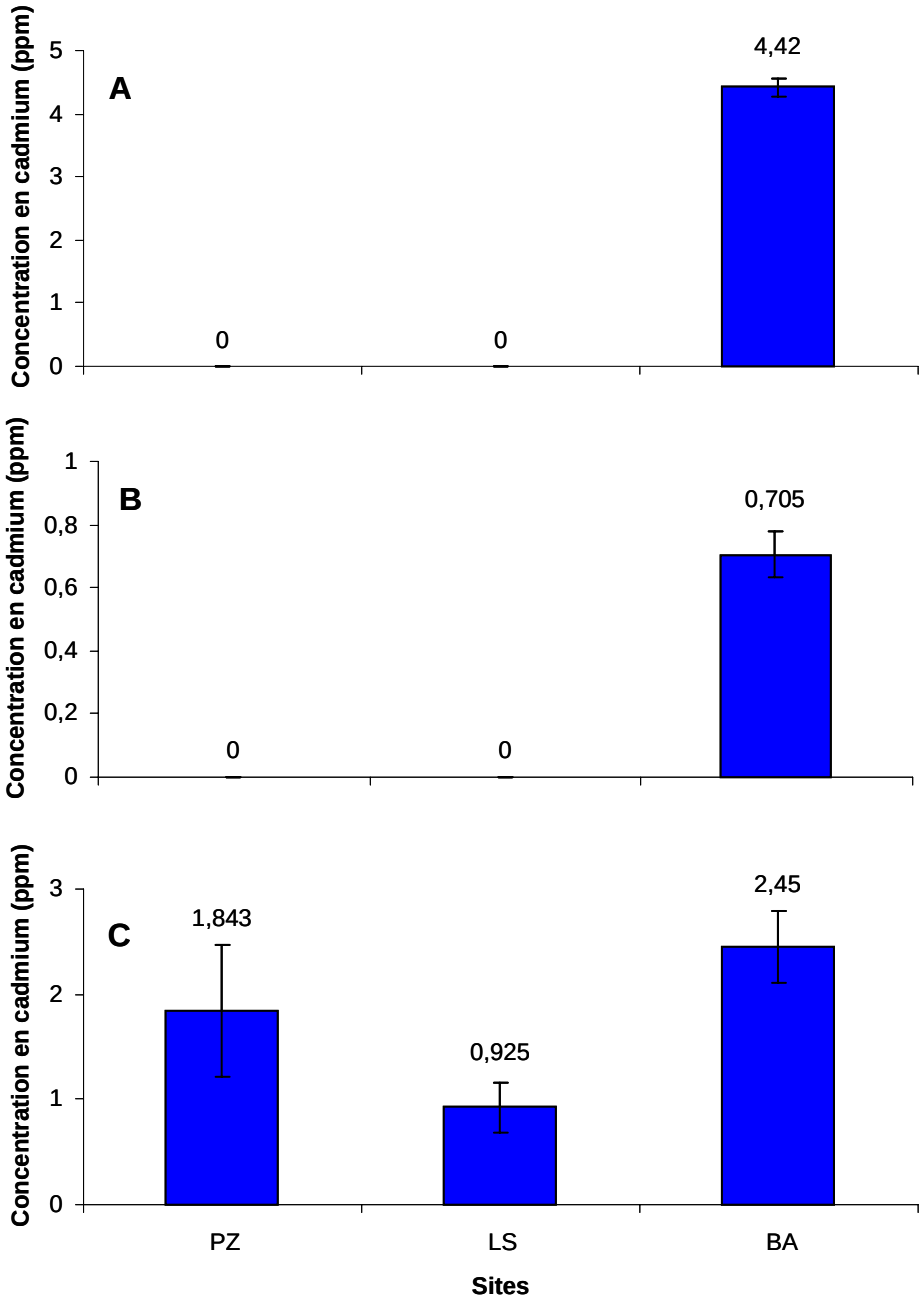


Figure 47: Dosage du cadmium total (A), disponible (B) et biodisponible (C) présent dans les échantillons de sols (A et B) et d'aiguilles de pins (C) prélevés en novembre 2006 et juin 2008 sur les sites de Paal (PZ), Lommel (LS) et Balen (BA).

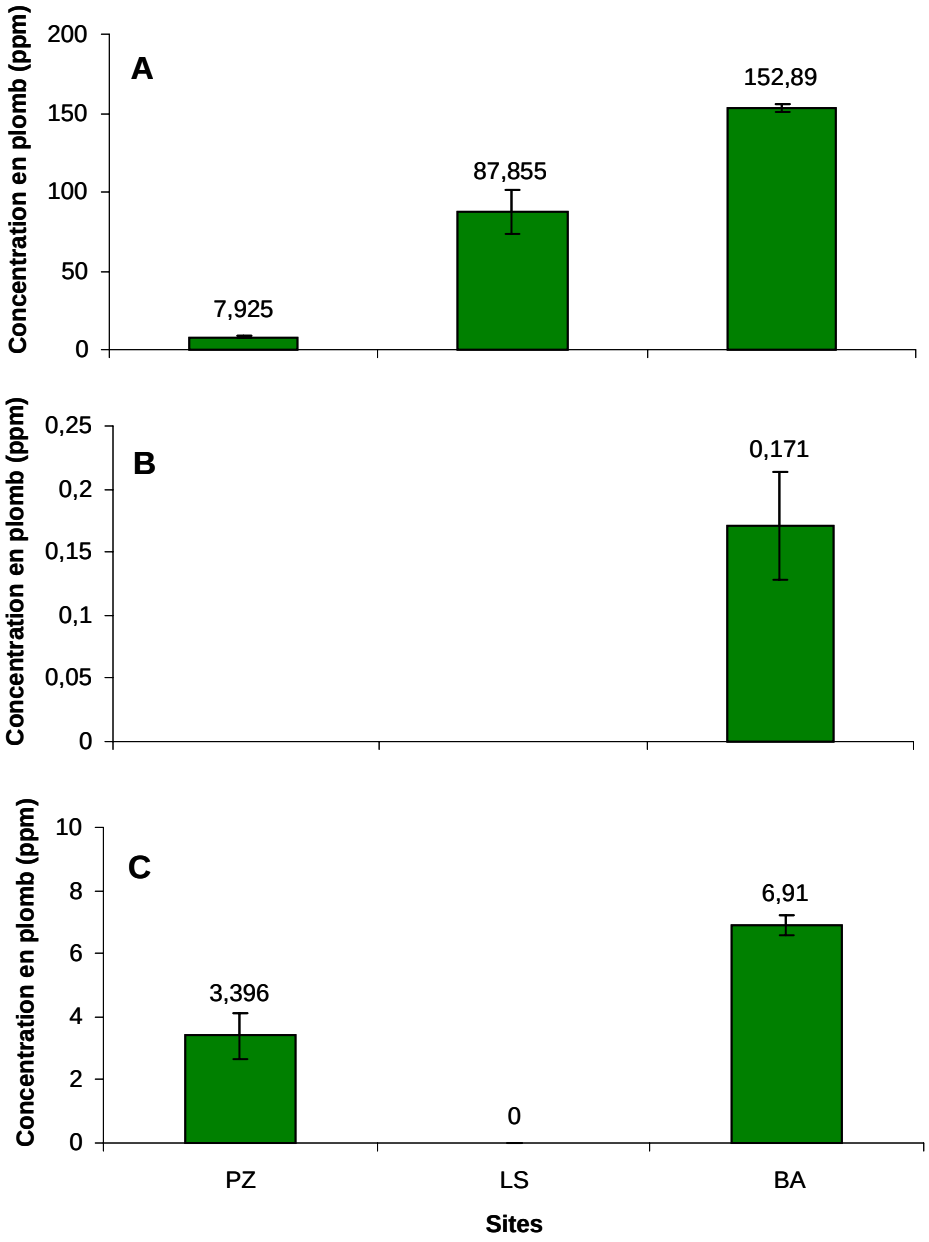


Figure 48: Dosage du plomb total (A), disponible (B) et biodisponible (C) présent dans les échantillons de sols (A et B) et d'aiguilles de pins (C) prélevés en novembre 2006 et juin 2008 sur les sites de Paal (PZ), Lommel (LS) et Balen (BA).

Na2EDTA 20 mM, NaCl 100 mM, SDS 2%, pH 9,0), 50 µL de β-mercaptoéthanol et 4 µL d'ARNt de levure à 10 mg.mL-1. Les quantités de SDS et de guanidine isothiocyanate ont été optimisées pour chacun des sols (cf Chapitre 1).

Le mélange est agité 5 min dans un broyeur Mikro Dismembrator U (B.Braun Biotech International) à 1600 agitations.min-1, puis centrifugé 5 min à 15000 rpm et à 4°C. Au surnageant est ajouté 1 mL de phénol acide (pH 5,0) / CHCl3 / alcool isoamylique (25/24/1 ; vol/vol/vol). Le mélange est agité 1 min au vortex à vitesse maximale, puis centrifugé 10 min à 15000 rpm et à 4°C. Cette extraction est réalisée 2 fois. Après transfert de la phase aqueuse dans un nouveau tube, 250 µL de CHCl3 / alcool isoamylique (24/1 ; vol/vol) sont ajoutés et le mélange est agité manuellement, puis centrifugé 10 min à 15000 rpm et à 4°C.

A la phase aqueuse sont ajoutés 0,1 vol d'acétate de sodium 3M (pH 5,2) et 2,5 vol d'éthanol absolu afin de précipiter les acides nucléiques. Après une incubation de 20 à 30 min à –70°C, le mélange est centrifugé 15 min à 15000 rpm et à 4°C. Le culot d'acides nucléiques est repris dans 40 µL d'eau «RNase-free» et 65 µL de LiCl 4M afin de précipiter sélectivement les ARN. Après une nuit à 4°C, les ARN sont précipités après une centrifugation de 15 min à 15000 rpm et les ADN contenus dans le surnageant sont précipités à l'éthanol et repris dans 100µL d'H_2O UP. Le culot d'ARN est traité à la DNase I (7 µL de DNase «RNase-free» (Fermentas) 1 $U.\mu L^{-1}$ + 2 µL de tampon + 11 µL d'H_2O «RNase-free») pendant 1h30 à 37°C. Les ARN sont précipités dans 2 vol d'isopropanol, centrifugés 8 min à 15000 rpm et lavés par 1 vol d'éthanol 70%. L'extrait brut d'ARN est séché sous vide puis repris dans 20 à 50 µL d'eau «RNase-free».

Une étape de purification supplémentaire sur micro-colonne contenant du Sephadex G-50 (ProbeQuant, Amersham) a été réalisée pour les extraits d'ARN et d'ADN issus du sol pollué de Balen. Cent µL d'extrait sont déposés sur la colonne. L'éluat obtenu est précipité par 0,1 vol d'acétate de sodium 3M (pH 5,2) et 2,5 vol d'éthanol absolu, lavé à l'éthanol 70%, séché et repris dans 20 à 50 µL d'eau «RNase-free».

Le kit d'extraction «RNA Power Soil» (Mobio) a également été utilisé pour une partie des échantillons de sol prélevés sur le site pollué.

Les ADN et les ARN obtenus après extraction sont quantifiés par mesure de leur absorbance à 260 nm au spectrophotomètre Nanodrop (Fig. 49). Une unité d'absorbance correspond à 40 µg d'ARN et à 50 µg d'ADN dans 1 ml (Sambrook *et al.*, 1989). La mesure du rapport d'absorbance $DO_{260\ nm}$ / $DO_{280\ nm}$ permet d'apprécier le degré de contamination par des protéines. Des rapports égaux à 1,8 pour les ADN et des rapports compris entre 1,8 et 2 pour les ARN indiquent l'absence de contaminations.

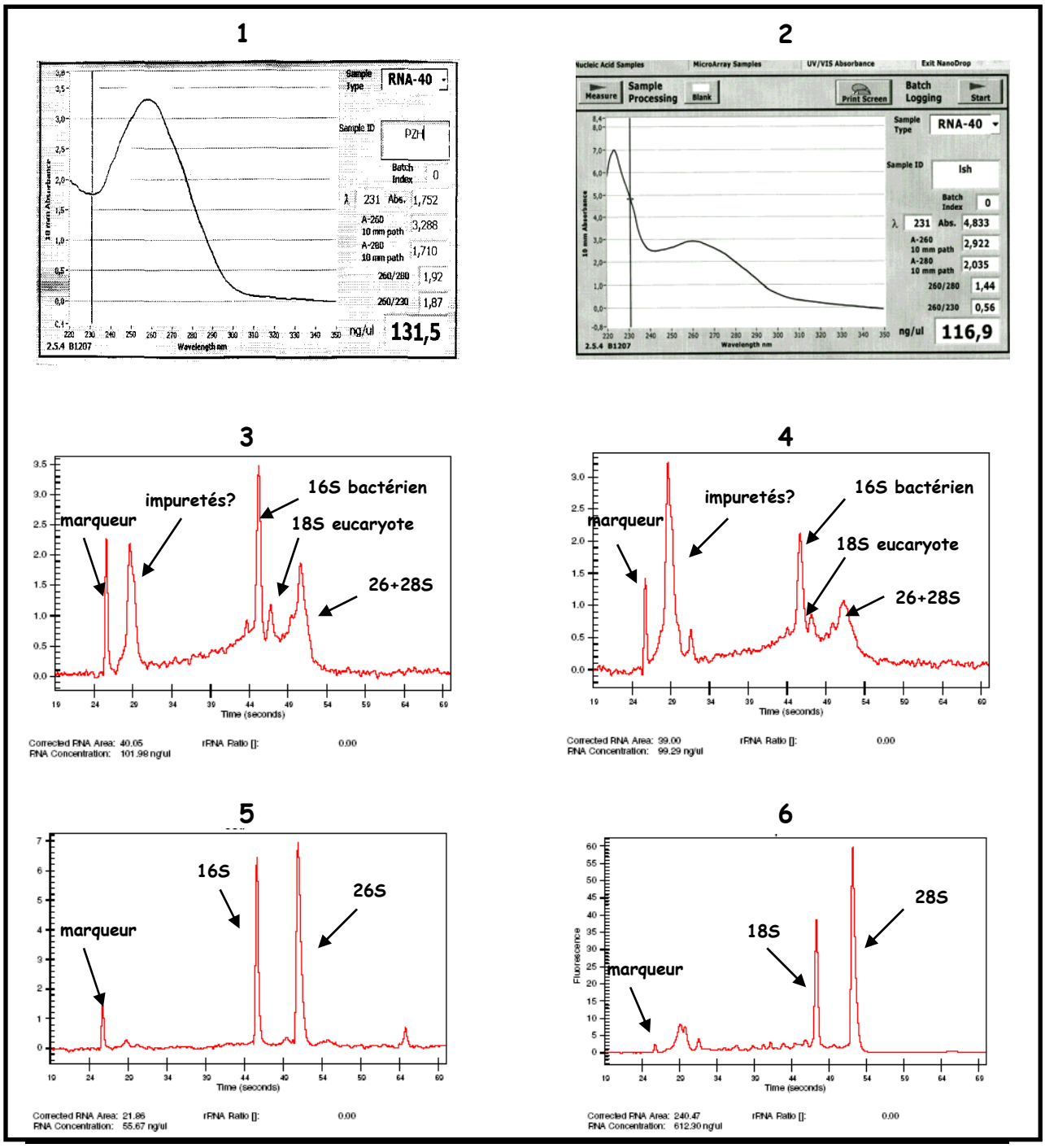


Figure 49: Spectres d'absorbance (1-2) et profils électrophorétiques (3-6) des solutions d'ARN extraits du sol de Paal (1 et 3) et du sol de Lommel (2 et 4). Les profils 5 et 6 correspondent respectivement aux ARN extraits de souches pures d'*E. coli* et du champignon *H. cylindrosporum*.

