

Présentation des projets financés au titre de l'édition 2008 du
Programme Génomique Microbienne à Grande Echelle

ACRONYME et titre du projet	Page
EVOGENO - Genomic evolution and dynamics during long-term bacterial adaptation.	2
EVOLDEEP - Evolutionary and population genomics of deep-sea uncultivated microorganisms	4
METASOIL - Metagenomic discovery and exploitation of the soil microbial community	6
GENOZYME - Mining genomes for novel biocatalysts	8
GENOPOP-GBS - Génomique des populations de <i>Streptococcus agalactiae</i> : diversification des génomes et adaptation à l'hôte. microbial community	10
CBME - Computational Biology for Metagenomics Experiments	12
BAMBI - Bio-prospection par une Approche Métagénomique de Biocatalyseurs chez des Invertébrés xylophages : analyse des microbiomes d'invertébrés du sol de la forêt malgache	14

Titre du projet

EVOGENO - Genomic evolution and dynamics during long-term bacterial adaptation

Résumé

La séquence de centaines de génomes bactériens a mis en évidence l'extraordinaire diversité et plasticité des réseaux métaboliques et de régulation. Cependant, la dynamique de l'évolution de ces diverses fonctions, et en particulier le lien entre les modifications génomiques, la diversité et la performance d'un organisme vivant restent des tâches difficiles à appréhender, mais fascinantes. Pour la plupart des microbiologistes, le terme « mécanisme » se rapporte aux interactions biochimiques ou régulatrices entre les gènes, les protéines et les métabolites dans une cellule. Toutes ces interactions doivent cependant être reliées au fitness, c'est-à-dire la mesure de la capacité reproductive d'un génotype, qui est le paramètre ultime du succès écologique. La dynamique de génomes entiers est une des clés des changements de fitness, conduisant in fine aux interactions entre génotypes et entre les génotypes et leur environnement. En reliant les changements génomiques aux changements de fitness, il est possible d'étudier comment la sélection naturelle est capable de remodeler et d'améliorer des génomes entiers, et quelles fonctions sont les plus plastiques au cours du temps évolutif. Un tel cadre évolutif est totalement complémentaire de la plupart des approches de « Biologie des systèmes », qui tentent de comprendre le fonctionnement global d'un organisme par l'analyse précise d'un clone de référence et procurent donc une description statique des réseaux métaboliques et de régulation. Disposer d'un contexte évolutif est cependant difficile, principalement par le fait que les événements adaptatifs permettant l'émergence de la structure génomique actuelle de l'organisme en question se sont produits à un temps inconnu dans le passé, dans des conditions inconnues, et sous des contraintes génomiques inconnues. Pour pallier à ces limitations, nous pouvons reproduire l'évolution en laboratoire dans des conditions contrôlées.

Dans ce projet, nous allons employer la plus longue expérience d'évolution en cours, pendant laquelle un ancêtre d'*Escherichia coli* a été utilisé pour propager douze populations dans un environnement défini pendant 40.000 générations. La dynamique de l'évolution des génomes sera analysée par séquençage Solexa du génome de 135 clones

évolués, isolés des différentes populations à chacun de dix temps d'évolution différents. Parmi les questions auxquelles ce projet permettra de répondre, nous pourrons comprendre les événements génétiques successifs conduisant à l'accroissement de la performance des bactéries dans leur environnement. Des analyses génétiques, phénotypiques et fonctionnelles rigoureuses seront combinées pour disséquer de façon complète les différents chemins adaptatifs dans ces populations. Nous étudierons comment l'apparition de mutations adaptatives est contrainte par la présence d'autres mutations, et comment elles affectent les réseaux métaboliques et de régulation. Nous analyserons ainsi pour la première fois à quel point les caractères génomiques peuvent évoluer et quelles sont les bases moléculaires de ces capacités évolutives.

