Les Cahiers Prospectives

Prospective génomique environnementale

UNE PROSPECTIVE DE L'INSTITUT ECOLOGIE ET ENVIRONNEMENT N°5 - OCT 2013

cnrs
dépasser les frontières
www.cnrs.fr


11
10
100
50
0
9
100
50
0
8
100
50
0
7
100
50
0
6
100
50

SOMMAIRE

Prospective

régionale environnementale

Coordination :
Denis Faure, Dominique Joly,
Sylvie Salamitou

AVANT-PROPOS

Avec le développement de nouvelles générations de séquençage (NGS), aussi appelées nouvelles technologies de séquençage (NTS), ce début de 21e siècle est marqué par la levée d'un verrou technologique majeur en biologie. Parce qu'elles augmentent significativement le niveau de production des séquences nucléotidiques par rapport au séquençage classique (Sanger), les NGS modifient de manière irréversible les recherches en écologie, évolution et sciences de l'environnement. Cette révolution technologique permet l'émergence d'un champ disciplinaire jusque-là inaccessible, la génomique* environnementale. Elle appréhende la structure et le fonctionnement des différentes composantes des génomes et métagénomes (ADN et ARN) de l'ensemble des organismes ; elle permet, en association avec d'autres approches à haut débit dites « omiques et métaomiques », une compréhension intégrée des populations, communautés et écosystèmes présents et passés. La génomique environnementale est considérée ici dans son acceptation la plus large intégrant les approches quantitative et qualitative de la complexité du vivant et de ses interactions aux différentes échelles de temps et d'espace.

Pour répondre aux défis technologiques, méthodologiques et scientifiques que représentent la « domestication » des NGS par et pour les laboratoires dans le domaine de la génomique environnementale, l'INEE du CNRS a engagé une série d'actions afin 1) de soutenir leur compétitivité nationale et internationale, 2) de développer leur expertise dans les domaines de la biodiversité (IPBES) et de l'évaluation des impacts environnementaux (REACH), et 3) de promouvoir des propositions d'applications potentielles comme l'ingénierie écologique, la modélisation prédictive ou le suivi de la qualité des écosystèmes et agrosystèmes. Dès 2010, un premier séminaire prospectif a été organisé à Chizé en association avec l'INRA, suivi en 2011 par la création du Réseau Thématique Pluridisciplinaire en Génomique Environnementale (RTP-GE). L'objectif de ce RTP était d'identifier et de mobiliser les communautés scientifiques de l'INEE et de ses partenaires concernés. Il s'agissait 1) de promouvoir un dialogue et des synergies entre différentes disciplines telles que biologie, écologie, évolution, paléobiologie et taxinomie, mais aussi chimie, bioinformatique, mathématique, biogéochimie, sociologie ou anthropologie,… 2) de développer et soutenir des échanges entre les communautés scientifiques étudiant les différents domaines du Vivant dont les Eucaryotes, Procaryotes et Virus, et 3) d'identifier les enjeux et les verrous méthodologiques des NGS posés aux différents champs thématiques afin de dégager des priorités conceptuelles et expérimentales susceptibles de déboucher sur des propositions de programmation de recherche.

Ce cahier de prospective en génomique environnementale reflète l'effervescence scientifique portée par les NGS au cœur des thématiques de l'INEE du CNRS. Il est l'aboutissement de la réflexion du RTP-GE à l'issue de 3 années d'activité. La rapidité d'évolution de ces technologies et les opportunités de science qu'elles génèrent impliquent de nouveaux paradigmes de travail. Ils changent radicalement les stratégies expérimentales d'étude de l'évolution, de la biodiversité et du fonctionnement des écosystèmes et systèmes biologiques complexes. A ce titre, ce document propose de nouvelles perspectives d'interactions et de partages de connaissances entre les acteurs des différents champs disciplinaires.

L'INEE, dans sa mission de faire émerger les sciences de l'environnement en tant que champ scientifique intégré a une position clé pour accompagner cette mutation. Le RTP-GE a identifié quatre domaines majeurs de recherche liés aux NGS, et qui sont déclinés dans l'ensemble de ses actions et dans ce document : 1) l'analyse de la biodiversité et des associations inter-organismes aux échelles individus, populations et communautés dans une approche intégrée des interactomes ; 2) l'identification des processus évolutifs et adaptatifs de l'ensemble des molécules du vivant impliquées, 3) la prise en compte des sociétés humaines dans leurs composantes économiques et culturelles pour l'étude de la dynamique des écosystèmes, y compris dans une approche rétrospective, et 4) l'anticipation de l'évolution des écosystèmes, y compris ceux d'origine anthropique, à l'aide de modèles prédictifs élaborés à partir de la compréhension de leurs propriétés aux échelles locales, régionales et planétaires.

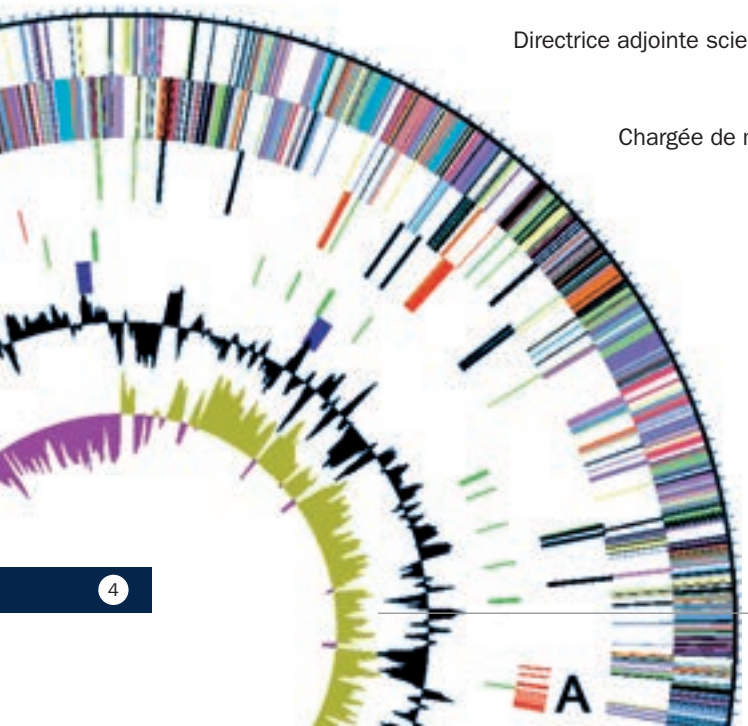
Les différentes actions entreprises par le RTP-GE à partir de ces quatre axes de recherche ont suscité un fort enthousiasme de la part d'une nouvelle communauté scientifique fédérée autour de la génomique environnementale, engouement qui témoigne d'un besoin manifeste de partage de connaissances et de savoir-faire sur les NGS. Le colloque de Lyon en 2011 a permis de consolider et de rendre visible un élan national dans le domaine de la génomique environnementale associant les laboratoires du CNRS et de ses partenaires. Un soutien direct aux laboratoires a été apporté sous la forme de deux appels à projets en génomique environnementale (APEGE 2012 et 2013), tandis qu'une école thématique (2012) a renforcé les interactions entre jeunes chercheurs et chercheurs confirmés et entre laboratoires relevant de champs disciplinaires différents. En 2013, cette dynamique se poursuit par un colloque en novembre à Rennes et des perspectives de valorisation des travaux de recherche dans des numéros spéciaux de revues à comité de lecture. D'autres actions sont d'ores et déjà programmées, notamment en 2014 dans le cadre de l'EMBnet en mai à Lyon, et du colloque conjoint entre la British Ecological Society et la Société Française d'Ecologie en décembre à Lille. `

Loin d'être un aboutissement, ce cahier de prospective, issu d'un travail collaboratif et pluridisciplinaire, a vocation à soutenir et amplifier des nouvelles dynamiques de recherche aux plans national et international.

Stéphanie Thiébault
Directrice de l'Institut écologie et environnement du CNRS

Martine Hossaert
Directrice adjointe scientifique de l'Institut écologie et environnement du CNRS

Dominique Joly
Chargée de mission de l'Institut écologie et environnement du CNRS



I INTRODUCTION

La génomique environnementale regroupe l'ensemble des connaissances acquises sur les organismes et écosystèmes présents et passés par l'analyse de la séquence des gènes, génomes, métagénomes, transcrits, transcriptomes et métatranscriptomes. Ainsi, en combinaison avec d'autres technologies et observations, la génomique environnementale informe sur la taxinomie* et la diversité des organismes actuels et fossiles (individus, populations, communautés), leur phylogénie* et évolution, leurs potentialités et capacités d'adaptation et d'acclimatation, leur biologie, leurs traits fonctionnels, et leurs interactions avec l'environnement dans ses dimensions biotique et abiotique.

Les **nouvelles générations (technologies) de séquençage** (NGS/NTS) de l'ADN permettent un niveau de production de données en (méta) génomique structurelle et fonctionnelle encore inimaginable il y a quelques années. Les NGS modifient profondément et durablement les stratégies expérimentales en évolution, biodiversité et écologie des organismes et écosystèmes présents et passés, mais aussi le **regard scientifique** porté sur les populations humaines, ainsi que la représentation du Vivant. Les NGS concernent **tous les domaines du vivant**, Archées, Eucaryotes et Bactéries ainsi que les Virus, et permettent d'accéder à des groupes taxinomiques encore inconnus. Cette révolution technologique crée de nouvelles opportunités scientifiques, renouvelle les itinéraires techniques et méthodologiques, et appelle de nouveaux besoins en formations initiales et permanentes des différentes communautés scientifiques relevant du CNRS-INEE et ses partenaires nationaux et internationaux.

Les **premières technologies de séquençage** ont émergé dans les années 1970 et ont été utilisées pour le séquençage des premiers génomes et métagénomes dès la fin des années 1990. Elles ont contribué à l'émergence de la génétique et génomique structurales et fonctionnelles, la phylogénie et la taxinomie moléculaires*, ainsi que le développement de marqueurs moléculaires de taxons ou code-barres. Les nouvelles générations (technologies) de séquençage (NGS/NTS) émergent au milieu des années 2000 avec la commercialisation de technologies dites de seconde génération : Solid par Life-technologies, Solexa par Illumina* et pyroséquençage 454* par Roche sont les plus utili-

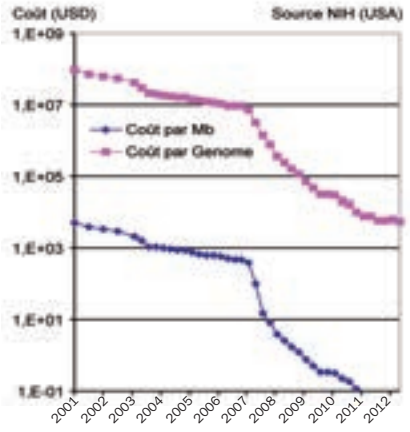


Figure 1A. Coût (USD) de séquençage par Mb et par génome

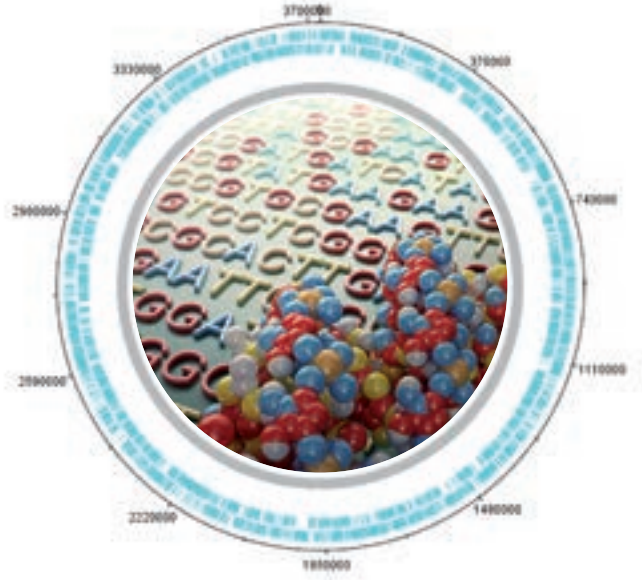
sées. L'émergence de ces nouvelles technologies assure la production d'un **nombre très important de séquences** ADN et simultanément une **baisse drastique de leur coût** (Figure 1A). La conjonction de ces deux caractéristiques a assuré une **propagation rapide des NGS** dans les laboratoires de recherche académique et du secteur privé liés aux sciences de la vie et de l'environnement.

Les NGS offrent aux sciences de l'environnement, de l'évolution, de l'écologie et de la biodiversité de **nouvelles opportunités scientifiques**, mais en contrepartie exigent **l'acquisition et la domestication de nouveaux outils** d'analyse, de manipulation, de stockage et d'échange des données issues des NGS. L'acquisition de ces savoir-faire s'inscrit dans un contexte de forte compétitivité nationale et internationale pour la production de connaissances, mais aussi d'enjeux sociétaux comme la biodiversité ou l'évaluation des impacts environnementaux des activités humaines. L'utilisation des NGS constitue également, pour des chercheurs et laboratoires questionnant différemment le Vivant et l'Environnement, une opportunité de **partage et d'échanges** de données, et de **fédération d'expertises** autour d'objets communs. Ce document de prospective en génomique environnementale est le fruit d'une réflexion et rédaction collectives de la communauté scientifique animée depuis 2011 par le Réseau Thématique Pluridisciplinaire en Génomique Environnementale (RTP-GE). Les actions du RTP-GE ainsi que la communauté scientifique concernée sont présentées dans les chapitres II et III. Les chapitres IV et XIII donnent une vue d'ensemble des **enjeux, verrous et propositions** en génomique environnementale, tandis que les autres chapitres précisent cette réflexion dans différents champs méthodologiques et scientifiques que sont l'accès à la production des données NGS (V), la qualité des données NGS (VI), la structure et la dynamique de la biodiversité (VII) et sa caractérisation (VIII), l'étude de l'évolution adaptative des gènes et génomes (IX), l'écologie fonctionnelle* des populations (X) et des communautés (XI), et la modélisation des écosystèmes (XII).