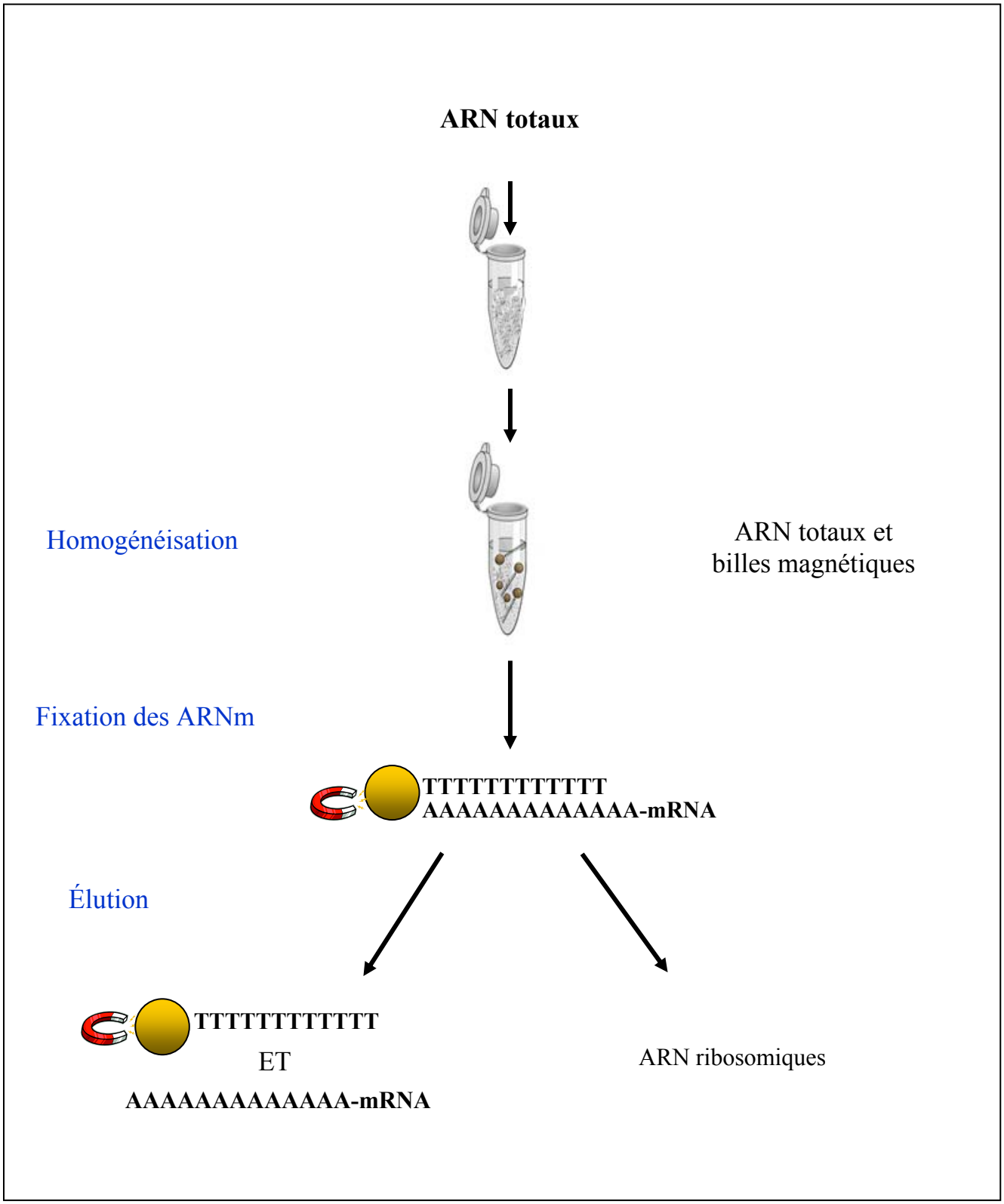


Figure 5: Principe de la méthode de purification des ARNm Eucaryotes à partir des ARN totaux par l'utilisation des billes magnétiques (Dynabeads Oligo (dT); Dynal). Les ARNm peuvent s'hybrider sur les billes magnétiques qui pourront être retenues par un aimant permettant ainsi la purification des ARNm après différents lavages.

La qualité et la quantité d'ARN ont également été appréciées après séparation par électrophorèse capillaire (Agilent RNA 6000 Nano Kit, Agilent Technologies) (Fig. 49). Les ARN sont marqués par un fluorophore (RNA Nano Dye Concentrate) et sont séparés suivant leur taille par électrophorèse au travers d'un réseau de capillaires interconnectés et remplis d'une matrice de polyacrylamide. Le profil obtenu permet notamment de déterminer un rapport entre ARNr 16S procaryotes et ARNr 18S eucaryotes ainsi que la concentration en ARN total.

3. Hybridation des ARNm eucaryotes polyadénylés sur billes magnétiques-dT_{25}

Le volume de l'extrait d'ARN total est réduit à 100 µL par évaporation sous vide (Speed-vac). Les ARN totaux sont mélangés à 100 µL d'une suspension de billes paramagnétiques, recouvertes d'oligo-dT25 dans une solution de concentration saline élevée selon le protocole préconisé par le fabriquant (Dynabeads mRNA Purification kit, Dynal Biotech) (Fig. 50). Après hybridation des ARNm aux oligo-dT , les billes sont attirées contre la paroi du tube par un aimant et les ARN non polyadénylés restés en solution sont prélevés et précipités par 2 vol d'isopropanol pendant 15 min à 4°C. Après une centrifugation de 15 min à 15000 rpm et à 4°C, le culot d'ARN non polyadénylés est lavé à l'éthanol 70% et repris dans 50 µL d'eau ultra pure «RNase-free». Les ARNm polyadénylés sont ensuite élués des billes par 20 µL d'eau ultra pure RNase-free à 65-80°C pendant 2 min.

4. Transcription inverse des ARNr

Cinq cent ng d'ARNr sont prélevés et mélangés soit à 2µL d'amorce eucaryote universelle Euk 516 ou SB (tableau 24) à 10 µM, soit à 1 µL d'hexanucléotides aléatoires à 10µM. Le mélange est complété à 11 µL avec de l'eau, puis chauffé à 70°C pendant 5 min et immédiatement placé dans la glace. En parallèle, un milieu réactionnel est préparé en mélangeant : 4 µL de tampon M-MuLV RT concentré 5X (Fermentas) ; 2 µL de dNTP à 2 mM ; 0,5 µL de Ribolock RNase Inhibitor à 40 U.µL-1 (Fermentas) ; 0,5 µL d'eau ultra pure «RNase free» et 2 µL de BSA à 5 mg.mL-1. Ce mélange est placé à 25°C pendant 5 min puis 1 µL d'enzyme RevertAid M-MuLV RT à 200 U.µL-1 est ajouté. Neuf µL de ce milieu réactionnel sont mélangés aux 11 µL d'échantillon d'ARNr et ce mélange est placé 10 min à

25°C, puis 1h30 à 42°C et enfin 10 min à 72°C pour arrêter la réaction. Les ADNc sont ensuite conservés à –20°C. Pour chaque extrait d'ARNr de chacun des sols, un minimum de 3 réactions indépendantes ont été réunies.

5. Amplification par PCR des ADNr 18S à partir de l'ADN de sol et des ARN de sol rétrotranscrits

Le couple d'amorces Euk516/Euk1A (Tableau 24) permet d'amplifier un fragment d'environ 500 pb situé en 5' de la séquence du gène 18S des ADNr de tous les eucaryotes (Fig. 51).

Le couple d'amorces S12.2/SB (Tableau 24) permet d'amplifier un fragment d'environ 700 pb situé en 3' de la séquence du gène 18S des ADNr de tous les eucaryotes (Fig. 51).

Le couple d'amorces 18F/18R (Tableau 24) permet d'amplifier un fragment d'environ 1800 pb soit la totalité du gène 18S des ADNr de tous les eucaryotes (Fig. 51).

Le milieu réactionnel (25 µL) est composé de : 14,9 µL d'eau ; 2,5 µL de tampon PCR concentré 10X (Tris-HCl 200 mM pH 8 ; KCl 500 mM) (Invitrogen) ; 0,75 µL de $MgCl_2$ à 25 mM; 1,25 µL de solution détergente W 1% (Invitrogen) ; 1,75 µL de dNTP (Fermentas) à 2 mM ; 0,5 µL d'un couple d'amorces à 10 µM ; 1,25 µL de BSA à 5 mg.mL-1 et 0,1 µL de Taq-DNA polymérase recombinante à 5 U.µL-1 (Invitrogen). Ce milieu réactionnel, additionné de 2 µL d'extrait brut ou purifié d'ADN métagénomique (10 à 20 ng) ou de 2 µL d'ARNr rétrotranscrits, est placé dans un thermocycleur PTC-200 et subit le programme suivant : une dénaturation initiale de 5min à 95°C suivie de 25 cycles d'amplification composés chacun d'une dénaturation de 45 sec à 95°C, d'une hybridation de 1 min à 56°C et d'une élongation de 45 sec à 72°C. Le programme s'achève par une élongation finale de 10 min à 72°C. La taille des amplifiats obtenus est vérifiée par électrophorèse sur gel d'agarose 1%. Pour chaque extrait d'ADNr ou d'ARNr rétrotranscrit de chacun des sols, un minimum de huit amplifications indépendantes ont été réunies. Les fragments d'intérêts sont découpés du gel et purifiés à l'aide du kit d'extraction sur gel Qiaquick (Qiagen).

6. Purification de fragments d'ADN sur gel d'agarose

L'extraction d'ADN de gels d'agarose a été réalisée à l'aide du kit QIAquick Gel extraction de Qiagen selon le protocole du fournisseur. La bande d'agarose contenant le

Séquences cibles	Amorces	Séquences nucléotidiques des amorces	Tm (°C)	Références
18S	Euk1A Euk516	5'-CTGGTTGATCCTGCCAG-3' 5'-ACCAGACTTGCCCTCC-3'	56	Diez *et al.*, 2001
18S	S12.2 SB	5'-GATYAGATACCGTCGTAGTC -3' 5'TGATCCTTCTGCAGGTTCACCTAC3'	56	Pawlowski, 2000
18S	18F 18R	5'-ACCTGGTTGATCCTGCCAG-3' 5'-TGATCCTTCYGCAGGTTCAC-3'	56	Moon-van der Staay *et al.*, 2001

Tableau 24: Séquences des amorces PCR utilisées pour l'amplification du gène 18S.

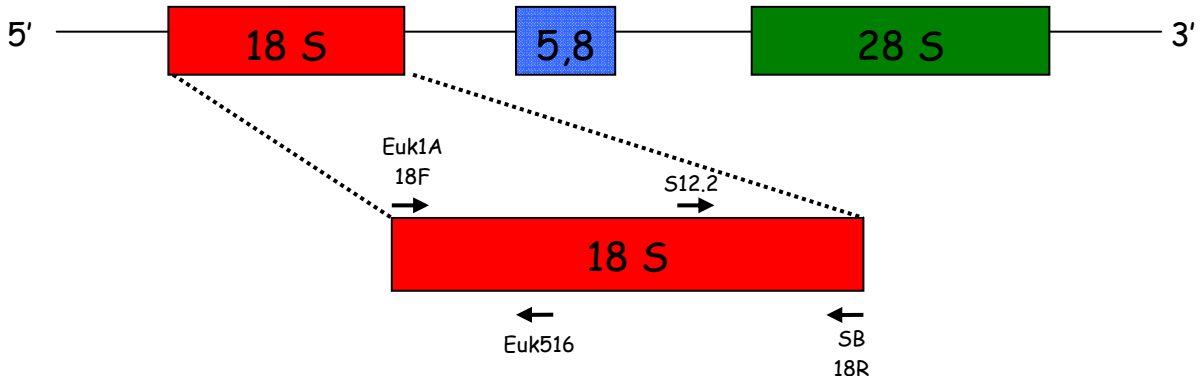


Figure 51: Représentation schématique de l'ADN ribosomique eucaryote et du positionnement des amorces utilisées pour l'amplification du gène 18S.

fragment d'ADN à extraire est récupérée dans un tube Eppendorf de 2ml. 3 volumes de tampon QG sont ajoutés au volume de la bande prélevée et le mélange est incubé 10 min à 50°C. L'échantillon est ensuite passé sur une colonne de purification selon le protocole du fournisseur. Après avoir lavé la colonne avec le tampon QG et PE, l'ADN est élué par 50 µl d'eau ultra-pure stérile. L'éluat est précipité par 0,1 volume de Na Acétate (3M pH 4,8) et 2,5 volumes d'éthanol absolu à -70°C pendant 30 min, puis est repris dans 15 µl d'eau ultra-pure stérile. L'ADN est quantifié par dosage spectrophotométrique à 260 nm au Nanodrop.

7. Banques d'ADN et d'ARN ribosomiques

7.1 Banque d'ADN ribosomiques

Les fragments de 500 et 700 pb issus des 8 répétitions indépendantes de PCR et purifiés ont été utilisés pour construire les deux banques d'ADN ribosomiques des Eucaryotes dans le vecteur plasmidique pCR4Blunt-TOPO (Invitrogen). La ligation a été réalisée en présence de 1 µl d'insert (30 ng/µl), 1 µl de solution saline, et 1 µl de vecteur (10 ng/µl) dans un volume réactionnel final de 6 µl. Le mélange réactionnel est homogénéisé et incubé 30 min à température ambiante. La durée d'incubation est choisie en fonction de la taille des produits PCR et de leur nombre. Un µL du mélange de ligation est ajouté à 25 µl de cellules électrocompétentes d'*E. coli* DH10B-T1^{R} (Invitrogen). Après un choc électrique de 2000 volts, 1 mL de milieu SOC liquide (Annexe) sont ajoutés dans les tubes préalablement refroidis dans la glace. Les cellules sont alors incubées 1 heure à 37°C. La totalité des cellules est étalée sur boites de milieu LB solide (Annexe) contenant de la kanamycine (50 µg/ml) et du X-Gal (80 mg/mL).