Partenaires

CNRS (partenaire coordinateur)
CEA Genoscope-Institut de Génomique
INSERM

Coordinateur

Dominique Schneider – CNRS
dominique.schneider@ujf-grenoble.fr

Aide de l'ANR

669 662 euros

Début et durée

Janvier 2009 - 48 mois

Référence

ANR-08-GENM-023

Titre du projet **EVOLDEEP - Evolutionary and population genomics of deep-sea uncultivated microorganisms**

Résumé

L'océan profond est l'un des écosystèmes terrestres les moins connus malgré le fait qu'il constitue l'un des plus vastes et que les activités des communautés microbiennes qui s'y développent sont fondamentales pour la complétion des cycles biogéochimiques globaux de notre planète. Des approches moléculaires de caractérisation de la diversité basées sur l'amplification, le clonage et le séquençage des gènes d'ARNr de la petite sous-unité du ribosome (SSU rRNA) ont révélé la présence de lignées de bactéries, d'archées et d'eucaryotes, souvent divergentes des lignées cultivées connues, dans le picoplancton marin des grandes profondeurs. Parmi celles-ci, des lignées spécifiques d'archées mésophiles, notamment les Crenarchaeota du Group I et les Euryarchaeota du Group II, mais aussi une diversité importante de bactéries comprenant, de manière surprenante, des phyla que l'on pensait typiques du sol comme les Acidobacteria, sembleraient quantitativement importantes dans le picoplancton profond. La métagénomique s'est révélée au cours des dernières années une stratégie puissante pour obtenir des informations complémentaires sur la structure, l'écologie et l'évolution des communautés microbiennes du plancton océanique en permettant l'accès aux gènes des microorganismes qui demeurent pour la plupart non cultivés. Toutefois, la majorité de ces études a été exclusivement focalisée sur la zone euphotique. Plusieurs partenaires impliqués dans ce projet ont réalisé des analyses métagénomiques pionnières sur les communautés planctoniques de l'océan profond lors des dernières années comprenant aussi bien des études à petite échelle (clones génomiques sélectionnés) que des études beaucoup plus massives à partir du séquençage aléatoire des extrémités des inserts des fosmidés des banques métagénomiques. Ces travaux nous ont mené à la construction de plusieurs banques métagénomiques contenant des inserts génomiques de grande taille comprenant environ 75000 clones (taille moyenne des inserts génomiques ~35 kbp, soit ~2.5 Gbp d'information génomique archivée) et provenant de différentes régions océaniques (les mers Ionienne et Adriatique dans le bassin Méditerranéen, l'Atlantique Sud et le

Front Polaire Antarctique, profondeurs entre 500 et 3000 m). Ces banques métagénomiques, qui sont bien caractérisées en terme de diversité des SSU rRNAs et qui ont permis les premières analyses métagénomiques comparatives du plancton océanique profond, sont en notre possession à l'ESE (UMR8079) à Orsay. Notre projet propose, d'une part, d'étendre nos analyses comparatives à la Mer de Marmara, dans le pourtour Méditerranéen et, d'autre part, d'exploiter nos banques métagénomiques du plancton océanique profond pour reconstruire des génomes complets ou des larges contigs génomiques des microorganismes marins non-cultivés, et plus particulièrement ceux des archées appartenant à des lignées cosmopolites des Crenarchaeota du Groupe I et des Euryarchaeota du Group II, qui sont abondantes dans nos banques. En plus des pistes sur leur métabolisme potentiel, nos résultats fourniront des informations importantes sur la structure génomique, la dynamique génomique et l'évolution chez ces organismes marins mal connus. Des analyses phylogénétiques à l'échelle du génome complet des archées mésophiles marines permettront de clarifier leur position phylogénétique, actuellement discutée, et de tester l'hypothèse selon laquelle ces organismes auraient acquis une grande proportion de leur dotation de gènes à partir d'autres organismes mésophiles lors de leur adaptation à des températures plus froides à partir d'un ancêtre hyperthermophile.