Nous remercions chaleureusement chacun des contributeurs de ce document de prospective en génomique environnementale, et plus particulièrement les membres du comité scientifique du RTP Génomique Environnementale, qui par leurs actions contribuent à faire de ce champ de recherche un espace de partage de savoirs et savoir-faire et de dialogues au sein des laboratoires et partenaires de l'INEE. Enfin, nous voudrions remercier Isabelle Poulain (INEE-CNRS) et Sylvie Apruzzèse-Serazin (Laboratoire Evolution, Génomes et Spéciation, CNRS, Gif-sur-Yvette) pour leur assistance régulière dans le fonctionnement du RTP, Samuel Mondy (Institut des Sciences du Végétal, CNRS, Gif-sur-Yvette) pour son aide dans l'exploitation des données du RTP-GE, ainsi que Conceiçao Silva et l'équipe du pôle communication de l'INEE-CNRS sans qui cet ouvrage n'aurait pu exister.

II ACTIONS DU RÉSEAU THÉMATIQUE PLURIDISCIPLINAIRE GÉNOMIQUE ENVIRONNEMENTALE

Coordinateurs : Denis Faure et Dominique Joly



Le CNRS-INEE a créé le Réseau Thématique Pluridisciplinaire Génomique Environnementale (RTP-GE 2011-2013) afin d'identifier la communauté scientifique active en génomique environnementale, d'améliorer sa visibilité et de soutenir ses propositions dans une perspective de programmation de la recherche, mais aussi de favoriser les échanges de savoirs et savoir-faire au sein de cette communauté en émergence. Au-delà de la mise en valeur d'actions disciplinaires relevant de l'écologie fonctionnelle, l'écologie des communautés, la biologie évolutive, la paléobiologie, la taxinomie moléculaire et la biodiversité, le RTP-GE a révélé une demande forte d'association d'expertises différentes sur un même objet biologique ou écosystème afin d'exploiter au mieux la richesse des données accessibles grâce aux NGS et autres approches omiques.

Le RTP-GE est animé par un bureau composé de Dominique Joly (CNRS, Coordinatrice du RTP-GE auprès de l'INEE), Denis Faure (CNRS, Responsable du RTP-GE), Pierre Peyret (Univ Clermont-Ferrand 1), François Pompanon (CNRS, Univ Grenoble 1) et Pascal Simonet (CNRS, Univ Lyon 1, Ecole Centrale), et soutenu par les avis d'un Comité Scientifique qui associe les membres du bureau et les personnalités suivantes : Colomban de Vargas (CNRS, Univ Paris 6, Roscoff), Michael Dubow (CNRS, Univ Paris-Sud), Mathieu Joron (CNRS, MNHN Paris), Line Le Gall (CNRS, MNHN Paris), Denis Le Paslier (CNRS, Génoscope Evry), Dominique Mouchiroud (CNRS, Univ Lyon 1), Francis Martin (INRA, Nancy), Eric Pelletier (CEA Génoscope, Evry), Guy Perrière (CNRS, Univ Lyon 1), Jean-Christophe Simon (INRA, Rennes), Pierre Taberlet (CNRS, Univ Grenoble 1),

Philippe Vandenkoornhuyse (CNRS, Univ Rennes 1), et Xavier Vekemans (CNRS, Univ Lille 1). Le bureau et le comité scientifique ont permis d'associer des expertises couvrant un large spectre de champs thématiques, organismes, écosystèmes, ainsi que les dernières avancées en technologies de séquençage et bioinformatique.

Les principales **actions nationales** du RTP-GE (Figure 2A) ont été l'organisation de deux colloques nationaux à Lyon et Rennes, une école thématique à Aussois, et deux appels à propositions en génomique environnementale (APEGE). Ces actions ont permis à la fois de recenser les laboratoires actifs en génomique environnementale, qui relèvent ou sont partenaires de l'INEE (voir chapitre III), mais aussi d'améliorer leur visibilité, stimuler les échanges de connaissances entre chercheurs et ingénieurs confirmés et en formation (doctorants et post-doctorants), et de recueillir les attentes des laboratoires et chercheurs dans ce domaine en émergence. Le RTP-GE a été fortement impliqué dans les Prospectives INEE d'Avignon en 2012, et notamment dans les ateliers en Ecologie prédictive et changements planétaires et en Génomique, et est à l'origine de la rédaction du présent document de prospective en génomique environnementale.

Les principales **actions internationales** du RTP-GE s'inscrivent au-delà de 2013 (Figure 2A), avec l'organisation en 2014 d'un symposium

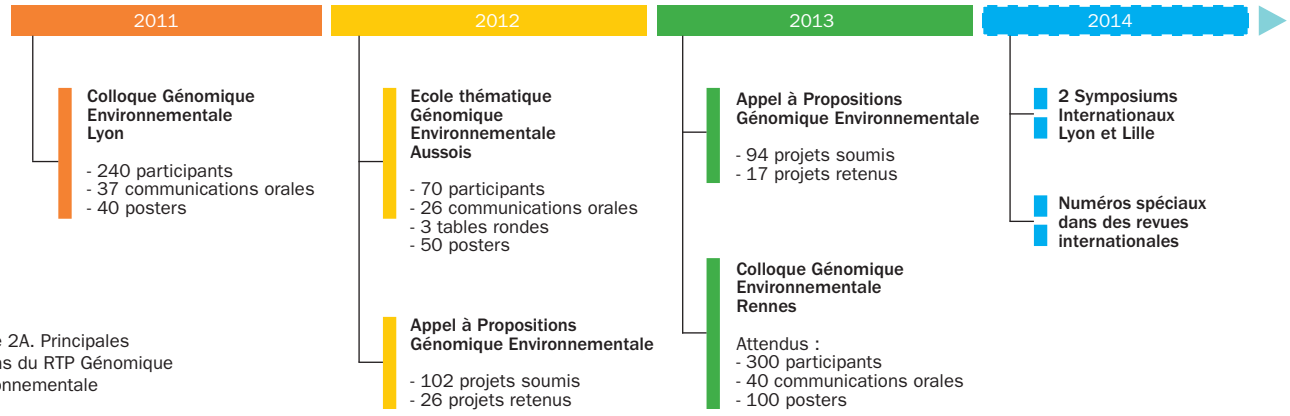


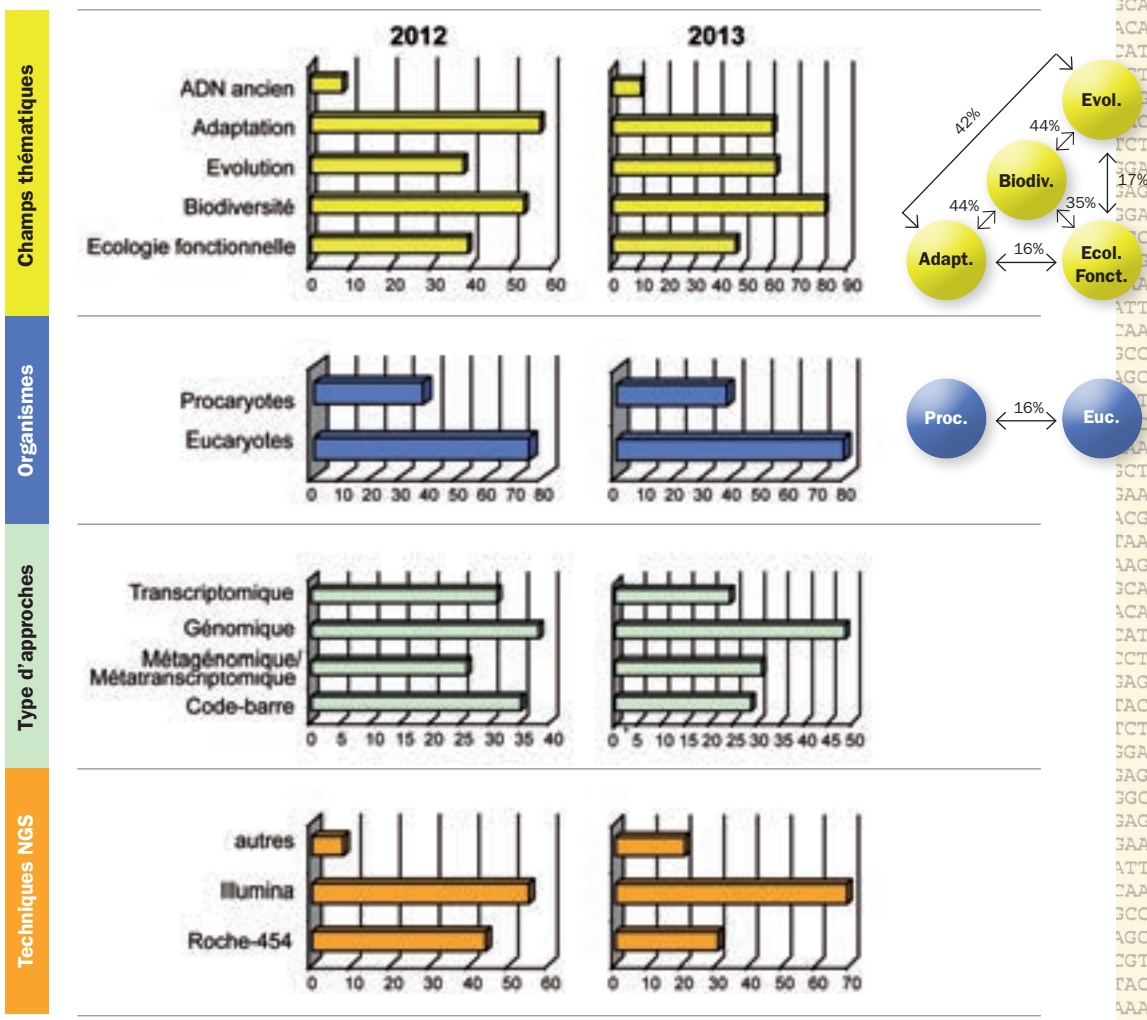
Figure 2A. Principales actions du RTP Génomique Environnementale

associé à la réunion annuelle EMBnet (Réseau Européen de Bioinformatique) à Lyon et d'un second associé au « British Ecological Society and Société Française d'Ecologie Joint Meeting » à Lille, mais aussi avec la proposition de numéros spéciaux « Environmental Genomics » (2014) dans une ou deux revue(s) internationale(s) à comité de lecture que sont Heredity (Nature Publishing Group) et Genetica (Springer).

Une mobilisation et une pression fortes de la communauté GE ont été révélées grâce aux appels à projets APEGE 2012 et 2013. Ceux-ci ont été portés par un financement du CNRS-INEE afin de soutenir des projets annuels mobilisant les approches de séquençage à haut débit (génomique/transcriptomique/métagénomique*/métatranscriptomique/barcode*) dans les domaines de l'écologie fonctionnelle et des interactions entre organismes, de l'écologie et de la biodiversité des communautés, de la dynamique et de l'évolution des interactions passées et présentes, de paléogénomique et ADN ancien et également de questions liées aux outils d'analyse des données NGS (bioinformatique/biostatistique/mathématique). Tous les objets biologiques ont été éligibles (Virus, Archées, Bactéries et Eucaryotes) ainsi que tous les environnements (marins, littoraux, continentaux, tempérés, tropicaux, polaires, etc.). Les projets devaient avoir pour objectif de répondre à une question biologique clairement identifiée. Le détail des appels à propositions est disponible sur le site du RTP-GE.

La pression sur ces deux appels à projets a été forte avec 102 et 94 projets déposés respectivement en 2012 et 2013 pour une pression financière de 1 800 et 1 100 k€. La différence entre les pressions financières de 2012 et 2013 est expliquée par la limite supérieure de la demande fixée dans l'appel d'offres à

Figure 2B.
Analyse des projets APEGE



20 k€/projet en 2012 et 13 k€/projet en 2013. Le comité scientifique du RTP-GE a joué un rôle clé dans la sélection des projets déposés. La liste complète des projets acceptés est disponible sur le site du RTP-GE. Les projets retenus couvrent l'ensemble des champs thématiques de ces appels à projets.

Un questionnaire à choix multiple rempli par les porteurs de projets APEGE a permis une **analyse de la typologie des projets déposés** pour quatre facteurs (Figure 2B) : le champ thématique (Ecologie fonctionnelle, Biodiversité, Evolution, Adaptation, ADN ancien), les organismes étudiés (Eucaryotes dont microorganismes*, Eucaryotes, et Procaryotes dont Bactéries, Achées et Virus), les méthodologies proposées (Transcriptomique, Génomique, Métagénomique/ Métatranscriptomique, et Code-barre), et les techniques NGS mobilisées (454-Roche, Illumina, autres). L'analyse de ces données réalisée par Samuel Mondy (CNRS, Gif-sur-Yvette) montre que : 1) les deux APEGE ont atteint la communauté scientifique ciblée dans l'expression de ces différents champs thématiques, considérant que la thématique ADN ancien représente moins de 10% des projets déposés ; 2) les questions abordées relèvent fréquemment de plusieurs champs thématiques comme le montre l'analyse de cooccurrence réalisées avec les données de 2013 ; 3) les objets étudiés sont majoritairement des eucaryotes (deux tiers) avec néanmoins 16% des projets de 2013 concernant à la fois des eucaryotes et procaryotes ou virus ; 4) les méthodologies reflètent la diversité des questions biologiques avec une poussée de la génomique en 2013 qui révèlent à ce jour une utilisation première des NGS pour l'étude de populations ou d'individus ; 5) enfin une poussée remarquée de l'approche Illumina au détriment du 454-Roche en 2013. Ce glissement montre l'évolution rapide de l'offre technologique et de ses usages. En résumé, les APEGE 2012 et 2013 ont révélé une **mobilisation forte et créative de la communauté GE** dans la diversité des champs thématiques et objets étudiés avec l'utilisation principalement de deux techniques NGS que sont 454-Roche et Illumina sur des plateformes publiques ou privées.