7.2 Banque d'ARN ribosomiques

Les fragments de 500 et 700 pb issus des 8 répétitions indépendantes de PCR et purifiés ont été utilisés pour construire les deux banques d'ARN ribosomique des Eucaryotes dans le vecteur plasmidique pCR4Blunt-TOPO comme décrit ci-dessus.

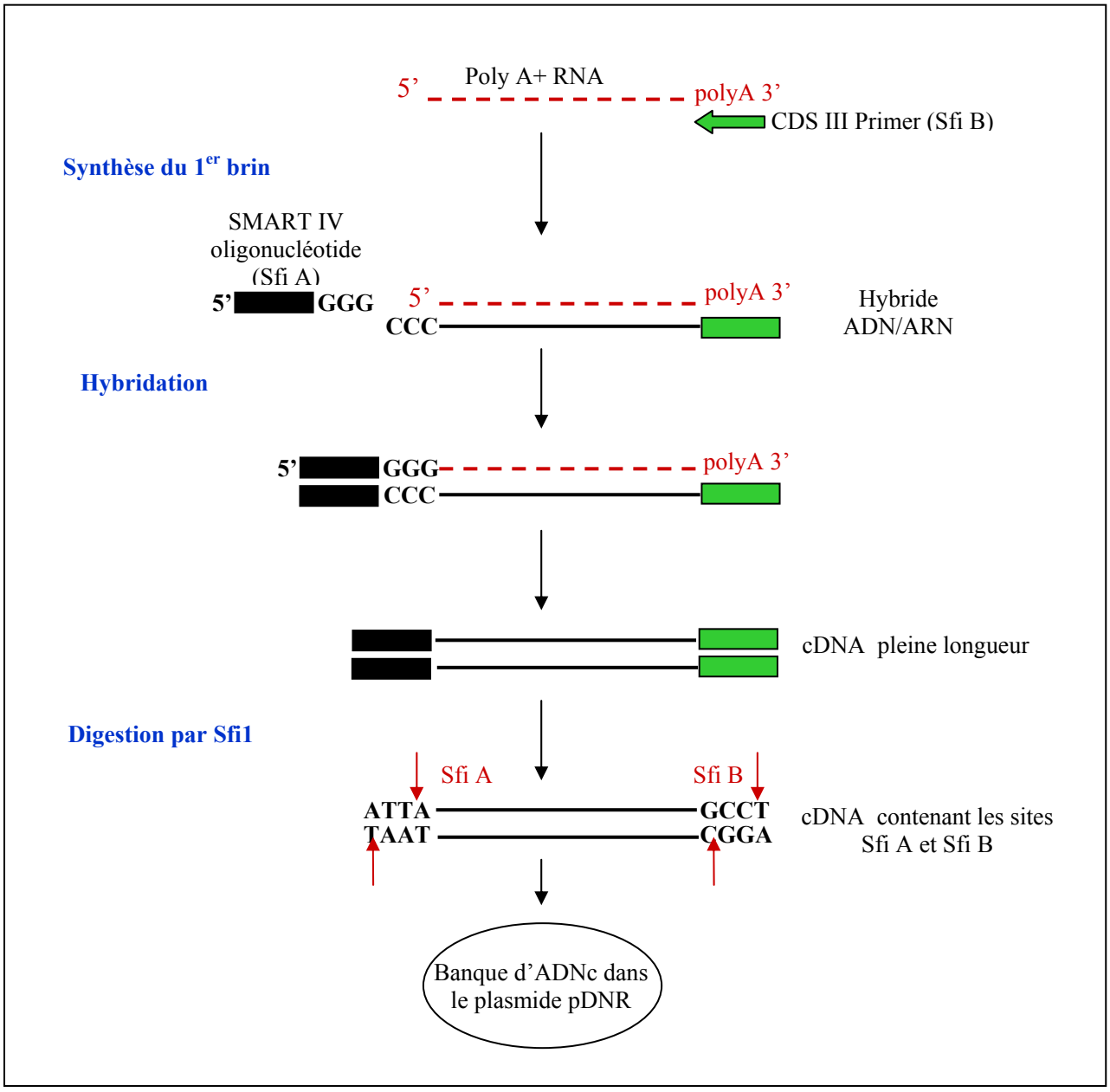


Figure 52: Principe de la technique d’amplification des ADNc pleine longueur avec le kit SMART (Clontech).
Cette technique permet d’obtenir des ADNc pleine longueur tout en partant d’une quantité infime de matériel. Les ARNm ont été purifiés sur billes magnétiques recouvertes d’Oligo(dT). Les ARNm ont été convertis en ADNc simple brin en utilisant la technologie SMART. Les ADNc sont ensuite amplifiés par LD-PCR (PCR longue distance). Les ADNc pleine longueur sont digérés par l’enzyme Sfi1. La création de sites asymétriques de part et d’autre de l’ADNc permet de réaliser un clonage directionnel dans un vecteur d’expression.

8. Banques d'ADN complémentaires

8.1 Obtention des ADNc

Les ADNc ont été synthétisés en utilisant le kit SMART (Clontech). Cette technologie, basée sur une étape de reverse transcription puis des étapes d'amplification par PCR, a été conçue afin d'obtenir des ADNc double brin pleine longueur tout en partant d'une quantité infime d'ARNm (Fig. 52). Environ 300 ng d'ARNm purifiés par affinité sur billes magnétiques recouvertes d'oligo (dT) repris dans 3 µl d'eau UP stérile sont mélangés à 1 µl d'oligonucléotides SMART IV et 1 µl d'amorce CDS III/ 3'. L'échantillon est dénaturé 2 min à 72°C et 2 min dans la glace. Sont ensuite ajoutés 2 µl de tampon approprié, 1 µl de DTT, 1 mM de chaque nucléotide (dATP, dCTP, dGTP, dTTP) et 1 µl de PowerScript Reverse Transcriptase. La synthèse de cDNA à lieu pendant 60 min à 42°C. La réaction est stoppée par le froid (5 min dans la glace).

2 µl de la réaction de reverse transcription ont été utilisés pour amplifier les ADN complémentaires par PCR longue distance (LD-PCR). L'amplification a été effectuée dans un volume final de 100 µl en présence de 10 µl du tampon approprié (10X Advantage 2 PCR buffer), 2 µl de "5' PCR Primer", 2 µl de "CDS III/ 3' PCR Primer", 200 µM de chaque nucléotide (dATP, dCTP, dGTP, dTTP), et une unité de DNA polymérase (50X Advantage 2 Polymerase Mix). Après une dénaturation initiale de 1 min à 95°C, l'amplification comprend 18 à 26 cycles du programme suivant : 15s de dénaturation à 95°C et 6 min d'élongation à 68°C. 50 µl des ADNc amplifiés sont digérés par 2 µl de Protéinase K (20 µg/µl) pendant 20 min à 45°C afin d'inactiver la DNA polymérase. Après deux extractions au phénol/SEVAG et une précipitation au Na acétate/ éthanol absolu, le culot d'ADNc est repris dans 79 µl d'eau UP stérile. Les ADNc sont ensuite digérés par 10 µl d'enzyme SfiI dans du tampon approprié en présence de BSA pendant 2 h à 50°C. Ensuite les ADN complémentaires sont fractionnés sur une colonne (CHROMA SPIN-400 ; Clontech). Seules les 4 fractions contenant des fragments supérieurs à 400 pb ont été collectées et regroupées. Après précipitation de ces 4 fractions par 0,1 volume de Na acétate (3 M ; pH 4,8) et 2,5 volumes d'éthanol absolu, 14 µl d'ADNc bordés chacune des sites Sfi1A & 1B ont été obtenus. La création de sites SfiI asymétriques de part et d'autre de l'ADNc permet de réaliser un clonage directionnel dans un vecteur d'expression de levure pFL61 dans lequel 2 sites de restriction Sfi IA et B ont été

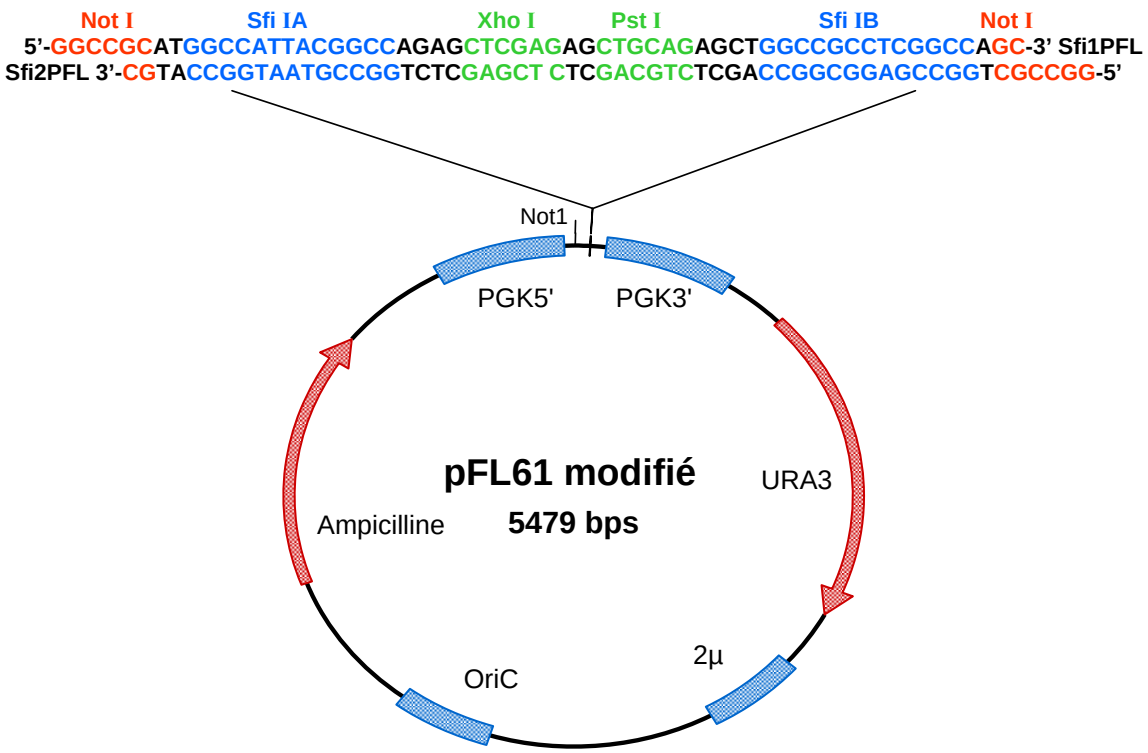


Figure 53: Carte de restriction du plasmide pFL61 modifié par insertion des séquences hybrides Sfi2PFL/Sfi1PFL dans un site Not1.
OriC, origine de réplication pour la multiplication dans *E. coli* ; 2µ, origine de réplication du plasmide 2µ de *S. cerevisiae* ; Ampicilline, marqueur de sélection pour le maintien du plasmide dans *E. coli* ; URA3, marqueur de sélection pour le maintien du plasmide dans *S. cerevisiae* ; PGK5', PGK3', séquences promotrice et terminatrice de la phosphoglycérokinase de *S. cerevisiae*.

Plasmides	Amorces	Séquences nucléotidiques des amorces	Tm(°C)
pFL61	PFLU PFLR	5'-CAGCTTCCAATTTCGTCACA-3' 5'-AAATACGCTGAACCCGAACA-3'	60
pDNR	M13-21 M13-Rev	5-TGTAAAACGACGGCCAGT-3' 5'-CAGGAAACAGCTATGACC-3'	55
pCR4Blunt-TOPO	M13-20 M13-Rev	5'-GTAAAACGACGGCCAG-3' 5'-CAGGAAACAGCTATGACC-3'	55

Tableau 25: Séquences des amorces PCR utilisées pour l'amplification des inserts contenu dans les plasmides pFL61, pDNR et pCR4Blunt-TOPO.

introduits entre le promoteur et le terminateur du gène codant pour la phosphoglycérate kinase (PGK) (Fig. 53).

8.2 Construction des banques d'ADNc

Les ligations des inserts (ADNc) aux vecteurs linéaires pDNR-LIB ou pFL61 (linéarisés par l'enzyme Sfi I) ont été réalisées à 15°C pendant 12 heures dans 0,5 µl de tampon de ligation ; 0,5 µl d'ATP et 1 U de T4-DNA ligase (Ozyme) dans un volume total de 5 µl. Pour chaque ligation, 100 ng de vecteur sont utilisés et la quantité d'insert est déterminée afin d'obtenir un rapport molaire insert/vecteur de 1:3.

Les banques ont été propagées dans E. coli DH10B"™T1R (de génotype : *F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 recA1 endA1 araD139 Δ(ara, leu)7697 galU galK - rpsL nupG tonA*).

Pour les banques d'ADNc construites dans pFL61, les bactéries ont été étalées sur milieu sélectif à raison de ca 65 c.f.u./cm^2. Après croissance à 37°C, les clones ont été récupérés dans du milieu LB par grattage et réunis afin d'extraire leurs plasmides par maxi-préparation (cf partie 10.2)

9. Méthodes moléculaires appliquées à *Escherichia coli*

9.1 Bactéries chimio-compétentes et transformation

La souche utilisée pour la transformation est la souche d'*E. coli* DH5α (Invitrogen) chimiquement compétente. L'ADN plasmidique est mélangé à 100 µl de suspension de bactéries compétentes. Après une incubation de 10 min à 4°C, les tubes sont placés 5 min à 37°C puis 1 ml de milieu LB liquide (Annexe) est ajouté. Ce mélange est alors incubé 1 heure à 37°C. Les transformants sont sélectionnés sur milieu LB contenant l'antibiotique approprié (Ampicilline à 100µg/mL ou Kanamycine à 50µg/mL) et du X-Gal à 40 µg.ml^{-1} si le vecteur contient le gène de la β-galactosidase.

9.2 Bactéries électrocompétentes et transformation

La souche utilisée pour la transformation par électroporation est la souche d'*E. coli* ElectroMAX DH10B ou DH10B"™T1R (Invitrogen). Pour transformer les bactéries, 100 à 200 ng d'ADN plasmidique contenus dans un volume de 1-2 µl sont ajoutés à 20 µl de cellules électrocompétentes. Les cuvettes d'électroporation stériles (1mm) sont refroidies sur la glace. Après mélange, la cuvette est placée dans un électroporateur (Eppendorf). Un courant de 2000 volts est alors appliqué et les cellules sont remises rapidement dans un tube froid stérile auquel est ajouté 1 ml de milieu SOC liquide (Annexe). Les cellules sont alors incubées 1 h à 37°C. Différentes quantités de cellules (de 50 à 200 µl) sont ensuite étalées sur un milieu sélectif approprié.