Partenaires CEPHYTEN Université Paris Sud (partenaire coordinateur)
CNRS DR4,
CNRS DR12
Université Miguel Hernandez

Coordinateur Purification Lopez-Garcia
puri.lopez@u-psud.fr

Aide de l'ANR 918 460 euros

Début et durée Janvier 2008 - 48 mois

Référence ANR-08-GENM-024

Titre du projet

METASOIL - Metagenomic discovery and exploitation of the soil microbial community

Résumé

Le sol de par sa communauté microbienne est sans conteste l'écosystème de notre planète qui présente la plus grande biodiversité. La principale limite à l'étude et à l'exploitation de ce très important écosystème demeure le faible niveau de cultivabilité des bactéries et même la difficulté à en extraire l'ADN metagénomique. Des avancées technologiques récentes repoussent toutefois ces limites qui du fait d'une combinaison avec les techniques de séquençage à haut débit vont permettre d'explorer le metagénome du sol dans son intégralité. Cette tâche va être l'objectif des dix prochaines années d'un consortium international, le présent projet en constituant sur 3 années l'étape d'initiation. Les objectifs de ce projet metasoil sont de prendre le leadership du projet international sur le metagénome du sol en créant une ressource génétique d'ordre metagénomique constituée à partir d'un sol unique, bien caractérisé. Les objectifs sont tant d'en identifier taxonomiquement les bactéries que leurs gènes fonctionnels afin d'explorer le potentiel adaptatif fonctionnel de la microflore en tant que ressources génétiques disponibles et éléments génétiques impliqués dans l'adaptation microbienne. Ce projet sera ainsi l'acte fondateur du consortium international en fournissant 5 Gb de séquences et 80 Gb d'ADN du sol cloné dans une banque de clones (objectif qui est à mettre en perspective des 3 GB du génome humain). Ce projet permettra à la recherche publique de constituer une ressource très importante avec des possibilités d'exploitation tant fondamentales qu'appliquées. Le sol ciblé est celui de la station expérimentale terrestre la plus étudiée au niveau mondial (Rothamsted), avec des prélèvements annuels depuis plus de 50 ans et un historique du site remontant à plus de 300 ans. La communauté bactérienne du sol sera extraite et séparée des autres constituants avant que l'ADN n'en soit extrait. Différentes techniques seront appliquées pour récupérer la plus grande diversité bactérienne possible et l'ADN sera séquencé par les méthodes combinées de Sanger et de pyroséquençage mais aussi cloné pour former une banque de plus de 2 millions de clones accessibles après criblages moléculaires et fonctionnels en vue de découvrir de nouveaux gènes et de nouvelles fonctions. Un point fondamental de ce projet concerne le stockage, l'accessibilité et l'utilisation des séquences

metagénomiques obtenues qui ont nécessité de réunir un partenariat pluridisciplinaire avec des expertises en bio-informatique, bio-statistique, dans les différentes méthodes liées à la metagénomique (constitution des banques de gènes, membranes haute densité) et microbiologie environnementale fondamentale. Les partenaires ont déjà collaboré ensemble en de multiples occasions et sont considérés comme des leaders dans leurs domaines respectifs. Avec l'initiation du consortium international ce groupe de partenaires est appelé à jouer dans un futur proche un rôle fondamental en microbiologie environnementale tant fondamentale qu'appliquée.

Partenaires

ECOLE CENTRALE DE LYON (partenaire coordinateur)
CNRS
LIBRAGEN
CEA
Université Louis Pasteur
Université de Pau et des Pays de l'Adour
CNRS
INRA
INRA
INRA

Coordinateur

Timothy Vogel
timothy.vogel@ec-lyon.fr

Aide de l'ANR

2 107 256 euros

Début et durée

Janvier 2008 - 36 mois

Référence

ANR-08-GENM-025

Titre du projet

GENOZYME - Mining genomes for novel biocatalysts

Résumé

Il est maintenant bien établi que les biocatalyseurs sont des alternatives intéressantes à la chimie de synthèse pour l'industrie. Les enzymes sont capables de catalyser des réactions qui ont le grand avantage d'être spécifiques du substrat, strictement énantiométriques et régiospécifiques. De plus, les enzymes opèrent dans des conditions de température et de pression peu élevées, compatibles avec la vie. En conséquence, la biocatalyse réduit les coûts environnementaux de la synthèse chimique classique et participe au développement d'une chimie « durable » qui devient de plus en plus une donnée importante dans la politique de développement des activités industrielles. La synthèse de produits chimiques grâce à des processus de biotechnologie qui utilisent les enzymes est nommée « biotechnologie blanche » et est utilisée par un nombre toujours croissant d'industriels pour la production de composés organiques, d'agents pour l'agriculture, de produits pharmaceutiques, d'additifs pour la nourriture ainsi que des bioplastiques (R. Hatti-Kaul et al. Trends in Biotechnol, 2007, 25, 119). Dans un récent rapport, McKinsey & Company ont montré que les processus biotechnologiques pourraient être utilisés dans la production de 10 à 20% des produits chimiques à l'horizon 2010 (Gupta et Raghava, 2007).