Les colloques et école thématique ont révélé une **demande d'échange d'expertises et de savoir-faire ainsi que de formation des personnels**. Si le colloque génomique environnementale de Rennes programmé en novembre 2013 est encore en cours d'organisation lors de la préparation de ce document, celui de Lyon en 2011 ainsi que l'Ecole Thématique Expert en Génomique Environnementale (ETEGE) à Aussois en 2012, ont permis de réunir de nombreux acteurs de la communauté scientifique concernée. Ces deux événements ont été possibles grâce au soutien du CNRS principalement et de l'INRA. Outre le recensement des laboratoires impliqués dans ces événements, ces rencontres ont permis des échanges et partages d'expériences importants dans le contexte de la domestication des outils et procédures NGS (de l'in vivo à l'in silico) nécessaires à leur valorisation. Ces rencontres ont révélé trois points forts qui caractérisent cette communauté en cours de constitution :
- une demande d'aide à **l'accès aux NGS** (capacités de production et d'analyse des données) ;
- une demande de la communauté GE pour le partage des données et des savoir-faire lors de colloques et de formations spécialisées ;
- une volonté forte de **dépasser les frontières disciplinaires** pour mieux utiliser les données NGS (mathématiques, bioinformatique, écologie, évolution, biologie fonctionnelle et structurale, intégration complexe et modélisation etc…) afin de faire face à une rude compétition internationale.

En conclusion, le **RTP-GE** s'est avéré un outil **mobilisateur** puissant de la communauté scientifique en génomique environnementale dont l'**ampleur et la vitalité** ont dépassé les attentes, et dont l'accompagnement mérite toutes les attentions futures dans un contexte scientifique et technique très **dynamique et compétitif**.

SITE INTERNET

RTP-GE : http://www.cnrs.fr/inee/recherche/actionsincitatives-RTP-Genoenvironnementale.htm

III COMMUNAUTÉ SCIENTIFIQUE ACTIVE EN GÉNOMIQUE ENVIRONNEMENTALE

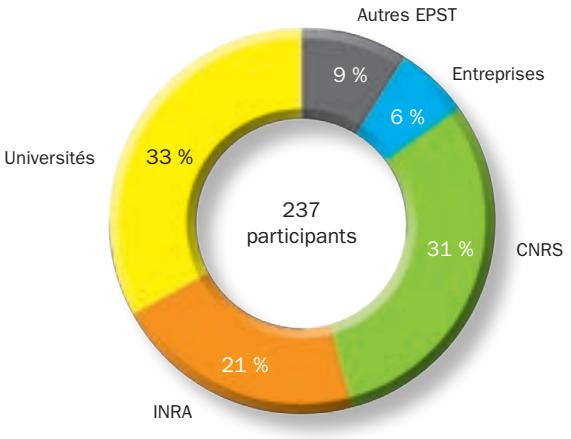
Coordinateurs : Dominique Joly et Denis Faure
Contributeurs : Catherine Boyen et Pascal Simonet

Répartie sur l'ensemble du territoire national la communauté scientifique concernée par la génomique environnementale est issue d'une douzaine d'organismes de recherche. Les principaux laboratoires impliqués sont sans surprise ceux qui bénéficient d'infrastructures nationales ou régionales mettant en oeuvre les NGS. 80% des laboratoires du CNRS-INEE sont concernés par cette thématique et participent activement aux différentes actions menées par le RTP-GE.

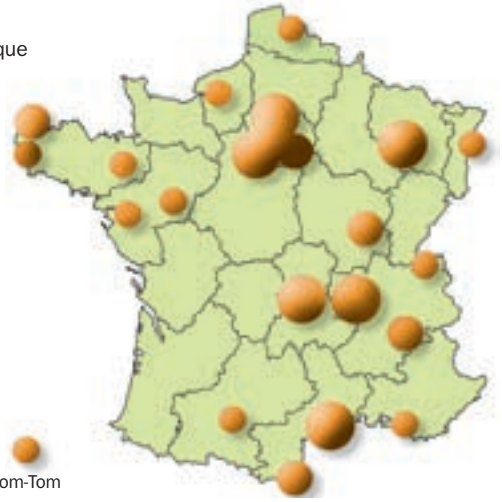
Répartition géographique

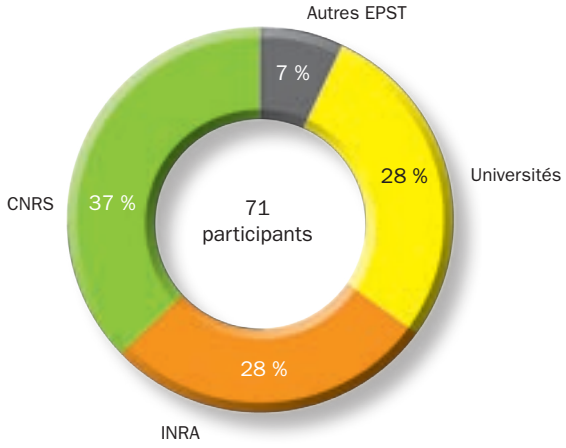
La distribution géographique des chercheurs et équipes qui ont répondu aux différentes actions du RTP-GE est largement répartie sur le territoire national (Figure 3A). Cette distribution est plus ou moins concentrée autour des plateformes de séquençage nationales et régionales regroupées dans l'infrastructure nationale France Génomique (Focus 3-1). Une plus faible représentation des acteurs locaux est observée en régions Centre, Poitou-Charentes et Aquitaine, notamment en ce qui concerne les participations au colloque de Lyon en 2011 et à l'école thématique d'Aussois en 2012. Ces régions sont celles qui ne disposent pas à l'heure actuelle d'infrastructures liées à France Génomique. Cependant, les laboratoires localisés dans ces régions ont participé aux appels à projets APEGE 2012 et 2013 ; ceci témoigne de recherches potentiellement structurantes pour le développement des NGS en génomique environnementale.

Colloque
Lyon 2011

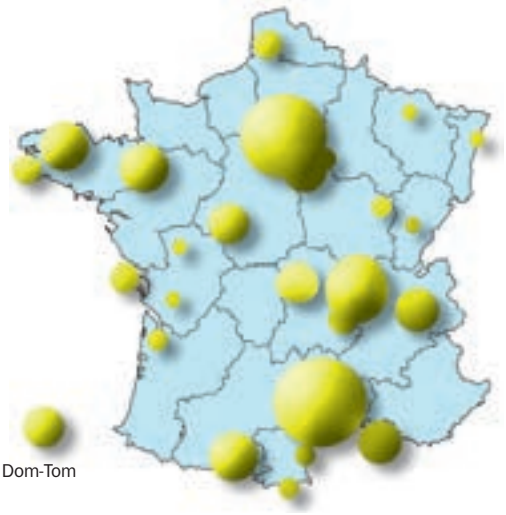


Ecole thématique
Aussois 2012





APEGEs
2012 & 2013



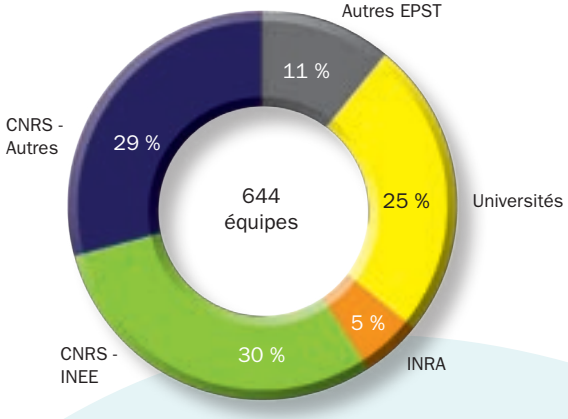


Figure 3A. Répartition géographique (par ville) et tutelle (principale et secondaire) de rattachement des personnels ayant participé aux actions menées par le RTP-GE (haut : colloque de Lyon 2011 ; milieu : école thématique d'Aussois en 2012 ; bas : APEGE 2012 et 2013). Le diamètre des cercles est proportionnel à la participation pour chacune des actions. Pour les projets APEGE, ont été comptabilisées toutes les équipes partenaires impliquées dans les dossiers.

Répartition structurelle

La communauté scientifique qui s'est manifestée lors des différentes actions organisées par le RTP-GE est issue, pour les trois quart, du **CNRS, de l'INRA et des Universités** (représentés en parts plus ou moins égales selon les actions). Le quart restant est représenté par d'autres organismes de recherche, MNHN, IRD, CIRAD, CEA, Génoscope, INSERM, INRIA, IRSTEA, IFREMER. Au CNRS, les principaux instituts partenaires des projets sont, par ordre décroissant d'importance en terme de nombre d'équipes associées, l'INSB (17%), l'INSU (10%), puis de façon plus anecdotique l'INSIS et l'INSMI (1% chacun). La prépondérance des deux premiers instituts n'est pas sans pertinence au regard des thématiques scientifiques portées par chacun d'eux. Le succès des APEGE 2012 et 2013 dépasse l'objectif initial d'amorçage et montre des besoins importants et croissants dans ce domaine.
Concernant l'INEE, **80% des laboratoires de l'institut** (52 sur 66 hors FR, GDR, UMS, UMI) ont participé à l'une ou l'autre des manifestations organisées par le RTP-GE et ont été notamment porteurs ou partenaires dans au moins un projet APEGE. La médiane du nombre de projets déposés par les laboratoires est de 2 avec un maximum de 7. Le nombre de projets est quasiment stable entre 2012 et 2013 (102 et 94 respectivement), ce qui témoigne, là encore, du **dynamisme important** de la communauté dans le domaine et d'une demande qui ne faiblit pas. Il est d'ailleurs à noter que de nombreux projets ont été renouvelés tant par rapport aux technologies NGS envisagées que par rapport aux questions scientifiques abordées, signes d'une **évolution rapide des concepts et des applications**. Le nombre d'équipes partenaires pour chaque projet varie fortement avec une moyenne de 2 laboratoires impliqués et 2 équipes participantes, le nombre de partenaires pouvant aller respectivement jusqu'à 4 et 5.

Complémentarité des communautés scientifiques

Depuis le séminaire de Chizé de 2010, il est très rapidement apparu que les différentes voies d'appréhension de la génomique environnementale (écologie, biodiversité, évolution) étaient souvent liées et que ces divers aspects étaient traités de façon très complémentaire. Certaines communautés de chercheurs, comme celle de l'écologie microbienne, étaient déjà bien identifiées comme partie prenante de la génomique environnementale, avec une expertise bien avancée dans l'utilisation des NGS. Les chercheurs concernés par l'écologie des communautés de microorganismes ont comme **priorité l'inventaire des organismes et des communautés** présentes et leurs **principales fonctionnalités**. D'autres communautés, notamment celles liées à la biodiversité et à l'écologie évolutive, étaient plus distantes par rapport à ce champ d'application, privilégiant l'utilisation d'outils et de concepts développés durant les dernières décennies, notamment en génétique des populations, et un usage plus restrictif des NGS. Pour cette dernière communauté, les NGS apportent des **moyens nouveaux d'approfondir des questions déjà anciennes**, sur des organismes bien identifiés. Le rapprochement de ces communautés a constitué un premier défi qu'a relevé le RTP-GE, qui s'est concrétisé, lors du colloque de Lyon, par un partenariat avec le réseau ECOMIC (ECOlogie MICrobienne) porté par l'INRA. Poussé par des usages communs avec des savoirs et des savoir-faire spécifiques, ce rapprochement s'est concrétisé lors des derniers projets APEGE où des projets collaboratifs issus de nouveaux partenariats ont vu le jour.

Une communauté scientifique très concernée par l'avènement des NGS est celle du consortium « **Bibliothèque du Vivant** » qui inclut 35 unités et 200 chercheurs. Cette structure partenariale entre le CNRS, l'INRA et le MNHN a pour objectif de définir un **cadre d'identification spécifique** sur différents modèles (principalement eucaryotes, mais avec une gamme d'organismes très large) afin de documenter la diversité génétique intra-spécifique, d'explorer les limites d'espèces, de caractériser ces dernières moléculairement avec une approche de type barcode, d'établir des phylogénies molécu-

FOCUS 3-1 : FRANCE GÉNOMIQUE

Cette infrastructure intégrée, de dimension européenne, lauréate des investissements d'avenir 2010, a été déployée à l'échelon national pour intégrer les capacités d'analyse des génomes de 2ème et 3ème générations et les traitements bioinformatiques des données NGS générées. Elle a pour ambition de fournir, dans le domaine de la génomique, un éventail de services permettant de renforcer la compétitivité des communautés nationales d'utilisateurs, chercheurs publics et industriels, et permettra la réalisation de grands projets internationaux. Il s'agit d'une infrastructure distribuée s'appuyant sur plusieurs centres aux compétences complémentaires ayant chacune leur(s) expertise(s) et technologie(s) spécifique(s) ainsi que les outils de bioinformatique ad hoc. La gouvernance est intégrée, avec une coordination des services et un accès unique à l'échelon national qui permet de couvrir tous les domaines des sciences du vivant (biodiversité, génomique médicale, génomique animale, génomique végétale, etc.). Cependant, une ouverture large est fortement attendue sur les organismes non modèles et les espèces non cultivées.