10. Extraction et purification d'ADN plasmidique bactérien

10.1 Purification rapide de plasmides :"Mini-préparations"

Les colonnes « NucleoSpin Plasmid » de Macherey-Nagel permettent d'obtenir rapidement de l'ordre de quelques µg d'ADN plasmidique super-enroulé sous une forme pure. Les bactéries contenant le plasmide sont cultivées 12 heures à 37°C dans 5 ml de milieu LB contenant l'antibiotique approprié. Elles sont ensuite centrifugées 10 min à 4°C à 10 000 rpm. Le culot est repris dans 250 µl de tampon A1 (Tris-HCl 50 mM pH 8 ; Na2EDTA 10 mM ; RNase A 100 µg.ml-1). 250 µl de tampon A2 (NaOH 200 mM ; SDS 1%) sont ajoutés et le mélange est laissé 5 min à température ambiante. Après l'addition de 350 µl de tampon A3 (KOAc 3M pH 5,5), la suspension est centrifugée 10 min à 15 000 rpm. Le surnageant est rapidement déposé sur colonne NucleoSpin. La colonne est lavée par 500 µl de tampon AW préchauffé à 50°C, puis par 600 µl de tampon A4. L'ADN est élué par 50 µl d'eau ultra-pure stérile puis quantifié.

10.2 Purification de plasmides :"Maxi préparations"

Les bactéries contenant les banques d'ADNc dans le plasmide pFL61 ont été grattées et réunies dans 250 mL de milieu LB. Après 1-2 h à 37°C, les cellules sont centrifugées 10 min à 4°C à 6000 rpm. Le culot est resuspendue dans 20 mL de Minilysat 1X (Glc 50mM, EDTA 10mM, Tris pH8 25mM). Après homogénéisation, 40 mL d'une solution de NaOH 0,2M et SDS 1% sont ajoutés et le mélange est remué doucement à l'aide d'une baguette en verre. 30 mL de K Acétate (3M pH 4,8) sont ajoutés et le mélange est laissé 10 à 30 min dans la glace. Après incubation, la suspension est centrifugée 25 min à 4°C à 10000 rpm et le surnageant est filtré à l'aide d'un entonnoir et d'un filtre papier. L'ADN est ensuite précipité avec 60 mL d'isopropanol pendant 10 à 30 min à température ambiante, puis centrifugé 30 min à 10000 rpm et repris dans 3 mL de TE pH 8.

Après une extraction au phénol/SEVAG, l'ADN est précipité par 0,1 volume de Na Acétate (3M pH 4,8) et 2,5 volumes d'éthanol absolu à -70°C pendant 30 min. Le culot d'ADN est rincé à l'éthanol à 70%, séché puis repris dans de l'eau ultra-pure stérile.

10.3 Modification du plasmide pFL61

Le plasmide pFL61 (Minet *et al.*, 1992) (plasmide se multipliant dans la levure et dans *E. coli*) (Fig. 53) a été modifié par incorporation de sites *Sfi*1 A et B dans son site de clonage au niveau du site *Not*1 en aval du promoteur PGK afin de permettre l'expression constitutive, dans la levure, des ADNc présentant des sites Sfi IA et B à leurs extrémités générées par le kit SMART cDNA Library Construction (Clontech). Le vecteur de clonage pFL61 est linéarisé par digestion enzymatique à l'aide de l'enzyme de restriction Not1. Après une extraction au phénol/SEVAG, l'ADN est précipité par 0,1 volume de Na Acétate (3M pH 4,8) et 2,5 volumes d'éthanol absolu à -70°C pendant 30 min. Le culot d'ADN est rincé à l'éthanol à 70%, séché puis repris dans de l'eau ultra-pure stérile.

Cent ng (0,028 pmole) du plasmide pFL61 digéré par l'enzyme de restriction Not I ont été ligués une nuit à 4°C à 2,8 pmoles des oligonucléotides SfipFL1 et SfipFL2 (Fig. 53) hybridés sur eux-mêmes (chauffés 1 min à 70°C puis refroidis à température ambiante) en présence de 1µL de ligase du phage T4 à 3 U.µL-1 (Fermentas) dans du tampon de ligation 10X.

La souche compétente E. coli DH5α a été transformée par 1 µL de mélange de ligation et étalée sur milieu gélosé LB + Amp (100 µg.mL-1). Une amplification par PCR de la zone de clonage a été réalisée avec les amorces PFLU et PFLR (Tableau 25) pour détecter la présence de la séquence hybride SfipFL1/SfipFL2 qui contient les sites Sfi IA et Sfi IB. L'insertion correcte des sites Sfi a été vérifiée par séquençage.

11. Analyse bioinformatique des séquences d'ADNr et d'ARNr 18S

Toutes les séquences ont été comparées aux données de la banque nr de Genebank (http://www.ncbi.nlm.nih.gov/) à l'aide du logiciel BLASTn. Seule les séquences présentant des homologies sur toute leur longueur avec des gènes d'ADNr 18S répertoriés ont été conservées. En parallèle, toutes les séquences ont été analysées par le logiciel Check Chimera (http: //rdp8.cme.msu.edu/cgis/chimera.cgi ?su=SSU) afin d'identifier les éventuelles séquences chimères artéfactuelles générées lors de la PCR. Des séquences d'organismes de références représentatifs de la diversité des grands règnes eucaryotes (champignons, protistes…) ont été ajoutées à la suite des séquences sélectionnées et cet ensemble de séquences à été aligné à l'aide du logiciel d'alignement CLUSTALW (http://www.ebi.ac.uk/). L'alignement ainsi généré a servi de base pour générer des arbres phylogénétiques à l'aide du logiciel PhyloWin en utilisant le modèle Neighbour Joining et les «p-distances». Les courbes de raréfaction ont été construites avec l'aide du logiciel Analytical rarefaction 1.3 de S. Holland (http://www.uga.edu/strata/software/). Les estimateurs SChao1 et SACE, tous deux basés sur l'abondance des taxons, ont été choisis pour estimer la richesse spécifique d'un sol à partir d'un échantillon de celui-ci. Ils ont été calculés grâce au logiciel développé par Kemp et Aller (http://www.also.org/lomethods/free/2004/0114a.html).

12. Analyses statistiques

Actuellement S_{Chao1} et S_{ACE} sont les deux estimateurs les plus couramment utilisés afin d'estimer la richesse des banques ribosomiques 16S et 18S. Ces méthodes estiment la richesse spécifique (nombre total de phylotypes) d'un échantillon environnemental de petite taille sans présupposer de modèles d'abondance particuliers. Elles sont basées sur des méthodes

statistiques de « Mark-Release-Recapture » (MRR) initialement utilisées pour estimer la taille des populations animales. Ces approches considèrent la proportion des phylotypes observés plus d'une fois (« recaptured ») par rapport à celle des phylotypes observées une seule fois (« singletons »). Ainsi, plus la diversité est faible, plus la probabilité de détecter plusieurs fois le même phylotype est élevée.

L'estimateur SChao1 (Chao, 1984) basé uniquement sur les phylotypes apparaissant 1 ou 2 fois se traduit par l'équation suivante :

$$S_{Chao1} = S_{obs} + \frac{F_1^2}{2(F_2+1)} - \frac{F_1 F_2}{2(F_2+1)^2}$$

On désignera par Sobs le nombre de phylotypes observées dans la banque ribosomique, et par F1 et F2 le nombre de phylotypes apparaissant une ou deux fois.
L'estimateur SACE se traduit par l'équation suivante :

$$S_{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{F_1}{C_{ACE}} \gamma_{ACE}^{\ 2}$$

On désignera par F1 le nombre de phylotypes apparaissant seulement une fois dans la banque ribosomique, par Srare le nombre de phylotypes apparaissant au maximum 10 fois et par Sabund le nombre de phylotypes apparaissant plus de 10 fois dans la banque ribosomique. γACE2 est un coefficient de variation de F1. ACE est un estimateur du « taux de couverture » de l'échantillon.

13. Levures

13.1 Souches utilisées

Les cellules de levures utilisées appartiennent à l'espèce *Saccharomyces cerevisiae* et sont issues de la souche BY4741 (Mat a ; his3Δ1 ; leu2Δ0 ; met15Δ0 ; ura3Δ0).

13.1.1 Souches sensibles au cadmium

Les souches sensibles au cadmium utilisées sont ΔYap1 (Mat a ; his3Δ1 ; leu2Δ0 ; met15Δ0 ; ura3Δ0 ; YML007wp::kanMX4) et ΔYcf1 (Mat a ; his3Δ1 ; leu2Δ0 ; met15Δ0 ; ura3Δ0; YDR135c::kanMX4).

Le gène *yap1* code pour un facteur de transcription de type bZIP (basic leucine zipper) impliqué dans la tolérance au stress oxydant. Il est activé par le peroxyde d'hydrogène (H2O2) qui induit plusieurs étapes de formation de ponts disulfures. Ce facteur transite alors du cytoplasme vers le noyau et active la transcription de nombreux gènes impliqués dans la résistance au stress oxydant tel que les thiorédoxines et les peroxyrédoxines.

Le gène *ycf1* code pour un transporteur vacuolaire du glutathion de type ABC (ATP-binding cassette). Il joue un rôle dans la détoxification de métaux tel que le cadmium en transportant le complexe glutathion-Cd du cytoplasme vers la vacuole.

13.1.2 Souche sensible au zinc

La souches sensible au zinc utilisée est ΔZrc1 (Mat a ; his3Δ1 ; leu2Δ0 ; met15Δ0 ; ura3Δ0 ; YMR243c::kanMX4).

Le gène *zrc1* code pour un transporteur tonoplastique du zinc. Il transporte le zinc du cytoplasme vers la vacuole où il est stocké.

13.2 Milieux de cultures

Les cellules de levures sont cultivées à 30°C. Elles sont entretenues sur un milieu complet YPG (Annexe 1).

La sélection des cellules de levures transformées avec le plasmide pFL61 modifié est effectuée sur un milieu minimum W0 (0,67% de « Yeast Nitrogen Base », 2% de glucose) complémenté par une solution d'acides aminés (Annexe 1) de l'adénine 1% et de la tyrosine 5%. Ce milieu est dépourvu d'uracile (marqueur de sélection).

Les tests de complémentation fonctionnelle des mutants de levures sensibles aux métaux sont réalisés sur un milieu minimum W0 complémenté en acides aminés et en adénine et supplémenté par du $CdCl_2$ (20µM pour ΔYap1, 40µM pour ΔYcf1) ou de $ZnSO_4$ (17,5 mM pour ΔZrc1). Il s'agit du milieu sélectif.

Pour confirmer que le phénotype de résistance résulte bien du gène porté sur le plasmide, ce dernier est éliminé en inoculant les transformants sur un milieu W0 complémenté en acides aminés, en adénine, en uracile et en acide 5-fluoro-orotique (FOA). L'acide 5-fluoro-orotique est un analogue structural d'un métabolite nécessaire à la synthèse de l'uracile. Son ajout dans le milieu de culture conduit à la synthèse du 5-fluoro-uracile qui est capable de s'incorporé dans l'ARN, d'induire une transcription erronée et donc de provoquer l'arrêt du cycle cellulaire. Les cellules vont donc utiliser l'uracile présent dans le milieu et rejeter leur plasmide.

13.3 Transformation de *Saccharomyces cerevisiae*.

13.3.1 Méthode TRAFO adaptée

Les cellules sont pré-cultivées dans 5 mL de milieu complet YPG à 28°C pendant deux à trois jours. Cinquante µL de cette préculture sont utilisés pour inoculer 50 mL de milieu YPG et sont placés pendant une nuit sous agitation 150 rpm à 28°C jusqu'à l'obtention d'environ 10^7 cellules/mL (DO 600 = 1-2). Après une centrifugation de 5 min à 3000 rpm, les cellules sont lavées par 25 mL d'eau ultrapure stérile, puis reprises dans 1mL d'eau et réparties dans 10 tubes eppendorf à raison de 100µL par tube. Les cellules sont centrifugés 30 sec à 6000 rpm, les surnageants sont éliminés et les culots sont repris dans 360µL d'un mélange contenant : 240µL de PEG 3350 à 50%, 36 µL de tampon lithium acétate (acétate de lithium pH 7,5 0,1 M), 100µg d'ADN de sperme de saumon et 500ng d'ADN plasmidique (banque ADNc). Après homogénéisation et incubation à 42°C pendant 40 min, les cellules sont centrifugées 30 sec à 6000 rpm et reprises dans 1 mL de YPG. Le mélange est placé à 30°C pendant 4h, puis centrifugé et repris dans 1mL d'eau ultrapure stérile. Les cellules sont étalées sur un milieu sélectif approprié.

13.3.2 Sélection des transformants résistants aux métaux

Les transformants sont sélectionnés après 3-4 jours de culture à 30°C sur un milieu sélectif approprié, puis re-isolés par stries d'épuisement sur le même milieu sélectif et cultivés pendant 3 jours à 30°C. Le plasmide est alors éliminé du transformant par culture sur milieu supplémenté en acide 5-fluoro-orotique. Après 3 jours de croissance à 30°C, les cellules sont

	dilution						dilution					
	1 DO_{600nm} = 1	**2** DO_{600nm} = 0.1	**3** DO_{600nm} = 0.01	**4** DO_{600nm} = 0.001	**5** DO_{600nm} = 0.0001	**6**	**7** DO_{600nm} = 1	**8** DO_{600nm} = 0.1	**9** DO_{600nm} = 0.01	**10** DO_{600nm} = 0.001	**11** DO_{600nm} = 0.0001	**12**
A	témoin BY4741						transformant G					
B	témoin ΔYAP ou ΔZRC …						transformant H					
C	transformant A						transformant I					
D	transformant B						transformant J					
E	transformant C						transformant K					
F	transformant D						transformant L					
G	transformant E						transformant M					
H	transformant F						transformant N					

Figure 54: Schéma d'une microplaque 96 puits dans laquelle sont réalisées les dilutions en cascade de chaque culture de « vrais » transformants de *S. cerevisiae*.

nombre de clones à extraire	**composition du mélange**	
	solution 1	**Zymolyase**
5	0,75 mL	9,75 µL
10	1,5 mL	19,5 µL
20	3 mL	39 µL
30	4,5 mL	58,5 µL
40	6 mL	78 µL
50	7,5 mL	97,5 µL

Tableau 26: Volumes de solution 1 et d'enzyme Zymolyase à ajouter lors de la purification de plasmide de levure avec le kit Zymoprep « Yeast Plasmid Extract1 » (Zymo Research).

placées sur un milieu W0 complémenté en acides aminés, en adénine, en uracile et supplémenté par le métal d'intérêt. A ce stade, les cellules qui ne repoussent pas sont considérées comme de « vrai » transformants.

13.3.3 Tests en gouttes

Une pré-culture de chaque « vrai » transformant est réalisée dans 5 mL de milieu minimum (supplémenté en uracile pour les souches sans plasmide) pendant 24h à 30°C. La DO à 600nm est ajustée à 1, puis 200µL de culture sont déposés dans une microplaque 96 puits stérile. Des dilutions au 1/10ème en cascades sont ensuite réalisées (Fig. 54).

5 µL de chaque dilution sont déposés sur un milieu sélectif gélosé (avec ou sans uracile) et sont incubés à 30°C pendant 3 à 5 jours. Deux répétitions sont effectuées pour chaque transformant.