Dans les cellules vivantes, la formation ou la rupture des liaisons C-C sont essentielles dans le catabolisme et l'anabolisme d'une grande variété de carbohydrates et de quelques cétoacides. La nature a donc produit de nombreuses enzymes qui sont capables de casser ou de créer des liaisons C-C (Samland and Sprenger, 2006). Ces enzymes suscitent l'intérêt des chimistes, par exemple, pour la production des monosaccharides qui nécessite de nombreuses étapes de protection et de déprotections en synthèse chimique traditionnelle. Les aldolases et les transcétoalases représentent deux des grandes familles d'enzymes qui font et défont les liaisons C-C. Ce projet a pour but d'élargir les possibles applications industrielles de ces enzymes par la découverte de nouvelles possibilités de réaliser des liaisons C-C. Les techniques de biologie moléculaires permettent de produire un grand nombre de protéines bactériennes qu'elles appartiennent au « monde » des bactéries cultivées ou non. Ces protéines peuvent alors être criblées pour découvrir celles qui catalysent des transformations chimiques intéressantes. Toutes

les étapes nécessaires à une telle analyse seront réalisées sur une plateforme à haut débit existante. Les candidats sélectionnés seront par la suite analysés pour évaluer leur potentiel industriel. A notre connaissance aucun programme n'est en cours pour découvrir de tels nouveaux biocatalyseurs.

Partenaires CEA Genoscope-Institut de Génomique (partenaire coordinateur)
Université Blaise Pascal
ISTHMUS

Coordinateur Marcel SALANOUBAT
salanou@genoscope.cns.fr

Aide de l'ANR 633 667 euros

Début et durée Janvier 2008 - 36 mois

Référence ANR-08-GENM-026

Titre du projet **Genopop-GBS - Génomique des populations de *Streptococcus agalactiae* : diversification des génomes et adaptation à l'hôte.**

Résumé

Streptococcus agalactiae est apparue comme la première cause d'infections néonatales dans la seconde moitié du 20^{ème} siècle. Cette bactérie est également un pathogène animal et a initialement été associée à des pertes économiques dans les élevages bovins laitiers puis dans des fermes piscicoles. Des études épidémiologiques et de génétique des populations basées sur le séquençage de sept gènes de ménage (MLST) ont permis de définir des complexes clonaux (CC) dominants responsables d'infections chez l'homme, d'identifier un clone hypervirulent (CC17) et de mettre en évidence l'importance des échanges génétiques dans l'évolution de cette espèce. Néanmoins, cette technique ne permet pas d'identifier les bases génétiques et fonctionnelles des spécificités des différents complexes clonaux. En réalisant une carte complète des polymorphismes nucléotidiques pour les huit génomes séquencés nous avons mis en évidence l'existence de transferts de grandes régions chromosomiques. Ces échanges ont un rôle important dans l'évolution de l'espèce *S. agalactiae* et ont probablement contribué à l'émergence des différents clones dominants. En parallèle, nous avons montré au laboratoire que ces échanges se font par conjugaison. Ces résultats infirment l'idée, précédemment admise, que les échanges génétiques au sein d'une espèce étaient essentiellement réalisés par transformation ou par transduction et impliquaient de courtes régions. De plus, ils font de *S. agalactiae* un modèle original pour étudier le rôle des transferts génétiques dans l'évolution d'une espèce bactérienne.