Le projet France-Génomique regroupe le Génoscope et le Centre National de Génotypage (CNG), les 7 plateformes régionales, les noeuds du réseau national de bioinformatique (APLIBIO en Ile-de-France et ReNaBi - Réseau National des plateformes de Bioinformatique), qui seront partenaires du prochain IFB (Institut Français de Bio-Informatique), infrastructure qui devrait voir le jour prochainement, et à court terme, le Très Grand Centre de Calcul (TGCC) du CEA à Bruyères-Le-Chatel, où seront implantés des équipements dédiés de stockage (initialement 5 Po) et de traitement bioinformatique (initialement 3000 coeurs) parfaitement sécurisés et évolutifs.

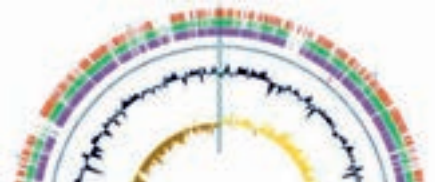
Ces infrastructures s'inscrivent dans les grands axes prioritaires de la recherche française recensés dans la Stratégie Nationale de Recherche et d'Innovation (SNRI), les feuilles de route française et européenne des Très Grandes infrastructures de Recherche (ESFRI), et les stratégies des Alliances concernées. L'ambition du programme France Génomique n'est clairement pas de rivaliser avec des structures étrangères comme le BJI chinois (qui a ouvert son centre européen à Copenhague en 2012) ou le JGI américain, mais d'apporter à la communauté française et ses partenaires internationaux des moyens et une expertise mutualisés.

Au total l'équipement prévu correspondra aux capacités actuelles du BGI (soit une trentaine de séquenceurs). L'effort sera porté non pas sur la capacité de production mais sur une expertise à 3 niveaux : 1) une veille technologique avec la mise au point de pipelines bioinformatiques permettant d'apporter à la communauté scientifique une expertise de haut niveau ; 2) un mode d'interactions avec la communauté via des collaborations sur projets ; 3) une capacité de financement de très grands projets (dont une partie sera dédiée à la génomique environnementale). Des financements de projets de moindre envergure soumis au fil de l'eau sont également prévus et leur mise en œuvre sera effectuée sur l'une des 7 plateformes régionales en fonction de ses spécificités techniques et scientifiques.

Figure 3B. Implantation nationale de l'infrastructure France Génomique. Localisation sur le territoire des plateformes nationales et régionales intégrées dans l'infrastructure.

laires entre espèces, et de donner les moyens de réaliser des inventaires moléculaires d'espèces dans un milieu donné. Il s'agit aussi de créer le lien indispensable entre la « bibliothèque moléculaire » de la biodiversité et les collections nationales de spécimens, ainsi que leurs bases de données associées. En outre, Bibliothèque du Vivant permet aux chercheurs de participer aux grands programmes internationaux comme « BarCode of Life » (voir chapitre VI). En 2011, ce consortium a soutenu 98 projets de séquençage pour une production d'environ 395.000 séquences de type Sanger au Génoscope. Aujourd'hui, l'émergence des NGS offre de nouvelles possibilités d'analyses multilocus et/ou génomiques des organismes vivants qui vont se substituer aux approches monolocus utilisant la technologie de type Sanger. Un rapprochement naturel entre la communauté Bibliothèque du Vivant et celle du RTP-GE s'est mis en place, afin de favoriser les échanges de pratiques, savoirs et savoir-faire et de promouvoir l'utilisation des NGS à toutes les échelles taxinomiques.

Les différents rapprochements entre communautés décrits précédemment, et auxquels a fortement participé le RTP-GE, montrent à quel point le **défi pour les prochaines années** sera pour la génomique environnementale de **renforcer l'interdisciplinarité** avec les bioinformaticiens, mathématiciens et statisticiens pour développer les **méthodes conceptuelles et théoriques** d'analyse et d'exploitation des données NGS, ainsi que des prédictions théoriques sur le devenir et les fonctions des organismes et des écosystèmes dans un environnement changeant. Des actions conjointes avec le réseau StatOmique pourront à ce titre être envisagées. Une meilleure intégration des niveaux d'analyse, de la population à la communauté, sera également à renforcer afin de prendre en compte l'environnement biologique.



Consortia nationaux et internationaux

Outre l'infrastructure nationale France Génomique qui rassemble l'équipement technologique au travers des plateformes et des réseaux de plateformes au niveau national, un certain nombre de réseaux et de consortia se constituent autour des questions de génomique environnementale, notamment au travers de l'utilisation des NGS, soit pour des questions méthodologiques (réseau bioinformatique, statistiques, etc), soit pour des questions d'objets de recherche ou de problématique environnementale, y compris vis-à-vis des anthroposystèmes ou des services écologiques. Un certain nombre de ces réseaux, aussi bien locaux que nationaux, sont brièvement mentionnés au fil des chapitres de ce cahier de prospective. Pour compléter le panorama national certains, non mentionnés ailleurs, sont présentés dans les Focus 3-2, 3-3, 3-4.

A l'international, différents consortia traitent de thèmes proches de ceux développés par le RTP-GE, notamment celui porté par l'European Science Fondation, Ecological and Evolutionary Functional Genomics (EuroEEFG). Son objectif est de rapprocher les écologues et les biologistes de l'évolution afin d'étudier : 1) le rôle des gènes et de leur régulation dans des processus variés tels que les réponses aux stress (changements environnementaux, interactions biotiques et contaminants), la relation entre génotype et phénotype (au niveau des traits d'histoire de vie, de l'adaptation, de la différenciation écotypique et de la spéciation), 2) le rôle des processus moléculaires sur l'architecture génomique des organismes et les processus de régulation, et 3) le rôle des changements évolutifs à l'échelle de l'écosystème. Ce programme finance huit projets collaboratifs

de recherche transfrontaliers dans lesquels la France est absente.

Une initiative plus large concerne le réseau international des observatoires génomiques, « Genomic Observatories (GO) » (type LTER, Long Term Ecological Research Network) qui regroupe le « Genomic Standards Consortium (GSC) » et le « Group on Earth Observations Biodiversity Observation Network (GEOBON) ». Les objectifs de ce réseau sont : 1) collecter des données génomiques, biophysiques et socio-économiques selon des standards partagés sur des sites d'intérêt et à long terme, 2) construire des modèles prédictifs d'évolution de la qualité et de la distribution des services écosystémiques, et 3) fournir formations, assistance technique, ressources et guide de bonnes pratiques via un site web en particulier pour les GO des pays du Sud. Quatorze observatoires sont actuellement identifiés dans ce réseau : 2 dans la région Indo-Pacifique (dont un en Polynésie française), 8 en Europe (dont celui de Roscoff), 1 en zone polaire, et 3 en Amérique.

FOCUS 3-2 : OCEANOMICS

Le projet OCEANOMICS – wOrld oCEAN biOressources, biotechnologies, and Earth-systeM servICeS - est un projet de recherche fondamentale et appliquée, lauréat du programme des « Investissements d'Avenir », dans sa section « Biotechnologies et Bioressources ». D'une durée de 7 ans, le projet fédère 10 partenaires académiques, 6 partenaires privés et de nombreux autres partenaires non-financés directement mais qui souhaitent collaborer.

Le projet OCEANOMICS vise à comprendre la biocomplexité et le potentiel biotechnologique du plus grand écosystème planétaire : le plancton océanique (Figure 3C). Il s'appuie sur les milliers de données et d'échantillons éco-morpho-génétiques récoltés lors de l'expédition Tara-Océans. OCEANOMICS va tout d'abord explorer cette collection unique qui couvre l'ensemble des communautés planctoniques (des virus aux animaux). Une combinaison de protocoles de séquençage et d'imagerie à très haut débit est mise en place pour extraire l'information de ces échantillons biologiques à plusieurs niveaux systémiques : DNA, RNA, phénotypes. Des comparaisons de ces nouvelles données aux métadonnées environnementales et aux nouveaux génomes et transcriptomes de souches/organismes planctoniques de référence séquencés dans le cadre du projet mèneront à une compréhension taxinomique, métabolique, et éco-systémique profonde de la structure, de la dynamique, et de l'évolution de la biodiversité planctonique.

Cette approche éco-systémique des océans révèle des enjeux considérables. Les 98% du volume de notre biosphère représentés par le plancton laissent imaginer l'énorme ressource potentielle en formes de vie encore inconnues et en composés bioactifs inexplorés. Une fois les connaissances de cette biodiversité approfondies, le projet OCEANOMICS s'orientera vers des collaborations avec ses partenaires privés afin 1) de transférer les nouvelles technologies et méthodes de séquençage et d'imagerie haut-débit à des études de cas en biomonitoring aquatique, 2) de procéder à du phénotypage d'échantillons environnementaux et de souches de choix pour l'analyse de leurs lipides, métabolites secondaires, et exométabolomes ; 3) cribler des souches de choix pour leurs composés bioactifs d'intérêt pharmaceutique, nutraceutique (en terme d'effet positif sur la santé), en aquaculture, cosmétique, et dans les secteurs de l'agriculture et de l'environnement. En parallèle de toutes ces activités scientifiques, OCEANOMICS servira de cas d'étude pour définir un modèle juridique équilibré pour la bio-prospection du plancton marin, un monde encore très peu utilisé, souvent au-delà des territoires nationaux, et en conséquence, à l'extrême limite des cadres réglementaires en vigueur.

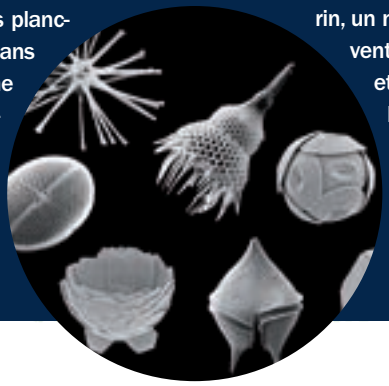


Figure 3C. Aperçu de la diversité morphologique observable chez les protistes marins.

FOCUS 3-3 : IDEALG Seaweed for the future

Les macro-algues marines ont émergé en tant que lignées photosynthétiques indépendantes il y a plus d'un milliard d'années lors de l'intense diversification des eucaryotes. La grande diversité de leurs phylogénies, de leurs modes de vie, de leurs cycles de vie, des composants métaboliques et cellulaires qu'elles synthétisent, ainsi que leurs caractéristiques génétiques en font des modèles particulièrement intéressants pour étudier les processus à l'origine de cette diversité et aussi un immense réservoir pour le développement de nouveaux produits et procédés. Notamment, les parois cellulaires de ces organismes contiennent une grande diversité de nouveaux polysaccharides qui n'ont pas d'équivalent chez les plantes terrestres.

Les macro-algues sont également essentielles pour le fonctionnement des écosystèmes côtiers. Par exemple les champs de laminaires forment d'immenses forêts sous-marines qui abritent un grand nombre d'espèces animales et végétales et qui jouent un rôle majeur dans la structuration de la biodiversité des côtes rocheuses des mers tempérées à froides. Ce sont des écosystèmes très productifs, souvent comparés aux forêts tropicales, qui ont une grande importance écologique mais aussi économique car elles sont exploitées pour leur teneur en alginates et en iode.

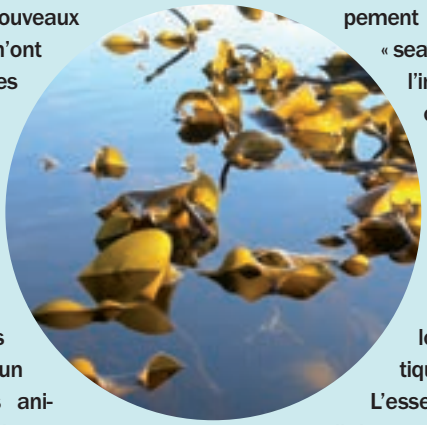


Figure 3D. Champ de laminaires sur les côtes de la Manche.

Le projet IDEALG lauréat de l'appel d'offres « Biotechnologie et Bioressources » des « Investissements d'Avenir » vise à valoriser au mieux ces végétaux marins dans un contexte de développement durable de la filière.

Il fédère la communauté scientifique et les acteurs privés autour de la thématique des grandes algues en Bretagne. Le projet s'intéresse à leur étude génomique et post-génomique afin de développer de nouveaux outils et méthodes permettant d'identifier et sélectionner des populations « ressources » locales ayant un intérêt industriel. Le développement de nouveaux outils de génétique (SNP*, QTL, marqueurs RAD*) et dans certains cas particuliers, la construction de cartes génétiques d'algues ont pour objectif de mieux comprendre les mécanismes fondamentaux de l'adaptation et des interactions phénotype / environnement ainsi que d'améliorer les processus de domestication de ces populations pour contribuer au développement biotechnologique de la filière. Un intérêt est aussi porté aux microorganismes associés aux algues afin de comprendre les interactions métaboliques entre les algues et les bactéries. L'ensemble de ces connaissances servira au développement de la filière de transformation des algues (dégradation, bioconversion, défenses naturelles…).

Un des objectifs du projet est le développement d'une plateforme virtuelle « seaweedomics » qui permettra l'intégration des données « omiques » depuis l'analyse des génomes d'algues et la métagénomique des bactéries associées, la connaissance des voies métaboliques et le phénotypage jusqu'aux développements bioinformatiques afférents.