13.4 Purification des plasmides de levures

L'ADN total est extrait des cellules de levures puis utilisé pour transformer des cellules électrocompétentes de *E. coli* afin de récupérer le plasmide contenu dans la levure.

Les cellules sont cultivées dans 5mL d'un milieu sélectif approprié pendant 48h à 28°C. Après une centrifugation de 5 min à 3000rpm, à température ambiante, le culot est lavé par 500µL d'eau stérile ultrapure, puis dans 200µL de tampon de lyse (2% Triton X100 ; 1% SDS ; 100mM NaCl ; 10 mM Tris-HCl pH 8 ; 1mM Na2EDTA). 200µL d'un mélange phénol/ chloroforme/ alcool isoamylique sont ajoutés et les cellules sont fortement agitées au vortex pendant 5 min en présence de 0,3g de billes de verre de diamètre 0,5mm. Les cellules sont à nouveau fortement agitées au vortex après l'ajout d'une solution de TE (10mM Tris-HCl pH 8 ; 1 mM Na2EDTA pH 8). Après une centrifugation de 3 min à 13000 rpm à température ambiante, les acides nucléiques sont précipités par 1 mL d'éthanol absolu. Le culot repris dans 400µL de TE pH 8 en présence de 3 µL de RnaseA (10mg/mL) est digéré pendant 5 min à 37°C. Après une re-précipitation avec 40µL de Na Acétate 3M pH 5,2 et 1 mL d'éthanol absolu, le culot est lavé, séché puis repris dans 20 µL de TE pH 8.

L'ADN total peut également être extrait avec le kit Zymoprep « Yeast Plasmid Extract1 » (Zymo Research). Les cellules sont cultivées dans 5mL d'un milieu sélectif

approprié pendant 48h à 28°C. Après centrifugation de 1 mL de culture à 11000 rpm pendant 2 min, le culot est repris dans 150 µL d'un mélange de solution 1 et d'enzyme zymolyase (Tableau 26).

Après incubation à 37°C pendant 15 à 60 min, 150 µL de solution 2 puis 150 µL de solution 3 sont ajoutées et homogénéisées. La suspension est centrifugée 2 min à 12000 rpm et le surnageant est transféré dans un nouveau tube. 400 µL d'isopropanol sont ajoutés et le mélange est centrifugé 8 min à 12000 rpm. Le culot est repris dans 35 µL d'eau UP stérile.

2 µL de cette préparation plasmidique sont utilisés pour transformer *E. coli*.

Bibliographie

Adl SM, Simpson AG, Farmer MA, *et al.*, 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol*. **52**, 399-451.

Adriano D, 1986. Trace elements in the terrestrial environment. *New York, USA: Springer Verlag*.

Aguilera A, Manrubia SC, Gomez F, *et al.*, 2006. Eukaryotic community distribution and its relationship to water physicochemical parameters in an extreme acidic environment, Rio Tinto (southwestern Spain). *Appl Environ Microbiol*. **72**, 5325-30.

Ahonen-Jonnarth U, Finlay R, Van-Hees P, *et al.*, 2000. Organic acids produced by mycorrhizal *Pinus sylvestris* exposed to elevated aluminium and heavy metal concentration. *New Phytologist*. **146**, 557-567.

Al-Hiyali S, McNeilly T et Bradshaw A, 1990. The effect of zinc contamination from electricity pylons. Contrasting patterns of evolution in five grass species. *New Phytologist*. **114**, 183-190.

Al-Hiyali S, McNeilly T, Bradshaw A, *et al.*, 1993. The effect of zinc contamination from electricity pylons. Genetic constraints on selection for zinc tolerance. *Heredity*. **70**, 22-32.

Altschul SF, Madden TL, Schaffer AA, *et al.*, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**, 3389-402.

Amann RI, Ludwig W et Schleifer KH, 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev*. **59**, 143-69.

Amaral Zettler LA, Gomez F, Zettler E, *et al.*, 2002. Microbiology: eukaryotic diversity in Spain's River of Fire. *Nature*. **417**, 137.

Andersson JO, 2005. Lateral gene transfer in eukaryotes. *Cell Mol Life Sci*. **62**, 1182-97.

Antonovics J, Bradshaw A et Turner R, 1971. Heavy metal tolerance in plants. *Advances in Ecological Research*. **7**, 1-85.

Archibald JM, 2005. Jumping genes and shrinking genomes-probing the evolution of eukaryotic photosynthesis with genomics. *IUBMB Life*. **57**, 539-47.

Arisue N, Hasegawa M et Hashimoto T, 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. *Mol Biol Evol*. **22**, 409-20.

Ashelford KE, Chuzhanova NA, Fry JC, *et al.*, 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Appl Environ Microbiol*. **72**, 5734-41.

Atkins P et Jones L, 1997. Chemestry - molecules, matter and change. *New York, USA: W.H. Freeman*.

Audia JP, Patton MC et Winkler HH, 2008. DNA microarray analysis of the heat shock transcriptome of the obligate intracytoplasmic pathogen *Rickettsia prowazekii*. *Appl Environ Microbiol*. **74**, 7809-12.

Bailly J, Fraissinet-Tachet L, Verner MC, *et al.*, 2007. Soil eukaryotic functional diversity, a metatranscriptomic approach. *Isme J*. **1**, 632-42.

Baker BJ, Hugenholtz P, Dawson SC, *et al.*, 2003. Extremely acidophilic protists from acid mine drainage host Rickettsiales-lineage endosymbionts that have intervening sequences in their 16S rRNA genes. *Appl Environ Microbiol*. **69**, 5512-8.

Baldauf SL, Roger AJ, Wenk-Siefert I, *et al.*, 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*. **290**, 972-7.

Bapteste E, Brinkmann H, Lee JA, *et al.*, 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci U S A*. **99**, 1414-9.

Barajas-Aceves M, Hassan M, Tinoco R, *et al.*, 2002. Effect of pollutants on the ergosterol content as indicator of fungal biomass. *J Microbiol Methods*. **50**, 227-36.

Bass D et Cavalier-Smith T, 2004. Phylum-specific environmental DNA analysis reveals remarkably high global biodiversity of Cercozoa (Protozoa). *Int J Syst Evol Microbiol*. **54**, 2393-404.

Beeby A, 2001. What do sentinels stand for? *Environ Pollut*. **112**, 285-98.

Béjà O, Spudich EN, Spudich JL, *et al.*, 2001. Proteorhodopsin phototrophy in the ocean. *Nature*. **411**, 786-9.

Bellion M, Courbot M, Jacob C, *et al.*, 2006. Extracellular and cellular mechanisms sustaining metal tolerance in ectomycorrhizal fungi. *FEMS Microbiol Lett*. **254**, 173-81.

Blaudez D, Botton B et Chalot M, 2000a. Cadmium uptake and subcellular compartmentation in the ectomycorrhizal fungus Paxillus involutus. *Microbiology*. **146 (Pt 5)**, 1109-17.

Blaudez D, Jacob C, Turnau K, *et al.*, 2000b. Differential responses of ectomycorrhizal fungal isolates to heavy metals *in vitro*. *Mycol. Res*. **104**, 1366-1371.

Boldrin F, Santovito G, Irato P, *et al.*, 2002. Metal interaction and regulation of *Tetrahymena pigmentosa* metallothionein genes. *Protist*. **153**, 283-91.

Bonkowski M, 2002. Protozoa and plant growth: trophic links and mutualism. *Eur J Protistol* **37**, 363-365.

Borrelly GP, Harrison MD, Robinson AK, *et al.*, 2002. Surplus zinc is handled by Zym1 metallothionein and Zhf endoplasmic reticulum transporter in *Schizosaccharomyces pombe*. *J Biol Chem*. **277**, 30394-400.

Bowers N, Pratt J, Beeson D, *et al.*, 1997. Comparative evaluation of soil toxicity using lettuce seeds and soil ciliates. *Environ Toxicol Chem*. **16**, 207-213.

Bradshaw A et McNeilly T, 1981. Evolution and pollution. *Studies in biology*. **146**, 1-76.

Brinkmann H, van der Giezen M, Zhou Y, *et al.*, 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol*. **54**, 743-57.

Bui ET, Bradley PJ et Johnson PJ, 1996. A common evolutionary origin for mitochondria and hydrogenosomes. *Proc Natl Acad Sci U S A*. **93**, 9651-6.

Bult CJ, White O, Olsen GJ, *et al.*, 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*. **273**, 1058-73.

Burki F, Shalchian-Tabrizi K, Minge M, *et al.*, 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One*. **2**, e790.

Campbell C, Warren A, Cameron C, *et al.*, 1997. Direct toxicity assessment of two soils amended with sewage sludge contaminated with heavy metals using a protozoan (*Colpoda steinii*) bioassay. *Chemosphere*. **34**, 501-514.

Cavalier-Smith T, 1993. Kingdom protozoa and its 18 phyla. *Microbiol Rev*. **57**, 953-94.

Cavalier-Smith T, 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol*. **52**, 297-354.

Cavalier-Smith T et Chao EE, 1996. Molecular phylogeny of the free-living archezoan *Trepomonas agilis* and the nature of the first eukaryote. *J Mol Evol*. **43**, 551-62.

Cavalier-Smith T et Chao EE, 2003. Phylogeny and classification of phylum Cercozoa (Protozoa). *Protist*. **154**, 341-58.

Cavalier-Smith T et von der Heyden S, 2007. Molecular phylogeny, scale evolution and taxonomy of centrohelid heliozoa. *Mol Phylogenet Evol*. **44**, 1186-203.

Chapman G et Dunlop S, 1981. Detoxication of zinc and cadmium by the freshwater protozoan *Tetrahymena pyriformis*. I. The effect of water hardness. *Environ Res*. **26**, 81-6.

Choe CP, Hwang UW et Kim W, 1999. Putative secondary structures of unusually long strepsipteran SSU rRNAs and its phylogenetic implications. *Mol Cells*. **9**, 191-9.
Chuang HW, Wang IW, Lin SY*, et al.*, 2009. Transcriptome analysis of cadmium response in *Ganoderma lucidum*. *FEMS Microbiol Lett*. **293**, 205-13.
Cobbett C et Goldsbrough P, 2002. Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. *Annu Rev Plant Biol*. **53**, 159-82.
Collins L et Penny D, 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol*. **22**, 1053-66.
Colpaert J et Van-Assche J, 1992. The effects of cadmium and the cadmium-zinc interaction on the axenic growth of ectomycorrhizal fungi. *Plant Soil*. **145**, 237-243.
Colpaert J, Vandenkoornhuyse P, Adriaensen K*, et al.*, 2000. Genetic variation and heavy metal tolerance in the ectomycorrhizal basidiomycete *Suillus luteus*. *New Phytologist*. **147**, 367-379.
Conklin DS, Culbertson MR et Kung C, 1994. Interactions between gene products involved in divalent cation transport in *Saccharomyces cerevisiae*. *Mol Gen Genet*. **244**, 303-11.
Conklin DS, McMaster JA, Culbertson MR*, et al.*, 1992. COT1, a gene involved in cobalt accumulation in *Saccharomyces cerevisiae*. *Mol Cell Biol*. **12**, 3678-88.
Courbot M, Diez L, Ruotolo R*, et al.*, 2004. Cadmium-responsive thiols in the ectomycorrhizal fungus *Paxillus involutus*. *Appl Environ Microbiol*. **70**, 7413-7.
Courtois S, Cappellano CM, Ball M*, et al.*, 2003. Recombinant environmental libraries provide access to microbial diversity for drug discovery from natural products. *Appl Environ Microbiol*. **69**, 49-55.
Courtois S, Frostegard A, Goransson P*, et al.*, 2001. Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. *Environ Microbiol*. **3**, 431-9.
Coyle P, Philcox JC, Carey LC*, et al.*, 2002. Metallothionein: the multipurpose protein. *Cell Mol Life Sci*. **59**, 627-47.
Cruz-Vasquez BH, Diaz-Cruz JM, Arino C*, et al.*, 2002. Study of Cd^{2+} complexation by the glutathione fragments Cys-Gly (CG) and gamma-Glu-Cys (gamma-EC) by differential pulse polarography. *Analyst*. **127**, 401-6.
Culotta VC, Howard WR et Liu XF, 1994. CRS5 encodes a metallothionein-like protein in *Saccharomyces cerevisiae*. *J Biol Chem*. **269**, 25295-302.
Cumming J, Swiger T, Kurnik B*, et al.*, 2001. Organic acid exudation by *Laccaria bicolor* and *Pisolithus tinctorius* exposed to aluminium *in vitro*. *Can J For Res*. **31**, 703-710.
Dacks JB, Marinets A, Ford Doolittle W*, et al.*, 2002. Analyses of RNA Polymerase II genes from free-living protists: phylogeny, long branch attraction, and the eukaryotic big bang. *Mol Biol Evol*. **19**, 830-40.
Daniel R, 2004. The soil metagenome-a rich resource for the discovery of novel natural products. *Curr Opin Biotechnol*. **15**, 199-204.
Daniel R, 2005. The metagenomics of soil. *Nat Rev Microbiol*. **3**, 470-8.
Dawson SC et Pace NR, 2002. Novel kingdom-level eukaryotic diversity in anoxic environments. *Proc Natl Acad Sci U S A*. **99**, 8324-9.
Deckert J, 2005. Cadmium toxicity in plants: is there any analogy to its carcinogenic effect in mammalian cells? *Biometals*. **18**, 475-81.
Demaneche S, David MM, Navarro E*, et al.*, 2009. Evaluation of functional gene enrichment in a soil metagenomic clone library. *J Microbiol Methods*. **76**, 105-7.
Diaz S, Amaro F, Rico D*, et al.*, 2007. *Tetrahymena* metallothioneins fall into two discrete subfamilies. *PLoS One*. **2**, e291.
Diaz S, Martin-Gonzalez A et Carlos Gutierrez J, 2006. Evaluation of heavy metal acute toxicity and bioaccumulation in soil ciliated protozoa. *Environ Int*. **32**, 711-7.

Diez B, Pedros-Alio C et Massana R, 2001. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl Environ Microbiol*. **67**, 2932-41.

Dinsdale EA, Edwards RA, Hall D, *et al.*, 2008. Functional metagenomic profiling of nine biomes. *Nature*. **452**, 629-32.

Dmuchowski W et Bytnerowicz A, 1995. Monitoring environmental pollution in Poland by chemical analysis of Scots pine (*Pinus sylvestris L.*) needles. *Environ Pollut*. **87**, 87-104.

Dondero F, Cavaletto M, Ghezzi AR, *et al.*, 2004. Biochemical characterization and quantitative gene expression analysis of the multi-stress inducible metallothionein from *Tetrahymena thermophila*. *Protist*. **155**, 157-68.

Doolittle WF, 1999. Phylogenetic classification and the universal tree. *Science*. **284**, 2124-9.

Duffus J, 2002. "Heavy metals" - A meaningless term? *Pure and Applied Chemestry*. **74**, 793-807.