Les nouvelles méthodes de séquençage haut débit permettent de passer de la génétique des populations à la génomique des populations. Notre objectif est, grâce au séquençage de 96 génomes de souches de *S. agalactiae* représentatives de la diversité des souches humaines et animales, de reconstruire les échanges génétiques qui ont permis l'émergence des différents clones. Sur la base de cette reconstruction, nous définirons les relations entre les complexes clonaux humains et animaux et chercherons à identifier les bases génomiques de l'adaptation à l'hôte. Enfin, nous compléterons les analyses bioinformatiques par des approches génétiques utilisant les outils développés au laboratoire, afin de préciser les mécanismes de transfert de matériel génétique. Ces différents résultats seront combinés pour proposer un modèle

global d'évolution de l'espèce *S. agalactiae*.

Les résultats attendus de ce travail sont une compréhension de l'évolution de cette bactérie pathogène importante et, notamment, des bases génomiques de la structuration en complexes clonaux. Cette étude devrait, en particulier, permettre de définir si l'apparition récente des infections à *S. agalactiae* correspond à l'émergence de clones particuliers comme le clone hypervirulent CC17. Dans le cadre du développement de vaccins contre *S. agalactiae*, ce travail permettra d'évaluer la couverture vaccinale et l'impact des transferts horizontaux dans l'échappement à un vaccin. De manière générale, ce travail permettra de développer les approches de génomique des populations, une nouvelle discipline à l'interface entre génomique et biologie des populations, basée sur l'utilisation des méthodes de séquençage haut débit, dans l'étude des bactéries pathogènes et de leur épidémiologie.

Partenaires

Institut Pasteur (partenaire coordinateur)
Institut Pasteur, Plateforme Intégration et Analyse Génomique

Coordinateur

Philippe GLASER
pglaser@pasteur.fr

Aide de l'ANR

278 393 euros

Début et durée

Janvier 2008 - 36 mois

Référence

ANR-08-GENM-027

Programme Génomique Microbienne à Grande Echelle

Edition 2008

Titre du projet **CBME - Computational Biology for Metagenomics Experiments**

Résumé

La génomique microbienne à grande échelle est un domaine où des progrès scientifiques importants et rapides sont attendus, avec à la clé des applications de grande portée, notamment dans les domaines de la chimie durable, la bioconversion et la santé. C'est un domaine où les possibilités ouvertes par les technologies de séquençage à haut débit sont particulièrement bien adaptées. La métagénomique bactérienne est devenue accessible et peut se décliner dans différents types d'écosystèmes. Cependant, la technologie ne résout pas tous les problèmes, et il y a des obstacles méthodologiques, dont certains sont parfois sous-estimés dans les projets proposés. En particulier les aspects de calcul bio-informatique et de traitement statistique méritent une attention particulière. Au delà des problèmes de gestion de données posés par leur volume important, la mise en relation de 10^7 à 10^8 fragments courts avec des catalogues de gènes ou de fragments longs comportant entre 10^5 et 10^6 éléments implique l'utilisation d'outils de "pattern matching" très performants bien adaptés. La présence de quelques espèces très abondantes risque de cacher des espèces plus rares mais qui pourraient se révéler plus intéressantes. Cette question est à relier au nombre d'échantillons analysés : plus ils sont nombreux plus on a de chance de trouver des espèces (ou des gènes ou des fragments longs) pertinents...mais plus l'étude coûte cher. Il y a donc clairement un problème de plan d'expérience à optimiser en fonction de ce que l'on sait de la répartition de l'abondance des espèces, mais aussi de la puissance des tests statistiques. De plus, l'analyse statistique elle-même pose des problèmes nouveaux: elle est rendue difficile par 3 éléments : (1) la multiplicité des tests de comparaison entre conditions, qui est de l'ordre du nombre d'éléments du catalogue pose le problème du contrôle des faux positifs. (2) il y a beaucoup plus de variables analysées (les éléments du catalogue) que d'individus (les échantillons) (3) on a des données de comptages, c'est à dire des données discrètes, alors que les analyses classiques sont faites sur des données continues. La combinaison de ces trois difficultés complique l'analyse statistique qui demande de construire des outils adaptés. L'objectif du projet CBME est de fournir des outils bioinformatiques et statistiques adaptés aux expériences de métagénomique et de donner des éléments pour déterminer le nombre d'échantillons nécessaires pour atteindre

avec une probabilité raisonnable les objectifs scientifiques visés par l'étude. C'est un projet général qui ne cible pas une expérience particulière de métagénomique. Cependant ce projet bénéficiera de l'implication de plusieurs de ses membres dans des projets de métagénomique dans différents écosystèmes.