L'essentiel de la biomasse d'algues produite en France est prélevé à partir de populations naturelles, la Bretagne se situant au cœur d'une région montrant une des plus fortes biodiversités algales au monde (hot-spot s'étendant du sud du Portugal jusqu'au nord de l'Angleterre et des Pays-Bas). Avec l'objectif de développer et diversifier l'utilisation des algues, le projet IDEALG vise aussi à promouvoir les technologies de production d'algues afin d'éviter une trop forte pression de récolte de ces populations naturelles. La culture d'espèces indigènes, non-modifiées et non-invasives représente les principaux critères de la feuille de route d'IDEALG. Le projet IDEALG porte un effort important sur l'étude des impacts de la récolte et de l'algoculture pouvant être générés sur l'environnement, mais aussi sur la société (acceptabilité) et sur l'activité économique des zones littorales. L'enjeu du projet repose sur l'intégration de cette filière à fort potentiel dans un contexte social, économique et environnemental durable.

FOCUS 3-4 : TERRAGENOME

Approches métagénomiques pour l'étude de la microflore du sol

En dépit de leur importance écologique, les microorganismes des sols sont encore très mal connus tant au niveau de leurs diversités taxinomique et fonctionnelle, que de la façon dont ils s'adaptent, évoluent et interagissent entre eux et avec leur environnement. C'est principalement dû au fait que seule une infime proportion des microorganismes qui composent les communautés microbiennes des sols, plus important réservoir de biodiversité, peut être cultivée in vitro. Grâce au développement conjoint des approches métagénomiques (extraction directe de l'ADN bactérien du sol) et des nouvelles méthodes de séquençage à très haut débit (NGS), l'étude des centaines de milliers d'espèces différentes des microorganismes des sols peut aujourd'hui être efficacement initiée. Un tel projet est cependant d'une ampleur considérable et seule une large mobilisation de la communauté scientifique internationale semble en mesure de relever un tel défi qui nécessite de plus une approche multidisciplinaire entre scientifiques spécialistes en microbiologie, écologie microbienne, biologie moléculaire, bioinformatique et physico-chimie des sols.

Figure 3E. Une parcelle à Rothamsted en Angleterre a été choisie par le consortium international « Terragenome » pour séquencer la totalité des génomes des bactéries de son sol.

Le projet « Terragenome » a eu pour objectif dès 2009 de fédérer la communauté scientifique internationale en vue de réaliser le séquençage complet du métagénome (l'ensemble des génomes microbiens) d'un sol de référence de la station agronomique expérimentale britannique à Rothamsted. Les travaux ont été initiés avec un projet français soutenu par l'ANR qui a généré les premiers séquençages d'ADN métagénomique et la construction d'une très importante banque d'ADN après qu'aient été en partie résolus les biais d'extraction. Aujourd'hui, « Terragenome » a élargi ses objectifs à d'autres sols et fédère les travaux de nombreux laboratoires, notamment au travers de colloques annuels visant à partager et harmoniser méthodes expérimentales et d'analyse bioinformatique avec un soutien financier assuré par des agences nationales comme le Thuinen Institute en Allemagne ou la NSF aux Etats-Unis.

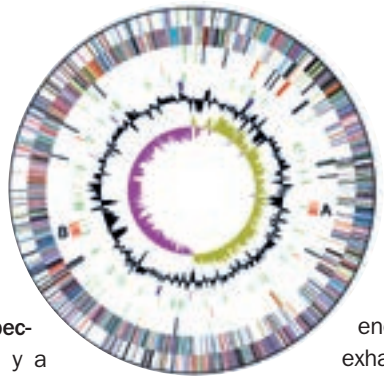
SITES INTERNET

Bibliothèque du Vivant : http://bdv.ups-tlse.fr/
France Génomique https://www.france-genomique.org
Idealg : http://www.idealg.ueb.eu/
Réseau BarCode of life : http://www.barcodeoflife.org/
Réseau Ecomic : http://www.dijon.inra.fr/ecomic/
Réseau StatOmique : http://vim-iip.jouy.inra.fr:8080/statomique/
Tara Oceans et Oceanomics : http://oceans.taraexpeditions.org/
Terragenome : http://www.terragenome.org/

IV ENJEUX, DÉFIS, VERROUS SCIENTIFIQUES ET PROSPECTIVES

Coordinateurs : Dominique Joly et Denis Faure

Les NGS redessinent le champ des possibles dans la recherche en écologie, évolution et sciences de l'environnement en permettant d'accéder à une fraction totalement inconnue de la biodiversité. Sa description comme sa compréhension sont devenues des enjeux majeurs de la biologie. Elles mobilisent tous les acteurs de la recherche au regard desquels les frontières disciplinaires et institutionnelles tombent. Le RTP-GE a participé à ce changement de paradigme en jouant un rôle moteur auprès de la communauté scientifique nationale et en suscitant des regards croisés sur notre appréhension de la complexité et la dynamique du vivant.



Les NGS ouvrent des perspectives, encore insoupçonnées il y a quelques années seulement, pour accéder de manière plus exhaustive que par le passé à la diversité génétique du monde vivant dans tous les domaines d'étude de la biodiversité moléculaire – sa description (code barre ADN, métagénomique, phylogénie moléculaire) comme sa compréhension (génomique évolutive, génétique des populations, écologie moléculaire, écologie des communautés). L'enjeu majeur est la caractérisation de la biodiversité dans toutes ses dimensions, de l'individu aux écosystèmes en passant par les populations et communautés. Les besoins portent sur la connaissance de la dynamique de cette biodiversité et l'évaluation des impacts environnementaux, qu'ils soient d'origine anthropique ou naturelle. Il faut noter cependant que la masse importante des données existantes concernent des organismes bien ciblés (ex. animaux : mammifères, oiseaux, poissons ; ex. végétaux : plantes, lignées, herbacées, algues, champignons) et nous sommes encore bien loin d'avoir une image exhaustive de la distribution, des comportements et des génomes de l'ensemble des organismes macroscopiques (ex. : insectes, crustacés, champignons) et microscopiques, qu'ils soient autonomes, parasitaires ou symbiotiques (virus, protozoaires, bactéries ou helminthes, ces derniers représentant 50 % de la diversité spécifique). Un effort particulier reste donc à faire pour coordonner et diversifier la collecte des données en écologie afin de mieux documenter la magnitude de la biodiversité.

Par ailleurs, l**a montée en puissance de ces technologies à haut débit** rend plus forte la nécessité de circuits courts et maîtrisés entre la conservation des ressources biologiques sur le long terme, leur analyse, l'établissement de bases de données (en connexion avec les informations associées – les « metadata », et les corrélations fonctionnelles, Hardisty et Roberts 2013), et enfin l'exploitation des résultats. Est concerné l'ensemble des échantillons environnementaux, dont

ceux représentatifs d'écosystèmes et de filières agro-alimentaires, afin de disposer de chroniques complètes permettant l'analyse rétrospective des évolutions des différents compartiments de l'environnement. La communauté française fortement impliquée dans toutes les actions du RTP-GE pourra ainsi améliorer de manière significative ses compétences en génomique-métagénomique écologique, systématique et phylogénie, ainsi que sa visibilité internationale, y compris dans le domaine de la conservation.

L'afflux massif de données (l'ère du BigData), et notamment celles en « omiques », génère des défis méthodologiques considérables en termes de gestion et de flux de données, d'analyse bioinformatique et de partage de données. Une approche pluridisciplinaire est donc indispensable à mettre en œuvre.

Les flux d'acquisition de données deviennent extrêmement rapides et volumineux ce qui pose le problème de leur gestion, stockage et exploitation statistique et mathématique. Des développements en bioinformatique et statistique sont à renforcer et soutenir, mais aussi les moyens de stockage et de calcul (implémentation d'outils capables de gérer de gros volumes de données, parallélisation, utilisation de grilles de calcul, etc…). Un défi particulier concerne le domaine de l'**open data** qui rend public de grandes quantités de données souvent inexploitées. Il s'agit alors de modéliser de manière intégrée à la fois la distribution de ce qui est mesuré et la distribution de l'effort d'observation. Un autre champ ouvert, qui pourrait s'appuyer sur les récents développements de la bioinformatique, est celui de l'analyse des interactions (réseau trophique, compétition inter et intra spécifique, etc…) à différentes échelles biogéographiques en tenant compte de covariables explicatives (traits, phylogénie, etc…).

Le traitement et l'exploitation des données nécessitent une **complémentarité de compétences** entre biologistes et bioinformaticiens qui devient ainsi un enjeu majeur de réussite par un partage de connaissances et de savoir-faire transdisciplinaires. Les données NGS sont typiquement fragmentées et bruitées, et sont par définition non-ciblées (le séquençage est réalisé en aveugle). Leur analyse implique la mise en œuvre de modèles et d'algorithmes nouveaux, qui prennent en compte la redondance dans les génomes et les biais expérimentaux associés aux technologies haut débit, tout en faisant face aux changements d'échelle permanents qu'elles induisent. Les méthodes actuelles prolifèrent au cas par cas, chacune développant des outils « maison », peu transposables et sans standard. Une activité intense de développements méthodologiques et logiciels est à soutenir fortement.

Le dernier aspect concerne **l'interfaçage et le partage des données**. Cette nouvelle façon de travailler engendre une véritable révolution culturelle, aussi bien au niveau des acteurs de la recherche que des institutions, afin de promouvoir la mise en ligne de jeux de données exploitables par autrui et la valorisation de cet investissement individuel dans les carrières scientifiques. Un **effort important** devra être fourni pour mutualiser des moyens de stockage via des accès sécurisés au niveau de plateformes régionales ou nationales, pour créer des bases de données ouvertes, normalisées, curées et révisées sur le long terme (voir les initiatives Ecological Data et DataOne), et enfin pour imposer des standards de qualité minimale définis par des consortia internationaux (Ecoinformatique).

A terme les données NGS doivent aussi être dirigées vers la **modélisation**. La principale difficulté vient de l'estimation des paramètres à prendre en compte et surtout de l'abondance de leur nombre (possibilité d'adopter des approches par automates soit temporisées, soit probabilistes). D'autres alternatives sont également à explorer via le data-mining qui consiste à extraire des connaissances à partir de grands jeux de données. Ici, un effort particulier devra être mis sur la synthèse des observations et leur interfaçage avec les autres niveaux d'organisation.

RÉFÉRENCES

Hardisty and Roberts. 2013. Data accessibility: Getting a handle on biological data. Nature 484: 318.

SITES INTERNET

DataOne : http://www.dataone.org

Ecoinformatique : http://knb.ecoinformatics.org/software/eml

Ecological Data : http://ecologicaldata.org

V ACCÈS ET PARTAGE DES DONNÉES NGS

Coordinateurs : Eric Pelletier et Guy Perrière

La recherche en génomique environnementale est en train de vivre une véritable révolution de l'information. Sont mis à disposition de la communauté scientifique des jeux de données de plus en plus importants, pour un nombre croissant d'organismes, ainsi que pour de nombreux gènes et écosystèmes. Ces nouveaux jeux de données, alliés à une puissance de calcul et à une sophistication des logiciels (rendues possibles par la mise en place de plateformes collaboratives comme R) offrent aujourd'hui une profondeur d'analyse qui n'était pas possible il y a encore dix ans. Toutefois l'avalanche de ces données dans le domaine des omiques est telle que l'on se heurte aujourd'hui aux difficultés qui en font leur succès en ce qui concerne leur accès et leur partage.

Accès aux banques de données NGS

La possibilité d'accéder rapidement et de façon exhaustive aux séquences génomiques est l'une des raisons du succès qu'a rencontré la bioinformatique au cours des trente dernières années. Or, l'arrivée en 2005 des premières méthodes à haut débit (Margulies *et al.* 2005), puis leur démocratisation dans les années 2010, a entraîné une véritable révolution dans ce domaine (Shendure et Lieberman Aiden 2012). En effet, si les instituts en charge de la maintenance des banques ont pu gérer pendant trois décennies un flux sans cesse croissant de données, ce n'est désormais plus le cas. Faisant face à un **déluge sans précédent** de séquences (génomiques ou transcriptomiques), ces organismes doivent désormais s'adapter et ici, comme dans de nombreux autres domaines des sciences et de la technologie, l'avenir appartient aux structures en réseau et non plus à de grands systèmes centralisés.

Depuis près de trente ans, les trois grandes banques généralistes de données collectant les séquences génomiques sont GenBank au National Center for Biotechnology Information (Benson *et al.* 2013), l'ENA (European Nucleotide Archive), ex-EMBL, à l'European Bioinformatics Institute (EBI, Cochrane *et al.* 2013) et la DDBJ (DNA Data Bank of Japan) au National Institute of Genetics (Ogasawara *et al.* 2013). Bien qu'à leurs débuts le contenu et la taille de ces trois banques étaient relativement différents, une **collaboration internationale** s'est rapidement établie et, depuis

25 ans, leur contenu est virtuellement identique : une séquence soumise à n'importe lequel des trois centres sera transmise avec un délai maximal de 24 heures aux deux autres.

La caractéristique principale de ces banques est qu'elles permettent d'**accéder librement** à la quasi-totalité des séquences biologiques obtenues par les laboratoires publics et privés. Cet accès non limité a en grande partie contribué aux avancées importantes obtenues par la bioinformatique au cours de son existence en tant que discipline scientifique. En effet, la question fondamentale en science de la reproductibilité des résultats a été facilitée par cette disponibilité immédiate et exhaustive des données.