Dumont MG, Radajewski SM, Miguez CB, *et al.*, 2006. Identification of a complete methane monooxygenase operon from soil by combining stable isotope probing and metagenomic analysis. *Environ Microbiol*. **8**, 1240-50.

Eckburg PB, Bik EM, Bernstein CN, *et al.*, 2005. Diversity of the human intestinal microbial flora. *Science*. **308**, 1635-8.

Edgcomb VP, Roger AJ, Simpson AG, *et al.*, 2001. Evolutionary relationships among "jakobid" flagellates as indicated by alpha- and beta-tubulin phylogenies. *Mol Biol Evol*. **18**, 514-22.

Eide DJ, Bridgham JT, Zhao Z, *et al.*, 1993. The vacuolar H(+)-ATPase of *Saccharomyces cerevisiae* is required for efficient copper detoxification, mitochondrial function, and iron metabolism. *Mol Gen Genet*. **241**, 447-56.

Ekelund F, Frederiksen HB et Ronn R, 2002. Population dynamics of active and total ciliate populations in arable soil amended with wheat. *Appl Environ Microbiol*. **68**, 1096-101.

Embley TM et Martin W, 2006. Eukaryotic evolution, changes and challenges. *Nature*. **440**, 623-30.

Eschenfeldt WH, Stols L, Rosenbaum H, *et al.*, 2001. DNA from uncultured organisms as a source of 2,5-diketo-D-gluconic acid reductases. *Appl Environ Microbiol*. **67**, 4206-14.

Fast NM, Kissinger JC, Roos DS, *et al.*, 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. *Mol Biol Evol*. **18**, 418-26.

Fauchon M, Lagniel G, Aude JC, *et al.*, 2002. Sulfur sparing in the yeast proteome in response to sulfur demand. *Mol Cell*. **9**, 713-23.

Favier A, 2003. La chimie dans les sciences médicales. *L'actualité chimique*. **269-270**, 103-108.

Felsenstein J, 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*. **39**, 783-791.

Fergusson J, 1990. The heavy elements. Chemestry, environmental impact and health effects. *Oxford, UK: Pergamon Press*.

Fierer N, Breitbart M, Nulton J, *et al.*, 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol*. **73**, 7059-66.

Floyd R, Abebe E, Papert A, *et al.*, 2002. Molecular barcodes for soil nematode identification. *Mol Ecol*. **11**, 839-50.

Fomina M, Hillier S, Charnock JM, *et al.*, 2005. Role of oxalic acid overexcretion in transformations of toxic metal minerals by *Beauveria caledonica*. *Appl Environ Microbiol*. **71**, 371-81.

Forbes V, 1998. Genetics of ecotoxicology. *Ann Arbor: Taylor and Francis*.

Foster R, 1988. Microenvironments of soil microorganisms. *Biol. Fertil. Soils*. **6**, 189-203.

Frank DN et Pace NR, 2008. Gastrointestinal microbiology enters the metagenomics era. *Curr Opin Gastroenterol*. **24**, 4-10.

Frias-Lopez J, Shi Y, Tyson GW, *et al.*, 2008. Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A*. **105**, 3805-10.

Gabor E, deVries E et Janssen D, 2003. Efficient recovery of environmental DNA for expression cloning by indirect methods. *FEMS Microbiol. Ecol.* **44**, 153-163.

Gabor EM, de Vries EJ et Janssen DB, 2004. Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases. *Environ Microbiol*. **6**, 948-58.

Gallego A, Martin-Gonzalez A, Ortega R, *et al.*, 2007. Flow cytometry assessment of cytotoxicity and reactive oxygen species generation by single and binary mixtures of cadmium, zinc and copper on populations of the ciliated protozoan *Tetrahymena thermophila*. *Chemosphere*. **68**, 647-61.

Galtier N, Gouy M et Gautier C, 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci*. **12**, 543-8.

Garcia JJ, Martinez-Ballarin E, Millan-Plano S, *et al.*, 2005. Effects of trace elements on membrane fluidity. *J Trace Elem Med Biol*. **19**, 19-22.

Gast C, Jansen E, Bierling J, *et al.*, 1988. Heavy metals in mushrooms and their relationships with soil characteristics. *Chemosphere*. **17**, 789-799.

Gilbert JA, Field D, Huang Y, *et al.*, 2008. Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS One*. **3**, e3042.

Gillespie DE, Brady SF, Bettermann AD, *et al.*, 2002. Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol*. **68**, 4301-6.

Gonzalez-Chavez MC, Carrillo-Gonzalez R, Wright SF, *et al.*, 2004. The role of glomalin, a protein produced by arbuscular mycorrhizal fungi, in sequestering potentially toxic elements. *Environ Pollut*. **130**, 317-23.

Grant S, Grant WD, Cowan DA, *et al.*, 2006. Identification of eukaryotic open reading frames in metagenomic cDNA libraries made from environmental samples. *Appl Environ Microbiol*. **72**, 135-43.

Guillou L, Viprey M, Chambouvet A, *et al.*, 2008. Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol*. **10**, 3349-65.

Gupta R, Beg QK et Lorenz P, 2002. Bacterial alkaline proteases: molecular approaches and industrial applications. *Appl Microbiol Biotechnol*. **59**, 15-32.

Habura A, Pawlowski J, Hanes SD, *et al.*, 2004a. Unexpected foraminiferal diversity revealed by small-subunit rDNA analysis of Antarctic sediment. *J Eukaryot Microbiol*. **51**, 173-9.

Habura A, Pawlowski J, Hanes SD, *et al.*, 2004b. Unexpected foraminiferal diversity revealed by small-subunit rDNA analysis of Antarctic sediment. *J Eukaryot Microbiol*. **51**, 173-9.

Hackett JD, Scheetz TE, Yoon HS, *et al.*, 2005. Insights into a dinoflagellate genome through expressed sequence tag analysis. *BMC Genomics*. **6**, 80.

Hackett JD, Yoon HS, Li S, *et al.*, 2007. Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. *Mol Biol Evol*. **24**, 1702-13.

Hall JL, 2002. Cellular mechanisms for heavy metal detoxification and tolerance. *J Exp Bot*. **53**, 1-11.

Hampl V, Hug L, Leigh JW, *et al.*, 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proc Natl Acad Sci U S A*. **106**, 3859-64.

Handelsman J, Rondon MR, Brady SF, *et al.*, 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*. **5**, R245-9.

Hashimoto T, Nakamura Y, Kamaishi T, *et al.*, 1995. Phylogenetic place of mitochondrion-lacking protozoan, Giardia lamblia, inferred from amino acid sequences of elongation factor 2. *Mol Biol Evol*. **12**, 782-93.

Henne A, Daniel R, Schmitz RA, *et al.*, 1999. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl Environ Microbiol*. **65**, 3901-7.

Henne A, Schmitz RA, Bomeke M, *et al.*, 2000. Screening of environmental DNA libraries for the presence of genes conferring lipolytic activity on *Escherichia coli*. *Appl Environ Microbiol*. **66**, 3113-6.

Hirt RP, Logsdon JM, Jr., Healy B, *et al.*, 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci U S A*. **96**, 580-5.

Högberg M et Högberg P, 2002. Extramatrical ectomycorrhizal mycelium contributes one third of microbial biomass and produces, together with associated roots, half of the dissolved organic carbon in a forest soil. *New Physiologist*. **154**, 791-795.

Holben WE, Jansson JK, Chelm BK, *et al.*, 1988. DNA Probe Method for the Detection of Specific Microorganisms in the Soil Bacterial Community. *Appl Environ Microbiol*. **54**, 703-711.

Hugenholtz P et Tyson GW, 2008. Microbiology: metagenomics. *Nature*. **455**, 481-3.

Hurt RA, Qiu X, Wu L, *et al.*, 2001. Simultaneous recovery of RNA and DNA from soils and sediments. *Appl Environ Microbiol*. **67**, 4495-503.

Jacob C, Courbot M, Brun A, *et al.*, 2001. Molecular cloning, characterization and regulation by cadmium of a superoxide dismutase from the ectomycorrhizal fungus *Paxillus involutus*. *Eur J Biochem*. **268**, 3223-32.

Jacob C, Courbot M, Martin F, *et al.*, 2004. Transcriptomic responses to cadmium in the ectomycorrhizal fungus *Paxillus involutus*. *FEBS Lett*. **576**, 423-7.

James TY, Kauff F, Schoch CL, *et al.*, 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*. **443**, 818-22.

Jékely G, 2005. Glimpsing over the event horizon: evolution of nuclear pores and envelope. *Cell Cycle*. **4**, 297-299.

Jeon S, Bunge J, Leslin C, *et al.*, 2008. Environmental rRNA inventories miss over half of protistan diversity. *BMC Microbiol*. **8**, 222.

Johnson MD, Tengs T, Oldach DW, *et al.*, 2004. Highly divergent SSU rRNA genes found in the marine ciliates *Myrionecta rubra* and *Mesodinium pulex*. *Protist*. **155**, 347-59.

Juste C, 1988. Appréciation de la mobilité et de la biodisponibilité des éléments en traces du sol. *Sci Sol*. **26**, 103-112.

Kamizono A, Nishizawa M, Teranishi Y, *et al.*, 1989. Identification of a gene conferring resistance to zinc and cadmium ions in the yeast *Saccharomyces cerevisiae*. *Mol Gen Genet*. **219**, 161-7.

Kempf P et Aller J, 2004. Estimating prokaryotic diversity: when are 16S rDNA libraries large enough? *Limnology and Oceanography: Methods*. **2**, 114-125.

Kim E, Simpson AG et Graham LE, 2006. Evolutionary relationships of apusomonads inferred from taxon-rich analyses of 6 nuclear encoded genes. *Mol Biol Evol*. **23**, 2455-66.

Kim SJ, Lee CM, Han BR, *et al.*, 2008. Characterization of a gene encoding cellulase from uncultured soil bacteria. *FEMS Microbiol Lett*. **282**, 44-51.

King N, Westbrook MJ, Young SL, *et al.*, 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*. **451**, 783-8.

Knietsch A, Bowien S, Whited G, *et al.*, 2003a. Identification and characterization of coenzyme B12-dependent glycerol dehydratase- and diol dehydratase-encoding genes from metagenomic DNA libraries derived from enrichment cultures. *Appl Environ Microbiol*. **69**, 3048-60.

Knietsch A, Waschkowitz T, Bowien S, *et al.*, 2003b. Metagenomes of complex microbial consortia derived from different soils as sources for novel genes conferring formation of carbonyls from short-chain polyols on *Escherichia coli*. *J Mol Microbiol Biotechnol*. **5**, 46-56.

Kozdroj J, Piotrowska-Seget Z et Krupa P, 2007. Mycorrhizal fungi and ectomycorrhiza associated bacteria isolated from an industrial desert soil protect pine seedlings against Cd(II) impact. *Ecotoxicology*. **16**, 449-56.

Kuhn DA, Gregory DA, Buchanan GE, Jr., *et al.*, 1978. Isolation, characterization, and numerical taxonomy of Simonsiella strains from the oral cavities of cats, dogs, sheep, and humans. *Arch Microbiol*. **118**, 235-41.

Kunkel TA et Loeb LA, 1979. On the fidelity of DNA replication. Effect of divalent metal ion activators and deoxyrionucleoside triphosphate pools on *in vitro* mutagenesis. *J Biol Chem*. **254**, 5718-25.

Kuroda K et Ueda M, 2006. Effective display of metallothionein tandem repeats on the bioadsorption of cadmium ion. *Appl Microbiol Biotechnol*. **70**, 458-63.

Kurokawa K, Itoh T, Kuwahara T, *et al.*, 2007. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res*. **14**, 169-81.

Lanfranco L, 2007. The fine-tuning of heavy metals in mycorrhizal fungi. *New Phytol*. **174**, 3-6.

Lanfranco L, Bolchi A, Ros EC, *et al.*, 2002. Differential expression of a metallothionein gene during the presymbiotic versus the symbiotic phase of an arbuscular mycorrhizal fungus. *Plant Physiol*. **130**, 58-67.

Lara E, Chatzinotas A et Simpson AG, 2006. Andalucia (n. gen.)-the deepest branch within jakobids (Jakobida; Excavata), based on morphological and molecular study of a new flagellate from soil. *J Eukaryot Microbiol*. **53**, 112-20.

Lara E, Moreira D et Lopez-Garcia P, 2009. The Environmental Clade LKM11 and Rozella Form the Deepest Branching Clade of Fungi. *Protist*.

Lawley B, Ripley S, Bridge P, *et al.*, 2004. Molecular analysis of geographic patterns of eukaryotic diversity in Antarctic soils. *Appl Environ Microbiol*. **70**, 5963-72.

Lecointre G et Le-Guyader H, 2001. Classification phylogénétique du vivant. *Belin, Paris*.

Lee J, Godon C, Lagniel G, *et al.*, 1999. Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. *J Biol Chem*. **274**, 16040-6.

Lee SW, Won K, Lim HK, *et al.*, 2004. Screening for novel lipolytic enzymes from uncultured soil microorganisms. *Appl Microbiol Biotechnol*. **65**, 720-6.

Lefevre E, Roussel B, Amblard C, *et al.*, 2008. The molecular diversity of freshwater picoeukaryotes reveals high occurrence of putative parasitoids in the plankton. *PLoS One*. **3**, e2324.

Leininger S, Urich T, Schloter M, *et al.*, 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*. **442**, 806-9.

Lesaulnier C, Papamichail D, McCorkle S, *et al.*, 2008. Elevated atmospheric CO2 affects soil microbial diversity associated with trembling aspen. *Environ Microbiol*. **10**, 926-41.

Li ZS, Lu YP, Zhen RG, *et al.*, 1997. A new pathway for vacuolar cadmium sequestration in *Saccharomyces cerevisiae*: YCF1-catalyzed transport of bis(glutathionato)cadmium. *Proc Natl Acad Sci U S A*. **94**, 42-7.

Liles MR, Manske BF, Bintrim SB, *et al.*, 2003. A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl Environ Microbiol*. **69**, 2684-91.

Liu J, Qu W et Kadiiska MB, 2009. Role of oxidative stress in cadmium toxicity and carcinogenesis. *Toxicol Appl Pharmacol*. **238**, 209-14.

Liu Y, Zhou J, Omelchenko MV, *et al.*, 2003. Transcriptome dynamics of *Deinococcus radiodurans* recovering from ionizing radiation. *Proc Natl Acad Sci U S A*. **100**, 4191-6.

Lopéz-Garcia P et Moreira D, 2008. Tracking microbial biodiversity through molecular and genomic ecology. *Research in Microbiology*. **159**, 67-73.

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, *et al.*, 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*. **409**, 603-7.

Ma J, Ryan P et Delhaize E, 2001. Aluminium tolerance in plants and the complexing role of organic acids. *Trends in Plant Science*. **6**, 273-278.

Maclean RC, Richardson DJ, LePardo R, *et al.*, 2004. The identification of *Naegleria fowleri* from water and soil samples by nested PCR. *Parasitol Res*. **93**, 211-7.