Partenaires INRA, AgroParisTech (partenaire coordinateur)
INRA

Coordinateur Jean Jacques DAUDIN
jean-jacques.daudin@agroparistech.fr

Aide de l'ANR 190 016 euros

Début et durée Janvier 2008 - 36 mois

Référence ANR-08-GENM-028

Titre du projet

BAMBI - Bio-prospection par une Approche Métagénomique de Biocatalyseurs chez des Invertébrés xylophages : analyse des microbiomes d'invertébrés du sol de la forêt malgache

Résumé

Le développement de la biotechnologie blanche pour la transformation des ressources en biomasse constitue un enjeu majeur pour le 21^{ème} siècle. De manière significative, la transformation de la biomasse est aussi un aspect clef de la vie terrestre. La plupart des procédés biogéniques se produisent au niveau du sol, là où la majeure partie de la biomasse se trouve. Les acteurs catalytiques majeurs de ces procédés sont des invertébrés tels que les termites et les vers de terre. Ces insectes sont des fournisseurs d'un arsenal de microorganismes qui catalysent, de manière concertée, l'hydrolyse des polymères tels que la cellulose, les hémicelluloses et les protéines et qui contribuent au biocyclage du C, N, et P. Il n'est donc pas surprenant que la demande actuelle pour des biotechnologies, notamment des enzymes pour la transformation de la biomasse, nous conduit vers l'étude des insectes xylophages. Ces études, qui portent souvent sur l'approche métagénomique et la caractérisation phylogénétique des microbiomes intestinaux, ont déjà permis d'apercevoir l'extrême richesse et le potentiel de cette stratégie pour la découverte de nouvelles enzymes. Cependant, compte tenu de la biodiversité immense, il reste énormément de travail à réaliser à la fois sur les microbiomes intestinaux des insectes et sur les microbiomes de leurs structures biogéniques (ex. termitières, turricules ...) et celui du sol. En effet, les invertébrés du sol n'adoptent pas tous la même stratégie nutritionnelle. Certains dégradent la biomasse en plusieurs étapes et font intervenir plusieurs microbiomes distincts, qui sont phylogénétiquement différents. Ceux-ci se localisent dans l'intestin, dans les structures biogéniques et dans le sol (c'est le cas des termites). D'autres insectes dégradent la biomasse en plusieurs étapes en utilisant un ou plusieurs microbiomes qui se ressemblent fortement. Ainsi, la dégradation de la biomasse dans le sol est un processus extrêmement complexe et assez peu compris. Par conséquent, pour exploiter ces processus à des fins biotechnologiques, il est indispensable de tenir compte de cette complexité et d'analyser d'autres microbiomes en plus de celui de l'intestin.

Dans ce projet nous proposons d'étudier les microbiomes associés à deux insectes (un termite et un vers de terre) qui se nourrissent sur les déchets du sol de la forêt tropicale malgache

dans le but d'isoler de nouvelles enzymes qui auront un fort potentiel pour des applications en bioraffinerie et, plus généralement, pour les procédés de biotechnologies blanches. Dans la première phase de l'étude, nous réaliserons une campagne de collecte des échantillons biologiques au Madagascar, puis nous préparerons plusieurs libraires métagénomiques. En particulier, pendant les 12 mois concernés par ce financement (starting grant), nous démontrons la pertinence de la méthode SIP (incorporation d'isotopes stables) pour l'enrichissement dans les banques métagénomiques d'ADN provenant de microorganismes fonctionnels, impliqués dans la dégradation de la matière lignocellulosique. Cette démonstration sera réalisée avec des vers de terre élevés sur un milieu enrichi en paille de blé marquée au ^{13}C .

Partenaires INSTITUT NATIONAL DES SCIENCES APPLIQUEES (partenaire coordinateur)
IRD
LIBRAGEN
CEA Genoscope-Institut de Génomique

Coordinateur Michael O'DONOHUE
michael.odonohue@insa-toulouse.fr

Aide de l'ANR 141 336 euros

Début et durée Janvier 2008 – 12 mois

Référence ANR-08-GENM-029