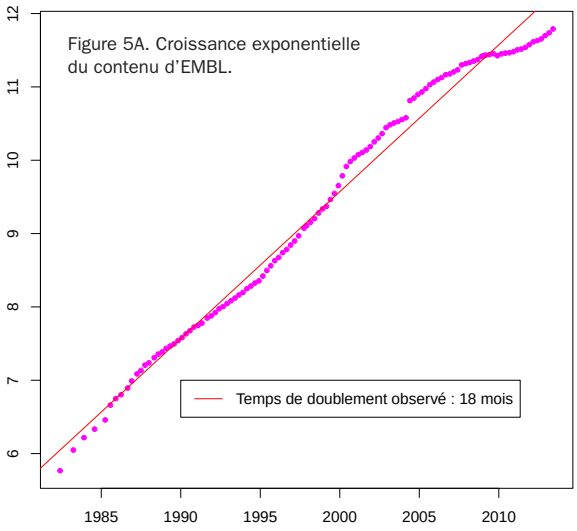


Figure 5A. Croissance exponentielle du contenu d'EMBL.

Le volume des données soumises à ces trois collections a crû de façon exponentielle, avec un **temps de doublement moyen de l'ordre de 18 mois** (Figure 5A). La période 2000-2010 a même vu ce temps de doublement diminuer avec le séquençage de nombreux génomes ou transcriptomes, dont le premier génome humain (Venter *et al.* 2001). Cependant, depuis 2010, un **changement de tendance** assez surprenant s'opère puisque c'est un allongement de ce temps de doublement que l'on observe. Une première explication de ce phénomène inattendu tient au fait que les centres en charge de la maintenance des banques sont de moins en moins capables de supporter les charges financières que représentent l'achat continuel de capacités supplémentaires de stockage ainsi que la maintenance des infrastructures associées.

Un autre problème est que les volumétries de données produites sont désormais telles qu'il n'est **plus possible de les transmettre** en un temps raisonnable aux centres de saisie via le réseau. Une solution de plus en plus utilisée, pour qui veut voir ses séquences figurer dans les banques, est d'expédier un disque dur sur lequel sont sauvegardées les séquences en question, puis de procéder au transfert sur place ! D'une certaine façon, il s'agit là d'un retour aux pratiques pré-internet puisque, jusque vers la fin des années 1980, c'est par envoi postal de supports physiques (bandes magnétiques ou disquettes) que se transféraient les séquences.

Enfin, la question de l'accès aux lectures* courtes, non annotées, est également un problème d'importance. Du fait de la quantité de séquences disponibles, les centres ne proposent plus un accès direct aux entrées individuelles, mais plutôt à des **archives compressées** pouvant contenir un grand nombre de lectures. La survie de ces archives a été remise plusieurs fois en question, ce service ayant déjà été supprimé une fois puis rétabli à l'EBI.

Dans ce contexte, de plus en plus de séquences ne sont **tout simplement pas envoyées** aux centres de saisie. Leur mise à disposition pour la communauté se fait par l'intermédiaire de **banques de données locales** mises en place dans le cadre de projets limités. La conséquence est qu'il existe désormais, outre les collections généralistes précitées, une **véritable pléthore** de banques spécialisées, qu'elles soient dédiées à un organisme ou à une problématique biologique particulière. La revue Nucleic Acids Research publie chaque année un numéro spécial consacré aux principales banques disponibles dans le monde. Dans son édition de janvier 2013, ce ne sont pas moins de 178 banques qui étaient recensées. Cependant, ce catalogue est très loin d'être exhaustif et le nombre de banques spécialisées existant est probablement beaucoup plus élevé.

En **perspectives**, du fait qu'une quantité croissante de séquences ne soient plus envoyées aux centres de saisie, les trois **collections généralistes ne peuvent plus être considérées comme exhaustives**. Or cette perte d'exhaustivité a d'ores et déjà des répercussions sur cette reproductibilité facile des résultats qui était l'apanage de la bioinformatique. C'est dans le but de pallier ce problème, que l'EBI a lancé,

début 2007, l'**initiative ELIXIR**. Cette initiative vise à fédérer les grands centres de bioinformatique (nationaux ou régionaux) dans un **réseau Européen**, chaque nœud ayant une spécificité thématique. Une telle réorganisation permettrait effectivement de répartir la charge, aussi bien en quantités de données à gérer qu'en termes d'infrastructures, pour ce qui est du stockage ou de l'archivage des séquences biologiques.

En cas de participation de la France à cette initiative, le **nœud national** serait représenté par l'IFB (Institut Français de Bioinformatique - Voir Focus 3-1). La structure de l'IFB est en fait proche de celle d'ELIXIR avec, d'une part, un hub central (l'IFB-core, localisé à Gif-sur-Yvette) et d'autre part, les six centres ReNaBi actuels.

Production de données NGS : enjeux techniques et académiques

Depuis 2005, l'avènement des nouvelles technologies de séquençage a bouleversé l'impact de la génomique sur de nombreux champs de la biologie. Avec aujourd'hui des coûts inférieurs à 50 USD le gigabase* et la multiplication des plateformes de production de données, les chercheurs et les laboratoires doivent opérer des choix stratégiques en ce qui concerne leur besoin en terme de séquençage.

Les NGS correspondent principalement à un **changement de paradigme.** Le séquençage par la méthode de Sanger se base sur une amplification de séquence par PCR avec des bases déoxy-NTP (qui permettent la poursuite de l'élongation) et dideoxy-NTP (qui, après incorporation, bloquent la poursuite de l'élongation), puis la séparation physique de produits de tailles différentes et leur détection optique permettant de reconstituer la succession des bases de la chaîne matrice. Les NGS couplent l'amplification des matrices sur support fixe avec détection optique simultanée. Ce couplage des opérations a permis de gagner **plusieurs ordres de grandeur dans la quantité** de lectures possibles simultanément (de 384 lectures de 900 pb en moyenne en 16 heures sur un appareil de type ABI 3730 XL en 2004 à aujourd'hui plus de 6 x 10^9 lectures de 100 pb en 12 jours sur un appareil de type Illumina HiSeq2000 aujourd'hui, pour un coût similaire). Ce changement d'échelle majeur a permis l'émergence de plusieurs types de machines, qui diffèrent par le mode de détection (optique comme pour Illumina et 454 ou électrochimique comme pour IonTorrent), par la quantité de séquences produites en un run* et par la durée de celui-ci.

Si le coût d'achat et de fonctionnement de tels appareils est maintenant à la portée d'un laboratoire classique (surtout pour les machines de débit modeste), il faut garder à l'esprit que la **technicité d'opération de ces appareils** est élevée, nécessite souvent des aménagements de locaux spécifiques, et surtout reste en perpétuelle évolution : les kits et les protocoles changent plusieurs fois par an, ce qui demande aux personnels en charge de maintenir une **veille technologique constante** pour conserver un usage optimal de la machine. S'équiper d'un séquenceur aujourd'hui demande une réflexion poussée, ces machines demandant souvent un usage continu pour maintenir leur état de fonctionnement.

Un autre aspect important doit être pris en considération : la partie informatique de la production de séquences. Les quantités de séquences produites sont élevées, ce qui nécessite un traitement en continu, des personnes compétentes dédiées, et un stockage des données. Ainsi, un run HiSeq2000 complet produit près de 1 To de données chaque quinzaine. Les estimations actuelles donnent le ratio coût de gestion informatique et d'analyse sur coût de production des séquences pour des valeurs de 10 à 100, selon le type de projet.

La **prochaine évolution** attendue dans le domaine des technologies de séquençage consiste en un **autre changement de paradigme** : la suppression de la phase d'amplification (séquençage de molécule unique). Les premiers prototypes commercialisés (Pacific Biosciences) ont un fonctionnement très délicat et le type de séquences produites (avec un taux d'erreur par base de 15%) reste encore instable. De plus, la quantité d'ADN nécessaire reste élevée (plusieurs microgrammes). Ce saut ne sera pas tant un saut quantitatif (difficile d'envisager prochainement de dépasser significativement les capacités de production des machines de type HiSeq d'Illumina) qu'un **saut qualitatif** : lectures très longues (plusieurs kilobases voire même dizaines de kilobases) à partir de faibles quantités de matériel. Différents types de séquences produites par différents appareils coexistent déjà, et chaque type de projet fait appel à plusieurs d'entre eux pour optimiser le résultat et la qualité de l'analyse qui suit.

Les capacités de production de séquences accessibles aux laboratoires de recherche sont variées. En ce qui concerne les moyens publics, les plateformes régionales et nationales sont en cours d'intégration partielle dans France Génomique (voir Focus 3-1), avec pour objectif d'optimiser la gestion, l'évolution et l'utilisation de ces moyens. Pour le projet Génome Humain (Lander *et al.* 2001, Heilig *et al.* 2003), la création du Génoscope – Centre National de Séquençage plaçait la France au 5ème rang mondial de capacité de production de séquences. **La place actuelle des moyens de production français est devenue modeste**. Au niveau international, de nombreuses structures ouvrent l'accès à leurs moyens de production à travers des appels à projets (BGI, JGI, Broad Institute, etc). Pour le volet privé, de nombreuses compagnies proposent leurs services, avec différents niveaux de traitements de données. Les laboratoires doivent accorder une attention particulière aux clauses de confidentialité et de propriété intellectuelle, surtout en ce qui concerne les travaux externalisés vers des sociétés et centres basés à l'étranger. Enfin, un aspect souvent négligé et mal évalué dans les projets portés par les chercheurs concerne les capacités de calcul et de stockage des données, ainsi que le personnel compétent pour analyser les données générées.

RÉFÉRENCES

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. Nucleic Acids Res 41:D36-42.

Cochrane G, *et al.* 2013. Facing growth in the European Nucleotide Archive. Nucleic Acids Res 41:D30-35.

Heilig R, *et al.* 2003 The DNA sequence and analysis of human chromosome 14. Nature 421:601–607.

Lander ES, *et al.* 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Margulies M, *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

Ogasawara O, Mashima J, Kodama Y, Kaminuma E, Nakamura Y, Okubo K, Takagi T. 2013. DDBJ new system and service refactoring. Nucleic Acids Res 41:D25-29.

Shendure, J, Lieberman Aiden E. 2012. The expanding scope of DNA sequencing. Nat Biotechnol 30:1084-1094.

Venter JC, *et al.* 2001. The sequence of the human genome. Science 291:1304-1351.

SITES INTERNET

454 : http://www.454.com/
BGI : http://www.genomics.cn/en/index
Broad Institute : https://www.broadinstitute.org/
DDBJ http://www.ddbj.nig.ac.jp/intro-e.html
ELIXIR : http://www.elixir-europe.org/
ENA http://www.ebi.ac.uk/ena/about/about
GenBank : http://www.ncbi.nlm.nih.gov/genbank/
Illumina : http://www.illumina.com/
Ion Torrent : http://www.iontorrent.com/
JGI : http://www.jgi.doe.gov/
NAR : http://nar.oxfordjournals.org/content/41/D1.toc
Pacific Biosciences : http://www.pacificbiosciences.com/

VI QUALITÉ DES DONNÉES NGS : DE LA SÉQUENCE AUX BASES DE DONNÉES

Coordinateur : Pierre Peyret
Contributeurs : Julie Aubert, Vincent Breton, François Enault, Line Le Gall, Denis Le Paslier, Tiphaine Martin, Guy Perrière, Eric Peyretaillade

La révolution technologique du séquençage avec les NGS autorise l'exploration de la diversité génétique du monde vivant provenant d'échantillons biologiques plus ou moins complexes depuis l'organisme isolé jusqu'à l'intégralité des communautés d'un écosystème. Cependant, ces évolutions se sont faites au détriment de la longueur et de la qualité des séquences posant de nouveaux problèmes d'analyse.

Le développement des plateformes de séquençage de deuxième génération a conduit à la production de données de séquences à des coûts très bas avec des débits considérables (Glenn 2011 ; Figure 6A). Ce **déluge de séquences** nécessite donc le développement de **nouveaux outils bioinformatiques** pour assurer un traitement optimal de l'information (Logares *et al.* 2012). Les besoins en espaces d'archivage et en capacités de calculs se trouvent démultipliés conduisant au déploiement de nouvelles infrastructures informatiques. La qualité des données générées et la pertinence des analyses sont un enjeu majeur de la (méta)génomique pour donner un nouvel éclairage aux mécanismes régissant le fonctionnement du vivant et à l'origine de cette formidable diversification créant ce subtil équilibre de la vie (Figure 6B).

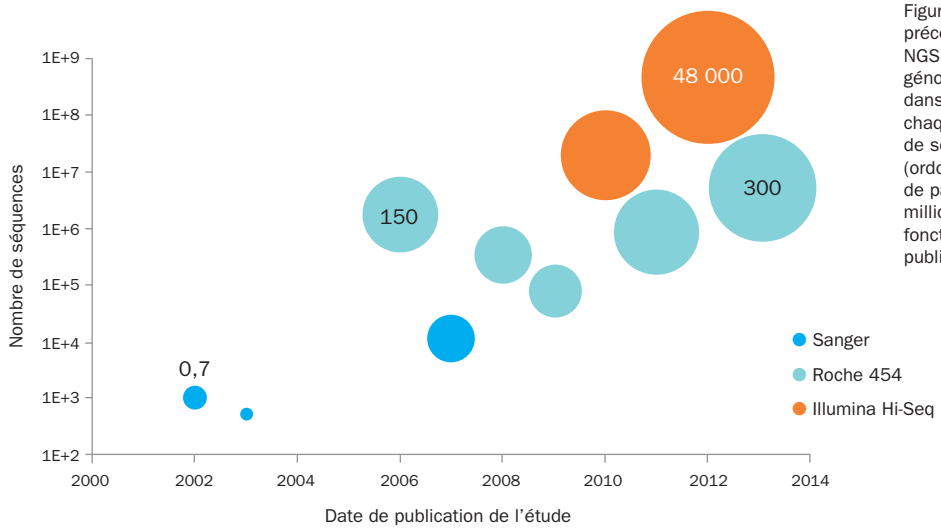


Figure 6A. Afflux sans précédent de données NGS : exemple des métagénomes viraux publiés dans la littérature. Pour chaque étude, le nombre de séquences étudiées (ordonnée) et la quantité de paire de bases (en million) sont reportés en fonction de l'année de publication (abcisse).