Majernik A, Gottschalk G et Daniel R, 2001. Screening of environmental DNA libraries for the presence of genes conferring Na(+)(Li(+))/H(+) antiporter activity on *Escherichia coli*: characterization of the recovered genes and the corresponding gene products. *J Bacteriol*. **183**, 6645-53.

Mangot JF, Lepere C, Bouvier C, *et al.*, 2009. Community structure and dynamics of small eukaryotes (<5 {micro}m) targeted by new oligonucleotide probes: a new insight into the lacustrine microbial food web. *Appl Environ Microbiol*.

Margulis L, 1975. Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. *Symp Soc Exp Biol*. 21-38.

Massana R, Terrado R, Forn I, *et al.*, 2006. Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. *Environ Microbiol*. **8**, 1515-22.

Mayumi D, Akutsu-Shigeno Y, Uchiyama H, *et al.*, 2008. Identification and characterization of novel poly(DL-lactic acid) depolymerases from metagenome. *Appl Microbiol Biotechnol*. **79**, 743-50.

McGrath KC, Thomas-Hall SR, Cheng CT, *et al.*, 2008. Isolation and analysis of mRNA from environmental microbial communities. *J Microbiol Methods*. **75**, 172-6.

Medini D, Serruto D, Parkhill J, *et al.*, 2008. Microbiology in the post-genomic era. *Nat Rev Microbiol*. **6**, 419-30.

Meharg A et Cairney J, 2000. Co-evolution of mycorrhizal symbionts and their hosts to metal-contamined environments. *Advances in Ecological Research*. **30**, 69-112.

Meharg AA, 2003. The mechanistic basis of interactions between mycorrhizal associations and toxic metal cations. *Mycol Res*. **107**, 1253-65.

Mehra RK et Winge DR, 1991. Metal ion resistance in fungi: molecular mechanisms and their regulated expression. *J Cell Biochem*. **45**, 30-40.

Mendez-Alvarez S, Rüfenacht K et Eggen R, 2000. The Oxidative Stress-Sensitive yap1 Null Strain of *Saccharomyces cerevisiae* Becomes Resistant Due to Increased

Carotenoid Levels upon the Introduction of the *Chlamydomonas reinhardtii* cDNA, Coding for the 60S Ribosomal Protein L10a. *Biochemical and Biophysical Research Communications*. **267**, 953-959.

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

Miller DN, Bryant JE, Madsen EL, *et al.*, 1999. Evaluation and optimization of DNA extraction and purification procedures for soil and sediment samples. *Appl Environ Microbiol*. **65**, 4715-24.

Mills HJ, Martinez RJ, Story S, *et al.*, 2005. Characterization of microbial community structure in Gulf of Mexico gas hydrates: comparative analysis of DNA- and RNA-derived clone libraries. *Appl Environ Microbiol*. **71**, 3235-47.

Minet M, Dufour ME et Lacroute F, 1992. Complementation of *Saccharomyces cerevisiae* auxotrophic mutants by Arabidopsis thaliana cDNAs. *Plant J*. **2**, 417-22.

Mirete S, de Figueras CG et Gonzalez-Pastor JE, 2007. Novel nickel resistance genes from the rhizosphere metagenome of plants adapted to acid mine drainage. *Appl Environ Microbiol*. **73**, 6001-11.

Mironov AA, Banin VV, Sesorova IS, *et al.*, 2007. Evolution of the endoplasmic reticulum and the Golgi complex. *Adv Exp Med Biol*. **607**, 61-72.

Moon-van der Staay SY, De Wachter R et Vaulot D, 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*. **409**, 607-10.

Moon-van der Staay SY, Tzeneva VA, van der Staay GW, *et al.*, 2006. Eukaryotic diversity in historical soil samples. *FEMS Microbiol Ecol*. **57**, 420-8.

Moreira D et Lopéz-Garcia P, 2002. The molecular ecology of microbial eukaryotes unveils a hidden diversity. *Trends in Microbiology*. **10**, 31-39.

Murasugi A, Wada C et Hayashi Y, 1981. Cadmium-binding peptide induced in fission yeast, *Schizosaccharomyces pombe*. *J Biochem*. **90**, 1561-4.

Muttray A et Mohn W, 1999. Quantitation of the population size and metabolic activity of a resin acid degrading bacterium in activated sludge using slot-blot hybridization to measure the rRNA:rDNA ratio. *Microbial Ecology*. **38**, 348-357.

Nara K, 2006a. Pioneer dwarf willow may facilitate tree succession by providing late colonizers with compatible ectomycorrhizal fungi in a primary successional volcanic desert. *New Phytologist*. **171**, 187-198.

Nara K, 2006b. Ectomycorrhizal networks and seedling establishment during early primary succession. *New Phytologist*. **169**, 169-178.

Nickrent DL et Sargent ML, 1991. An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. *Nucleic Acids Res*. **19**, 227-35.

Nieboer E et Richardson D, 1980. The replacement of the nondescript term 'heavy metals' by a biologically and chemically significant classification of metal ions. *Environmental Pollution Series B, Chemical and Physical*. **1**, 3-26.

Nikolaev SI, Berney C, Petrov NB, *et al.*, 2006. Phylogenetic position of Multicilia marina and the evolution of Amoebozoa. *Int J Syst Evol Microbiol*. **56**, 1449-58.

Nishimura K, Igarashi K et Kakinuma Y, 1998. Proton gradient-driven nickel uptake by vacuolar membrane vesicles of *Saccharomyces cerevisiae*. *J Bacteriol*. **180**, 1962-4.

Not F, Valentin K, Romari K, *et al.*, 2007. Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science*. **315**, 253-5.

Nowack EC, Melkonian M et Glockner G, 2008. Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol*. **18**, 410-8.

O'Brien HE, Parrent JL, Jackson JA, *et al.*, 2005. Fungal community analysis by large-scale sequencing of environmental samples. *Appl Environ Microbiol*. **71**, 5544-50.
Ochiai E, 1987. General principles of biochemistry of the elements. *New York: Plenum Press*.
Ogram A, Sayler G et Barkay T, 1987. The extraction and purification of microbial DNA from sediments. *J. Microbiol. Methods*. **7**, 57-66.
Pace NR, 1997. A molecular view of microbial diversity and the biosphere. *Science*. **276**, 734-40.
Pace NR, 2009. Problems with "procaryote". *J Bacteriol*. **191**, 2008-11.
Pagani A, Villarreal L, Capdevila M, *et al.*, 2007. The *Saccharomyces cerevisiae* Crs5 Metallothionein metal-binding abilities and its role in the response to zinc overload. *Mol Microbiol*. **63**, 256-69.
Parfrey LW, Barbero E, Lasser E, *et al.*, 2006. Evaluating support for the current classification of eukaryotic diversity. *PLoS Genet*. **2**, e220.
Park J, Lee J et Jung J, 2003. Cadmium uptake capacity of two strains of *Saccharomyces cerevisiae* cells. *Enz Microb Technol*. **33**, 371-378.
Pathak GP, Ehrenreich A, Losi A, *et al.*, 2009. Novel blue light-sensitive proteins from a metagenomic approach. *Environ Microbiol*. **11**, 2388-99.
Patron NJ, Inagaki Y et Keeling PJ, 2007. Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host lineages. *Curr Biol*. **17**, 887-91.
Paul E et Clark F, 1989. Soil microbiology and biochemistry. *Academic Press, San Diego, CA*.
Pawlowski J, 2000. Introduction to the molecular systematics of foraminifera. *Micropaleontology* **46**, 1–12.
Pawlowski J et Burki F, 2009. Untangling the phylogeny of amoeboid protists. *J Eukaryot Microbiol*. **56**, 16-25.
Pazour GJ, Agrin N, Leszyk J, *et al.*, 2005. Proteomic analysis of a eukaryotic cilium. *J Cell Biol*. **170**, 103-13.
Philippe H et Germot A, 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. *Mol Biol Evol*. **17**, 830-4.
Poretsky RS, Bano N, Buchan A, *et al.*, 2005. Analysis of microbial gene transcripts in environmental samples. *Appl Environ Microbiol*. **71**, 4121-6.
Poretsky RS, Hewson I, Sun S, *et al.*, 2009. Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol*. **11**, 1358-75.
Porter TM, Schadt CW, Rizvi L, *et al.*, 2008. Widespread occurrence and phylogenetic placement of a soil clone group adds a prominent new branch to the fungal tree of life. *Mol Phylogenet Evol*. **46**, 635-44.
Precigou S, Goulas P et Duran R, 2001. Rapid and specific identification of nitrile hydratase (NHase)-encoding genes in soil samples by polymerase chain reaction. *FEMS Microbiol Lett*. **204**, 155-61.
Prescott DM, 2000. Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nat Rev Genet*. **1**, 191-8.
Pruesse E, Quast C, Knittel K, *et al.*, 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*. **35**, 7188-96.
Qiu J, Guo Z, Liu H, *et al.*, 2008. DNA microarray-based global transcriptional profiling of *Yersinia pestis* in multicellularity. *J Microbiol*. **46**, 557-63.
Radajewski S, McDonald IR et Murrell JC, 2003. Stable-isotope probing of nucleic acids: a window to the function of uncultured microorganisms. *Curr Opin Biotechnol*. **14**, 296-302.

Reyes-Prieto A, Weber AP et Bhattacharya D, 2007. The origin and establishment of the plastid in algae and plants. *Annu Rev Genet*. **41**, 147-68.

Richards TA, Vepritskiy AA, Gouliamova DE*, et al.*, 2005. The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol*. **7**, 1413-25.

Riesenfeld CS, Goodman RM et Handelsman J, 2004a. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol*. **6**, 981-9.

Riesenfeld CS, Goodman RM et Handelsman J, 2004b. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol*. **6**, 981-9.

Rodriguez-Ezpeleta N, Brinkmann H, Burey SC*, et al.*, 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Curr Biol*. **15**, 1325-30.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B*, et al.*, 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol*. **56**, 389-99.

Rondon MR, August PR, Bettermann AD*, et al.*, 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl Environ Microbiol*. **66**, 2541-7.

Rowlands T, Baumann P et Jackson SP, 1994. The TATA-binding protein: a general transcription factor in eukaryotes and archaebacteria. *Science*. **264**, 1326-9.

Rühling A et Söderström B, 1990. Changes in fruitbody production of mycorrhizal and litter decomposing macromycetes in heavy metal polluted coniferous forests in North Sweden. *Water Air Soil Pollut* **49**, 375-387.

Ruotolo R, Marchini G et Ottonello S, 2008. Membrane transporters and protein traffic networks differentially affecting metal tolerance: a genomic phenotyping study in yeast. *Genome Biol*. **9**, R67.

Saitou N et Nei M, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. **4**, 406-25.

Sambrook J, Fritsch F et Maniatis T, 1989. Molecular cloning: a laboratory manual. *Cold Spring Harbor, New York*.

Sanità-di-Toppi L et Gabbrielli R, 1999. Response to cadmium in higher plant. *Env. Exp. Bot*. **41**, 105-130.

Sapp J, 1994. Evolution by association: a history of symbiosis. *(Oxford Univ. Press, New York)*.

Schat H, Llugany M et Bernhard R, 2000. Metal-specific patterns of tolerance, uptake and transport of heavy metals in hyperaccumulating and nonhyperaccumulating metallophytes. *Phytoremediation of contaminated soil and water. Lewis Publishers, Boca Raton, FL, USA*. 171-188.

Schutzendubel A et Polle A, 2002. Plant responses to abiotic stresses: heavy metal-induced oxidative stress and protection by mycorrhization. *J Exp Bot*. **53**, 1351-65.

Sebat JL, Colwell FS et Crawford RL, 2003. Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl Environ Microbiol*. **69**, 4927-34.

Sessitsch A, Gyamfi S, Stralis-Pavese N, *et al.*, 2002. RNA isolation from soil for bacterial community and functional analysis: evaluation of different extraction and soil conservation protocols. *J Microbiol Methods*. **51**, 171-9.

Shalchian-Tabrizi K, Kauserud H, Massana R, *et al.*, 2007. Analysis of environmental 18S ribosomal RNA sequences reveals unknown diversity of the cosmopolitan phylum *Telonemia*. *Protist*. **158**, 173-80.

Shamrock VJ et Lindsey GG, 2008. A compensatory increase in trehalose synthesis in response to desiccation stress in *Saccharomyces cerevisiae* cells lacking the heat shock protein Hsp12p. *Can J Microbiol*. **54**, 559-68.

Shang Y, Song X, Bowen J, *et al.*, 2002. A robust inducible-repressible promoter greatly facilitates gene knockouts, conditional expression, and overexpression of homologous and heterologous genes in *Tetrahymena thermophila*. *Proc Natl Acad Sci U S A*. **99**, 3734-9.

Shiraishi E, Inouhe M, Joho M, *et al.*, 2000. The cadmium-resistant gene, CAD2, which is a mutated putative copper-transporter gene (PCA1), controls the intracellular cadmium-level in the yeast *S. cerevisiae*. *Curr Genet*. **37**, 79-86.

Simon N, LeBot N, Marie D, *et al.*, 1995. Fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes to identify small phytoplankton by flow cytometry. *Appl Environ Microbiol*. **61**, 2506-13.

Simpson AG, 2003. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *Int J Syst Evol Microbiol*. **53**, 1759-77.

Simpson AG, Inagaki Y et Roger AJ, 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. *Mol Biol Evol*. **23**, 615-25.

Slapeta J, Moreira D et Lopez-Garcia P, 2005. The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proc Biol Sci*. **272**, 2073-81.

Smit E, Leeflang P, Gommans S, *et al.*, 2001. Diversity and seasonal fluctuations of the dominant members of the bacterial soil community in a wheat field as determined by cultivation and molecular methods. *Appl Environ Microbiol*. **67**, 2284-91.

Steenkamp ET, Wright J et Baldauf SL, 2006. The protistan origins of animals and fungi. *Mol Biol Evol*. **23**, 93-106.

Steffan RJ, Goksoyr J, Bej AK, *et al.*, 1988. Recovery of DNA from soils and sediments. *Appl Environ Microbiol*. **54**, 2908-15.

Stiller JW et Hall BD, 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol Biol Evol*. **16**, 1270-9.

Stoeck T, Hayward B, Taylor GT, *et al.*, 2006. A multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. *Protist*. **157**, 31-43.

Stoeck T, Taylor GT et Epstein SS, 2003. Novel eukaryotes from the permanently anoxic Cariaco Basin (Caribbean Sea). *Appl Environ Microbiol*. **69**, 5656-63.

Stohs SJ et Bagchi D, 1995. Oxidative mechanisms in the toxicity of metal ions. *Free Radic Biol Med*. **18**, 321-36.

Stommel M, Mann P et Franken P, 2001. EST-library construction using spore RNA of the arbuscular mycorrhizal fungus *Gigaspora rosea*. *Mycorrhiza*. **10**, 281-285.