Figure 6B. Qualité des données. L'explosion du séquençage nouvelle génération entraîne un déluge de données, qui doivent être traitées efficacement par des approches innovantes de bioinformatique, sous peine d'être submergé par l'information et rester aveugle devant l'extraordinaire diversité du monde vivant à la base du fonctionnement des écosystèmes.

La qualité des données issues des nouvelles techniques de séquençage est liée non seulement à la technologie utilisée mais également, dans certains cas, aux étapes amont ayant permis l'obtention du matériel génétique à séquencer. Ainsi, les algorithmes bioinformatiques doivent éliminer les régions non informatives biologiquement, détecter les régions de mauvaise qualité, identifier les erreurs de séquençage ainsi que les séquences issues d'artéfacts de manipulation (Quince *et al.* 2011). Des séquences de mauvaise qualité peuvent en effet, compromettre les analyses ultérieures (assemblage*, annotation) mais également surestimer une diversité non représentative des organismes étudiés ou des environnements explorés conduisant à des interprétations erronées. Au final, les séquences de qualité retenues, permettront de refléter le plus fidèlement possible l'information génétique initiale issue des échantillons et de mettre en place des traitements statistiques pour tester les hypothèses initialement posées (Focus 6-1).

La qualité des affiliations est un verrou méthodologique et scientifique pour appréhender la diversité des organismes. La diversité du monde vivant ne permet pas de caractériser les organismes sur de simples caractères phénotypiques. Ainsi, l'identification des organismes passe par l'analyse de gènes phylogénétiquement informatifs pouvant être isolés facilement et permettant d'établir des relations de parenté. Cette approche d'identification des organismes par « étiquetage » est communément appelée barcoding (voir chap. VIII). Les parentés phylogénétiques ne sont pas toujours testées du fait notamment de la nécessité de capacités de calcul importantes pour la reconstruction des arbres. Cependant, les développements récents de méthodes ne nécessitant pas de recalcul complet des arbres (e.g. pplacer) devraient permettre d'améliorer la situation. Cette approche a entraîné une non-affiliation de nombreuses séquences ou une affiliation erronée par simple recherche de similarité de séquences pouvant par voie de conséquence se propager. L'utilisation de plusieurs marqueurs phylogénétiques et/ou de génomes complets ouvrent la voie d'une **phylogénomique** plus résolutive précisant les identifications et les parentés des organismes avec des applications récentes sur des données environnementales (Chivian *et al.* 2013). Enfin, l'utilisation d'une **nomenclature officielle** définissant le nom des espèces est également essentielle même si les taxinomies évoluent régulièrement (Yarza *et al.* 2013).

FOCUS 6-1

Quelques règles de bonne conduite pour améliorer la qualité des résultats d'analyse différentielle à partir de données NGS

Les technologies de séquençage à très haut débit sont des outils puissants pour explorer de nouvelles pistes de recherche dans de nombreux domaines de la biologie. Afin d'obtenir des résultats pertinents à partir de ces masses de données, des outils informatiques et statistiques adaptés sont nécessaires, mais une bonne réflexion préalable aux expériences reste indispensable (Figure 6C). Est présentée ici la recherche de régions d'intérêt (ou de rang taxinomique) différentiellement exprimées (ou différentiellement abondantes) entre plusieurs conditions à partir de données de comptage RNA-Seq (ou métagénomiques).

Chaque étape, de la production au traitement des données, a un impact non négligeable sur les suivantes. Il est donc important d'explorer les données pour adapter la stratégie d'analyse statistique à la fois à la question biologique d'intérêt et aux données recueillies. La normalisation consiste à détecter les biais techniques et à les corriger en vue de rendre les échantillons comparables. Elle est propre à chaque technologie et à chaque plateforme. Cette étape est délicate car elle revient à une modification des données brutes. Il est donc important qu'elle se limite au strict nécessaire. Certains biais peuvent être éliminés par un plan d'expérience adapté, un protocole expérimental judicieusement choisi et un traitement bioinformatique efficace. Le biais principal est la différence de profondeur de séquençage (nombre total de lectures alignées sur les différentes régions d'intérêt) entre les échantillons. Dillies *et al.* (2013) ont montré que les méthodes basées sur une taille de banque efficace, définie à partir de régions d'intérêt peu variables d'un échantillon à l'autre, sont les plus adaptées. Elles sont efficaces même en cas de répertoires d'ARNm exprimés très différemment. Si des biais de type échantillons spécifiques dus à la teneur en GC sont observés, une normalisation supplémentaire peut s'avérer nécessaire. L'analyse différentielle consiste, quant à elle, à l'aide d'un test statistique, à déterminer les régions d'intérêt statistiquement significatives à un seuil choisi. Les méthodes spécifiques au RNA-Seq notamment DESeq (Anders et Huber 2010) ont été développées dans le cadre d'un faible nombre de répétitions (moins de 5) par condition. Soneson et Delorenzi (2013) ont montré que lorsque le nombre de répétitions augmente, ces méthodes ne sont plus forcément les plus puissantes.

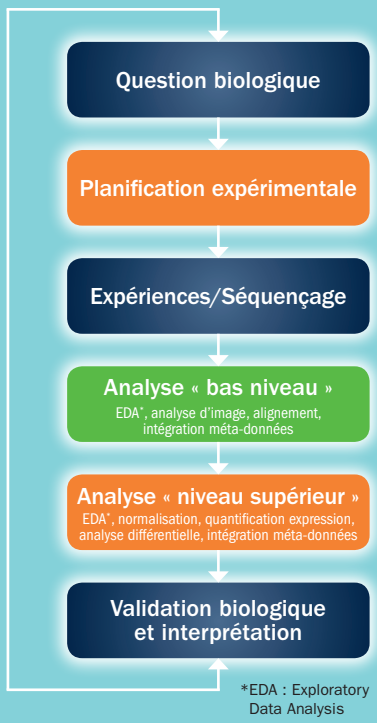


Figure 6C. Procédure d'analyse des données NGS.

Pour tirer le maximum de conclusions pertinentes de ces expériences à haut débit, il est important d'intégrer dès le démarrage d'un projet les différents acteurs : biologistes, bioinformaticiens, statisticiens ; de planifier en amont les expériences et d'anticiper les biais possibles ; d'adapter les méthodes d'analyse aux questions possibles et d'interpréter ces données en connaissance de cause.

FOCUS 6-2

Stratégies d'analyse des métagénomes viraux

Même si leur influence sur l'écologie de notre planète et l'évolution des organismes est reconnue, les virus de l'environnement restent largement méconnus. Afin de les étudier sans mise en culture, les approches métagénomiques sont appliquées aux communautés virales depuis une dizaine d'années (Edwards et Rohwer 2005). Si extraire les particules encapsidées d'un échantillon n'est pas trivial, l'analyse bioinformatique des données est la principale étape limitante de telles études. En effet, les serveurs généralistes de traitement des données de métagénomique comme Mg-Rast et Camera sont adaptés aux données microbiennes mais pas aux données virales. Les spécificités des génomes viraux, en particulier l'absence de gène marqueur universel et le faible taux de séquences affiliées (moins de 30%), rendent nécessaire l'utilisation d'outils qui leur sont adaptés.

Seuls deux serveurs Web spécifiques aux viromes permettent à l'heure actuelle de réaliser des analyses complètes de ces jeux de données : VIROME (Wommack *et al.* 2012) et Metavir (Roux *et al.* 2011). VIROME propose une affiliation de chaque séquence métagénomique par comparaison à des génomes de référence* mais aussi à des séquences environnementales. L'utilisation de ces dernières permet d'annoter plus de séquences et de quantifier l'abondance des séquences dans différents écosystèmes. Metavir tente également de faire face au faible taux de séquences affiliées au travers d'analyses utilisant l'ensemble des jeux de données : 1) estimation de la diversité génétique à partir d'une clusterisation des séquences et 2) comparaison des viromes entre eux à partir de plusieurs méthodologies (comparaisons directes des séquences ou de leurs fréquences en k-mers).

L'analyse de métagénomes viraux étudiés jusqu'à présent a montré que les communautés virales sont très diversifiées et riches, le type d'environnement semblant constituer le principal facteur influençant la composition de ces communautés. En plus du faible taux de séquences affiliées, les phylogénies de gènes marqueurs de différentes familles virales réalisées par Metavir montrent que les séquences des viromes sont le plus souvent éloignées des séquences de référence disponibles, témoignant du fait que la grande majorité des virus de l'environnement nous est encore inconnue à l'heure actuelle.

FOCUS 6-2 (Suite)

Les technologies de séquençage récentes (e.g. Illumina HiSeq2000) sont cependant en train de modifier le type même de données générées. En effet, la quantité de données produites (jusqu'à 50 Gb) rend possible (et indispensable) un assemblage des séquences brutes, ce qui permet d'obtenir de larges fragments d'ADN génomique voire des génomes complets. L'analyse de ces larges fragments nécessite à son tour de nombreux développements spécifiques et une nouvelle version de Metavir est en cours de développement pour permettre une annotation des milliers de contigs* obtenus ainsi que leur visualisation (Figure 6D). Au-delà d'un accroissement des capacités de calcul du fait de l'augmentation des quantités de données générées, il convient d'adapter les méthodes d'analyse de la composition en espèces et en gènes, de comparaison entre viromes, ou encore d'analyse du contexte génomique (co-occurence et proximité des gènes sur les fragments génomiques) afin d'identifier de nouveaux liens fonctionnels entre gènes et ainsi d'annoter les nombreux gènes viraux pour lesquels la fonction reste inconnue.

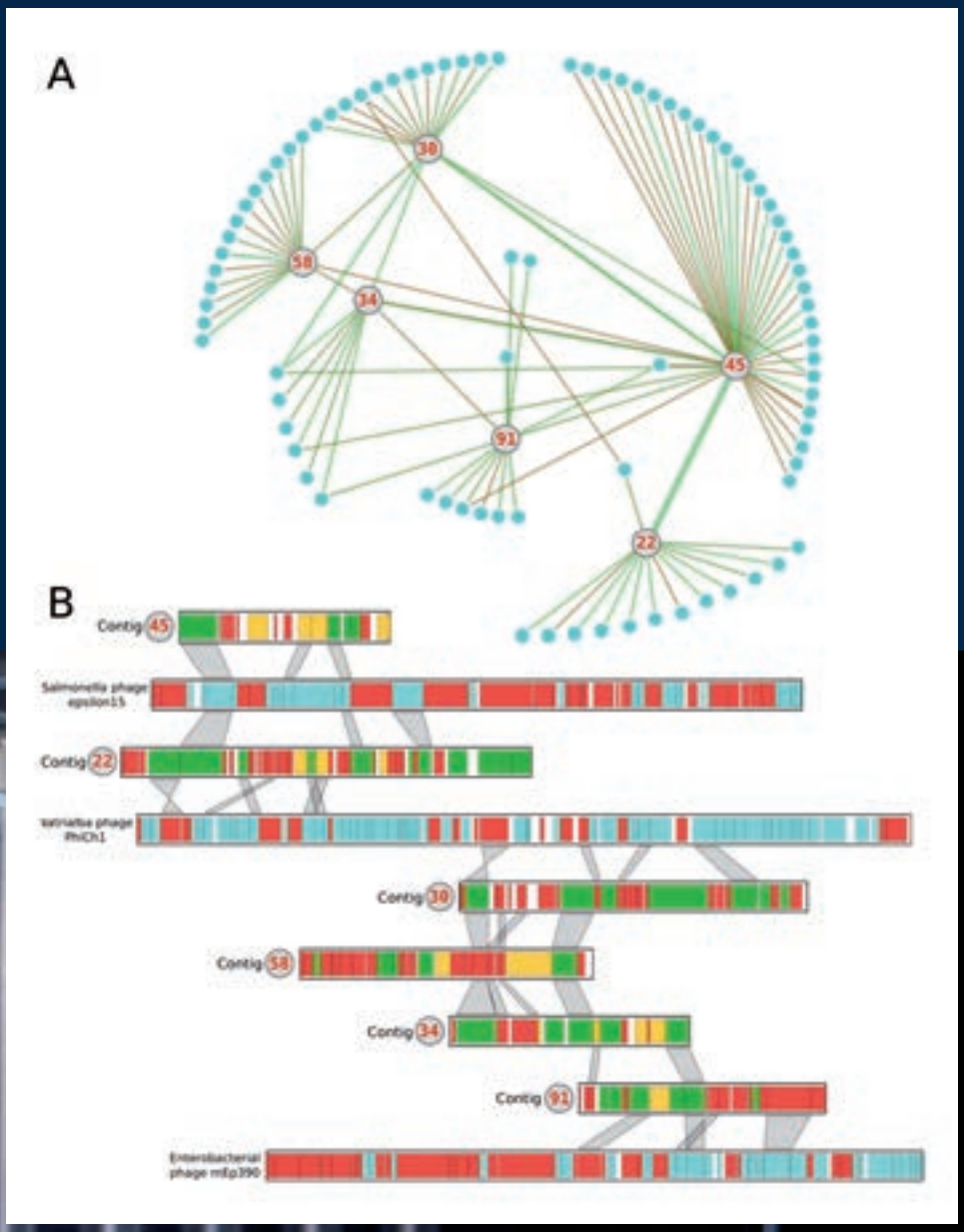


Figure 6D. Exemple de visualisation sur Metavir de quelques contigs assemblés à partir d'un virome généré par Illumina HiSeq 2000. (A) Réseau composé de 6 contigs (cercles gris) et de génomes de référence (cercles bleus), les arcs représentant les similarités entre ces séquences génomiques. (B) Cartes génomiques et similarités entre les 6 contigs et 3 virus de référence.