Tokuda G et Watanabe H, 2007. Hidden cellulases in termites: revision of an old hypothesis. *Biol Lett*. **3**, 336-9.

Torsvik V, Daae FL, Sandaa RA, *et al.*, 1998. Novel techniques for analysing microbial diversity in natural and perturbed environments. *J Biotechnol*. **64**, 53-62.

Torsvik V et Ovreas L, 2002. Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol*. **5**, 240-5.

Trassy C et Mermet J, 1984. Les Applications Analytiques des Plasmas HF. *Technique et Documentation (Lavoisier), Paris*.

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

Turnau K, Kottle I, Dexheimer J, *et al.*, 1994. Element distribution in *Pisolithus tinctorius* mycelium treated with cadmium dust. *Bot Acta*. **74**, 137-142.

Turnbaugh PJ, Ley RE, Hamady M, *et al.*, 2007. The human microbiome project. *Nature*. **449**, 804-10.

Tyson GW, Chapman J, Hugenholtz P*, et al.*, 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. **428**, 37-43.

Uchiyama T, Abe T, Ikemura T*, et al.*, 2005. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol*. **23**, 88-93.

Urich T, Lanzen A, Qi J*, et al.*, 2008. Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One*. **3**, e2527.

Valster RM, Wullings BA, Bakker G, *et al.*, 2009. Free-living protozoa in two unchlorinated drinking water supplies, identified by phylogenic analysis of 18S rRNA gene sequences. *Appl Environ Microbiol*. **75**, 4736-46.

Van-Assche F et Clijsters H, 1990. Effects of metals on enzyme activity in plants. *Plant, Cell and Environment*. **13**, 195-206.

Van de Peer Y, Baldauf SL, Doolittle WF*, et al.*, 2000. An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *J Mol Evol*. **51**, 565-76.

van der Lelie D, Schwitzguebel JP, Glass DJ*, et al.*, 2001. Assessing phytoremediation's progress in the United States and Europe. *Environ Sci Technol*. **35**, 446A-452A.

Vandenkoornhuyse P, Baldauf SL, Leyval C*, et al.*, 2002. Extensive fungal diversity in plant roots. *Science*. **295**, 2051.

Vangronsveld J, Colpaert JV et Van Tichelen KK, 1996. Reclamation of a bare industrial area contaminated by non-ferrous metals: physico-chemical and biological evaluation of the durability of soil treatment and revegetation. *Environ Pollut*. **94**, 131-40.

Vangronsveld J, Van Assche F et Clijsters H, 1995. Reclamation of a bare industrial area contaminated by non-ferrous metals: in situ metal immobilization and revegetation. *Environ Pollut*. **87**, 51-9.

Vellai T et Vida G, 1999. The origin of eukaryotes: the difference between prokaryotic and eukaryotic cells. *Proc. Biol. Sci.* **266**, 1571-1577.

Venter JC, Remington K, Heidelberg JF*, et al.*, 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. **304**, 66-74.

Vido K, Spector D, Lagniel G*, et al.*, 2001. A proteome analysis of the cadmium response in *Saccharomyces cerevisiae*. *J Biol Chem*. **276**, 8469-74.

Viprey M, Guillou L, Ferreol M*, et al.*, 2008. Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environ Microbiol*. **10**, 1804-22.

Vivas A, Barea JM, Biro B*, et al.*, 2006. Effectiveness of autochthonous bacterium and mycorrhizal fungus on *Trifolium* growth, symbiotic development and soil enzymatic activities in Zn contaminated soil. *J Appl Microbiol*. **100**, 587-98.

Vogel T, Simonet P, Jansson J*, et al.*, 2009. TerraGenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology*. **7**, 252.

Voget S, Leggewie C, Uesbeck A*, et al.*, 2003. Prospecting for novel biocatalysts in a soil metagenome. *Appl Environ Microbiol*. **69**, 6235-42.

Voiblet C, Duplessis S, Encelot N*, et al.*, 2001. Identification of symbiosis-regulated genes in *Eucalyptus globulus-Pisolithus tinctorius* ectomycorrhiza by differential hybridization of arrayed cDNAs. *Plant J*. **25**, 181-91.

Waisberg M, Joseph P, Hale B*, et al.*, 2003. Molecular and cellular mechanisms of cadmium carcinogenesis. *Toxicology*. **192**, 95-117.

Wang B et Qiu YL, 2006. Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza*. **16**, 299-363.

Wang GY, Graziani E, Waters B, *et al.*, 2000. Novel natural products from soil DNA libraries in a streptomycete host. *Org Lett*. **2**, 2401-4.

Warnecke F, Luginbuhl P, Ivanova N, *et al.*, 2007. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*. **450**, 560-5.

Watanabe M et Suzuki T, 2002. Involvement of reactive oxygen stress in cadmium-induced cellular damage in *Euglena gracilis*. *Comp Biochem Physiol C Toxicol Pharmacol*. **131**, 491-500.

Weber M, Trampczynska A et Clemens S, 2006. Comparative transcriptome analysis of toxic metal responses in *Arabidopsis thaliana* and the Cd(2+)-hypertolerant facultative metallophyte *Arabidopsis halleri*. *Plant Cell Environ*. **29**, 950-63.

Weissman Z, Berdicevsky I, Cavari BZ, *et al.*, 2000. The high copper tolerance of *Candida albicans* is mediated by a P-type ATPase. *Proc Natl Acad Sci U S A*. **97**, 3520-5.

Wellington EM, Berry A et Krsek M, 2003. Resolving functional diversity in relation to microbial community structure in soil: exploiting genomics and stable isotope probing. *Curr Opin Microbiol*. **6**, 295-301.

Wemmie JA, Szczypka MS, Thiele DJ, *et al.*, 1994. Cadmium tolerance mediated by the yeast AP-1 protein requires the presence of an ATP-binding cassette transporter-encoding gene, YCF1. *J Biol Chem*. **269**, 32592-7.

Whittaker RH, 1969. New concepts of kingdoms or organisms. Evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science*. **163**, 150-60.

Wilkinson D et Dickinson N, 1995. Metal resistance in trees - the role of mycorrhizae. *Oikos*. **72**, 298-300.

Woese CR, 1987. Bacterial evolution. *Microbiol Rev*. **51**, 221-71.

Woese CR et Fox GE, 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*. **74**, 5088-90.

Wu AL et Moye-Rowley WS, 1994. GSH1, which encodes gamma-glutamylcysteine synthetase, is a target gene for YAP-1 transcriptional regulation. *Mol Cell Biol*. **14**, 5832-9.

Wu L, Thompson DK, Li G, *et al.*, 2001. Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Appl Environ Microbiol*. **67**, 5780-90.

Xu Z, Bowers N et Pratt J, 1997. Variation in morphology, ecology and toxicological responses of *Colpoda inflata* (Stokes) collected from five biogeographic realms. *Eur J Protistol* **33**, 136-144.

Yergeau E, Bokhorst S, Huiskes AH, *et al.*, 2007. Size and structure of bacterial, fungal and nematode communities along an Antarctic environmental gradient. *FEMS Microbiol Ecol*. **59**, 436-51.

Yilmaz S et Zengin M, 2004. Monitoring environmental pollution in Erzurum by chemical analysis of Scots pine (*Pinus sylvestris L*.) needles. *Environ Int*. **29**, 1041-7.

Yun J, Kang S, Park S, *et al.*, 2004. Characterization of a novel amylolytic enzyme encoded by a gene from a soil-derived metagenomic library. *Appl Environ Microbiol*. **70**, 7229-35.

Zhou J, Bruns MA et Tiedje JM, 1996. DNA recovery from soils of diverse composition. *Appl Environ Microbiol*. **62**, 316-22.

Zhou J et Thompson DK, 2002a. Challenges in applying microarrays to environmental studies. *Curr Opin Biotechnol*. **13**, 204-7.

Zhou J, Xia B, Treves DS, *et al.*, 2002b. Spatial and resource factors influencing high microbial diversity in soil. *Appl Environ Microbiol*. **68**, 326-34.

Annexe 1

Milieu LB (Luria-Bertani) (d'après Sambrook *et al.*, 1989)

Nacl	10 $g.L^{-1}$
Extrait de levure	5 $g.L^{-1}$
Tryptone	10 $g.L^{-1}$
Agar	15 $g.L^{-1}$

Ajuster le pH à 7 avec du NaOH 1M.
Autoclaver 20 minutes à 120°C.

Milieu SOC

Bactotryptone	20 $g.L^{-1}$
Extrait de levure	5 $g.L^{-1}$
Glucose	20 mM

Ajuster le pH à 7,2-7,4 et autoclaver 20 minutes à 120°C.
Les solutions suivantes sont stérilisées par filtration et ajoutées après autoclavage.

NaCl	10 mM
KCl	2,5 mM
$MgCl_2$	10 mM
$MgSO_4$	10 mM

Milieu YPG

Extrait de levure	10 $g.L^{-1}$
Bactopeptone	10 $g.L^{-1}$
Glucose	20 $g.L^{-1}$
Agar	20 $g.L^{-1}$

Autoclaver 30 minutes à 110°C.

Milieu YNB

Yeast Nitrogen Base	6,7$g.L^{-1}$
Glucose	20 $g.L^{-1}$
Adénine	1 $mL.L^{-1}$
Uracile	4 $mL.L^{-1}$ (sauf pour le milieu sélectif correspondant)
Tyrosine	5 $mL.L^{-1}$

Ajuster le pH à 5,9 avant de mettre l'agar (20$g.L^{-1}$).
Autoclaver 30 minutes à 110°C.

Après autoclave, laisser refroidir jusqu'à 50-60 °C et ajouter les solutions suivantes (stériliser par filtration) :
Acides aminés complet 10 $mL.L^{-1}$

Arginine	$2g.L^{-1}$
Histidine	$2g.L^{-1}$
Isoleucine	$6g.L^{-1}$
Leucine	$6g.L^{-1}$
Lysine	$4g.L^{-1}$
Méthionine	$2g.L^{-1}$
Phénylalanine	$6g.L^{-1}$
Thréonine	$5g.L^{-1}$
Tryptophane	$4g.L^{-1}$
Valine	$15g.L^{-1}$

Annexe 2

Test KHI-2	Fungi	Metazoa	Amoebozoa	Alveolata	Rhizaria	Heterokonta	Excavata	Plantae	Autres
ADNr+ARNr 3' Balen vs Paal	**0,160**	**$3{,}87.10^{-4}$**	**$7{,}61.10^{-7}$**	**0,023**	**0,007**	**0,614**	**0,081**	**$3{,}84.10^{-3}$**	**$2{,}54.10^{-4}$**
ADNr+ARNr 5' Balen vs Paal	**0,610**	**$2{,}36.10^{-3}$**	**0,069**	**0,123**	**0,406**	**0,960**	**0,1286**	**$1{,}76.10^{-9}$**	**0,378**
ADNr+ARNr 5' Balen vs Lommel	**$1{,}4.10^{-10}$**	**$1{,}63.10^{-3}$**	**0,583**	**$2{,}13.10^{-9}$**	**$5{,}5.10^{-3}$**	**0,938**	**0,032**	**0,072**	**0,938**
ADNr+ARNr 5' Paal vs Lommel	**$6{,}4.10^{-12}$**	**0,283**	**0,513**	**$4{,}93.10^{-6}$**	**$4{,}1.10^{-4}$**	**0,962**	**$2{,}7.10^{-4}$**	**0,026**	**0,629**
Balen 5' ADNr/ARNr	**$1{,}2.10^{-9}$**	**$1{,}21.10^{-7}$**	**0,864**	**0,258**	**$1{,}71.10^{-6}$**	**0,575**	**0,971**	**0,913**	**0,085**
Balen 3' ADNr/ARNr	**0,049**	**0,571**	**0,774**	**0,416**	**0,194**	**0,990**	**0,181**	**0,031**	**$3{,}38.10^{-3}$**
Paal 5' ADNr/ARNr	**$4{,}9.10^{-3}$**	**$2{,}29.10^{-5}$**	**$4{,}06.10^{-9}$**	**0,550**	**$1{,}12.10^{-9}$**	**0,937**	**0,122**	**0,049**	**0,249**
Paal 3' ADNr/ARNr	**$1{,}9.10^{-6}$**	**0,330**	**$3{,}19.10^{-13}$**	**$9{,}1.10^{-4}$**	**0,221**	**0,943**	**0,302**	**$1{,}60.10^{-5}$**	**0,343**
Lommel 5' ADNr/ARNr	**0,679**	**0,306**	**0,071**	**0,004**	**0,001**	**0,380**	**0,284**	**0,0187**	**0,380**

Valeurs des tests de KHI-2 obtenues par comparaison des jeux de séquences ribosomiques disponibles pour chaque site. Les ADNr et les ARNr cumulés représentant un même phylum eucaryote ont été comparés entre site pour chaque portion du gène 18S analysée. Les ADNr et les ARNr représentant un même phylum eucaryote ont été comparés entre eux pour chaque site et pour chaque portion du gène 18S analysée.

Test KHI-2	Protéines annotées	Protéines hypothétiques conservées	Nouvelles protéines hypothétiques
ADNc Balen / ADNc Paal	**$6{,}17.10^{-9}$**	**$2{,}59.10^{-8}$**	**$1{,}32.10^{-20}$**

Valeurs des tests de KHI-2 obtenues par comparaison des jeux de séquences d'ADNc de Balen et de Paal pour chaque catégorie d'ADNc. Les protéines annotées, correspondent aux protéines avec une fonction biologique connue ; les protéines hypothétiques conservées, correspondent aux protéines de fonction inconnue déposées dans les bases de données ; et les nouvelles protéines hypothétiques, correspondent aux ADNc qui ne donnent aucune réponse lors d'un Blastx contre les bases de données publiques (e-value < $1e^{-5}$).

Test KHI-2	A	B	C	D	E	F	G	H
ADNc Balen / ADNc Paal	**0,655**	**0,416**	**$8{,}24.10^{-5}$**	**0,034**	**0,330**	**$5{,}37.10^{-3}$**	**0,842**	**0,138**

Test KHI-2	I	J	K	L	M	N	O
ADNc Balen / ADNc Paal	**0,992**	**0,894**	**$3{,}7.10^{-6}$**	**0,159**	**0,147**	**0,973**	**0,329**

Valeurs des tests de KHI-2 obtenues par comparaison des jeux de séquences d'ADNc annotés de Balen et de Paal pour chaque groupe fonctionnel d'ADNc.
A, métabolisme des acides nucléiques ; B, métabolisme des protéines ; C, métabolisme des acides aminés et dérivés ; D, métabolisme des acides gras et des lipides ; E, métabolisme des glucides ; F, métabolisme secondaire ; G, autres métabolismes (K, N, S…) ; H, cofacteurs, vitamines, groupes prosthétiques, pigments ; I, trafic cellulaire ; J, contrôle du cycle cellulaire ; K, respiration ; L, réponses au stress ; M, virulence ; N, paroi cellulaire ; O, autres fonctions