FOCUS 6-3

Qualité des annotations

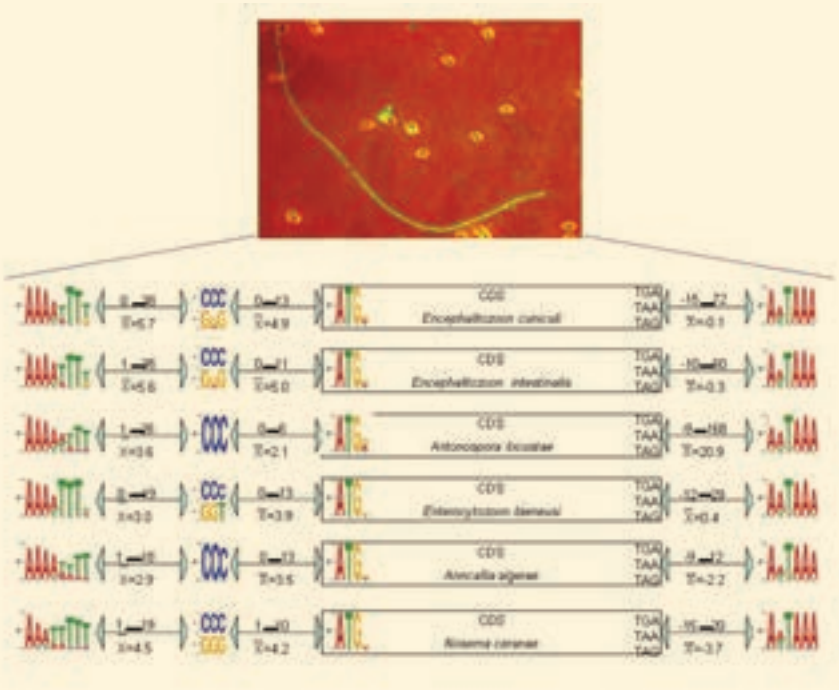


Figure 6E. Annotation expertisée des séquences de microsporidies, pathogènes intracellulaires obligatoires. Visualisation des spores de microsporidies et du système d'invasion appelé tube polaire. Signaux de régulations utilisés pour la détection des CDS (Coding DNA Sequence) lors de l'annotation.

La prédiction de la structure des gènes reste l'un des problèmes les plus importants et passionnants de la biologie computationnelle (Brent 2008). Les approches basées sur l'exploitation des données extrinsèques ont considérablement augmenté la qualité des prédictions de novo. Néanmoins, ces données n'étant pas toujours disponibles, l'emploi d'algorithmes de prédiction ab initio utilisant les données intrinsèques de la séquence reste l'unique solution. Cependant, malgré des progrès conséquents, ces méthodes ne sont pas capables de produire un catalogue in extenso de l'ensemble des gènes (Brent 2008). Afin d'améliorer la prédiction, une définition plus fine des règles définissant la présence ou l'absence d'un gène sur une portion d'ADN doit être envisagée.

Pour illustrer ce propos, l'exemple de l'annotation du génome des microsporidies sera choisi. En effet, ces organismes, pour lesquels près de 1500 espèces réparties en 187 genres ont été décrites, sont des parasites capables d'infester des protozoaires et la plupart des organismes pluricellulaires (invertébrés et vertébrés). Les approches de génomique sont donc idéales pour une meilleure compréhension de ces organismes retrouvés dans tout type d'environnement mais ne pouvant pas être isolés facilement. A l'heure actuelle, un nombre réduit de génomes a été en partie ou entièrement séquencé et annoté en utilisant des méthodes que l'on peut qualifier de généralistes. Ces approches restent encore peu efficaces comme le montre par exemple la prédiction erronée de près de 30% des gènes du génome d'*Encephalitozoon cuniculi*. Ce sont des travaux exploitant conjointement les données intrinsèques et extrinsèques mais également les signaux de régulation de l'expression des gènes, qui ont permis l'élaboration d'une nouvelle méthode à la fois originale et innovante d'annotation des séquences de ces organismes (Figure 6E) (Peyretaillade *et al.* 2012)

L'utilisation de cette méthode a permis d'assurer de manière efficace la ré-annotation de génomes microsporidiens en identifiant des gènes non ou mal prédits. Pour exemple, citons l'identification de 387 gènes annotés au sein du génome de l'espèce *Enterocytozoon bieneusi* qui correspondent en réalité à des séquences d'espèces bactériennes appartenant au genre *Pseudomonas* ayant comme origine une contamination et non un transfert. Par ailleurs, cette approche s'est également montrée pertinente pour assurer la caractérisation de gènes n'ayant pas pu être identifiés en raison notamment de leur trop faible taille. En effet, bien que ces petits gènes puissent exercer des fonctions importantes, leur identification reste une tâche ardue et relève le plus souvent du hasard. L'exploitation de l'ensemble des caractéristiques des séquences a également permis une ré-annotation pertinente des codons d'initiation de la traduction. En effet, l'identification de ces codons qui permettent de définir la séquence exacte de chaque protéine représente un autre challenge majeur de l'annotation des gènes.

Pour conclure, l'amélioration des algorithmes d'annotation doit donc passer par l'analyse de grands fragments d'ADN ou de génomes constituant des données de référence. Aussi la constitution de bases de données de référence associées à des algorithmes dédiés performants assurera des annotations dynamiques expertisées propageant des données de qualité.

La qualité des annotations des génomes pâtit de l'augmentation exponentielle du nombre de génomes séquencés. Paradoxalement, les connaissances restent encore minimes au regard de l'extraordinaire volume de données produites et de la diversité du monde vivant. La métagénomique fait abstraction des difficultés d'isolement des microorganismes pour étudier simultanément l'ensemble des génomes des organismes constituant une communauté, révélant ainsi une diversité encore insoupçonnée (Temperton et Giovannoni 2012). Par exemple, les données NGS ont permis de révéler un monde à part avec une abondance de virus incroyable dans tous les environnements explorés (Focus 6-2). Cependant, l'extraordinaire complexité des communautés microbiennes de la plupart des écosystèmes rend difficile, voire impossible, la reconstruction des génomes à partir de séquences de petites tailles même si elles sont générées de façon massive. En effet, les logiciels d'assemblage de séquençage et les capacités informatiques ne sont pas suffisamment performants pour réaliser un assemblage *de novo** complet (Namiki *et al.* 2012). Se pose le problème de la détection de gènes au sein de millions de petits fragments d'ADN qui peuvent de plus porter des erreurs de séquençage. Malgré ces difficultés, des logiciels très populaires permettent de donner les grandes orientations d'affiliation, d'annotation, de comparaison de (méta)génomes et ainsi d'appliquer des analyses statistiques plus ou moins élaborées pour la comparaison des données (Caporaso *et al.* 2010). Néanmoins, la qualité des annotations étant un point crucial pour constituer des données de référence, des efforts conséquents doivent être entrepris pour le développement de stratégies d'annotation efficaces et pertinentes (Focus 6-3). Il est également important d'améliorer la qualité des génomes de référence par le développement de méthodes de classement (binning) efficaces (Albertsen *et al.* 2013).

La qualité des bases de données est essentielle à la valorisation des données issues des NGS. La première source de données nucléiques assurant le partage et la diffusion de la connaissance (International Nucleotide Sequence Database Collaboration) a comme origine une collaboration entre les bases japonaise (DNA Data Bank of Japan), américaine (GenBank) et européenne (European Nucleotide Archive, voir chapitre V). Les collections de données nécessitent l'utilisation de standards assurant la qualité de soumission pour un partage et une réutilisation efficace. De telles initiatives ont été menées par le Genomic Standards Consortium (GSC) pour les données de génomique et métagénomique mais également par le Consortium pour le Barcode of Life (voir chapitre III). Ce consortium vise à identifier des organismes (animaux, plantes et champignons) par de courtes séquences d'ADN de gènes choisis comme standards. Ainsi, l'objectif du projet Barcode Of Life Database (BOLD) est de mettre à disposition une plateforme web permettant d'identifier un organisme par comparaison avec une base de données, devant atteindre à terme 5 millions de spécimens représentant près d'un demi-million d'espèces, en lien avec leur géo-référencement et leur conservation dans des collections d'Histoire Naturelle. Des bases de données spécialisées ont également été développées pour identifier les microorganismes procaryotes (RDP, Greengenes, SILVA). Ainsi, des bases de données spécialisées se multiplient, qu'elles soient dédiées à des analyses phylogénétiques, (méta)génomiques ou métaboliques. Néanmoins, l'explosion des données de séquences nécessite de nouvelles évolutions informatiques facilitant les transferts, la sécurisation, le stockage et les analyses (Focus 6-4).

En perspectives, pour que l'abondance de données ne tue pas la qualité des données rendant de fait invisible la biodiversité étudiée, il devient évident d'amplifier les recherches interdisciplinaires associant, la biologie, l'informatique, la bioinformatique et les mathématiques. Pour poursuivre cette révolution dans l'acquisition de connaissances, **l'expertise est un point clef de la réussite** assurant la production, l'administration et l'analyse de données de qualité qui doit passer obligatoirement par une formation en adéquation avec les enjeux actuels. Les innovations technologiques (séquençage de troisième génération, isolement de cellules, capture de gènes, infrastructures informatiques) en association avec le développement d'algorithmes bioinformatiques de plus en plus performants vont sans nul doute se poursuivre et s'accélérer. Ceci permettra de reconstruire et d'analyser de nombreux génomes de référence qui soutiendront une nouvelle vision du monde vivant en tenant compte des interactions et des échanges entre organismes.

FOCUS 6-4

Architectures informatiques en relation avec les données massives de séquençage

Depuis quelques années, la biologie fait face à un déluge de données provenant essentiellement des progrès du séquençage à haut débit. Il est important de distinguer deux niveaux d'infrastructure informatique en génomique (voir le site web e-Infrastructures) : 1) la production des données brutes qui réclame des moyens importants de stockage et d'archivage et des moyens relativement modérés de calcul ; 2) le traitement/service/support bio-informatique permettant de produire des données filtrées à plus forte valeur ajoutée et exploitables par les communautés intéressées. Ces deux niveaux nécessitent des volumes de stockage d'égale importance mais les ressources nécessaires pour le calcul et la création de données temporaires sont beaucoup plus conséquentes pour le second.

La production des données de séquence est actuellement décentralisée sur plusieurs sites académiques en France, dans plusieurs pays étrangers mais aussi sur un certain nombre de plateformes de services privées. La démocratisation des technologies implique à court terme une explosion des données produites, mais également du nombre de lieux de production avec le risque de fragmentation et de duplication des moyens informatiques et des savoir-faire. Une plus grande coordination en termes d'infrastructures et d'investissements apparaît donc nécessaire. La France est en retard dans la structuration de ses infrastructures informatiques ou e-infrastructures pour la génomique par rapport à d'autres pays. Un effort de concentration des moyens sur un nombre restreint de centres régionaux a été initié dans le cadre du Groupement d'Intérêt Scientifique IBISA et doit être poursuivi dans la perspective d'une infrastructure informatique globale cohérente.

Les besoins de calcul pour le traitement des données massives de séquençage relèvent aussi bien du calcul intensif (High Performance Computing ou HPC) que du traitement à haut débit (High Throughput Computing ou HTC) (Wandelt *et al.* 2012). Paradoxalement, les centres de calcul nationaux sont encore peu utilisés par la communauté de la génomique. Cependant, l'utilisation des grilles informatiques a été explorée très tôt par la communauté française de bioinformatique. L'initiative GRISBI « Grilles pour la bioinformatique » adossée au réseau RENABI, au GIS IBISA et au GIS France Grilles a permis de tester l'utilisation de la grille pour répondre de manière plus harmonieuse et collaborative aux enjeux de la bioinformatique, comme par exemple la gestion des données de projets de séquençage de nouvelle génération. Cette étude a mis en avant la nécessité de prendre en compte plusieurs facteurs. Le premier facteur à gérer est le stockage/archivage des données. Les données doivent être déplacées, dupliquées et modifiées sur différents sites : lieu de production, lieu d'archivage, lieu d'analyse, lieu d'exploitation et de visualisation par les biologistes. Une technologie telle que iRODS (Chiang *et al.* 2011) permet un stockage virtuel entre différents sites et ainsi une accession simplifiée et partagée à distance aux données avec une politique de gestion des métadonnées associées. Un deuxième facteur à prendre en compte est la qualité du réseau afin d'assurer un transfert des données rapide et sécurisé entre les différents sites. Un troisième et dernier facteur pour l'exploitation des données de séquençage à haut débit est de fournir aux biologistes la capacité de lancer des chaines de traitement sur des supercalculateurs, des grilles ou des clouds selon leurs besoins.

Pour répondre aux enjeux liés à l'émergence des nouvelles technologies de haut débit en biologie, la France doit se doter d'une e-infrastructure physique inter-organismes de stockage et de calcul évolutive permettant de combiner les fonctionnements d'un cœur national, de centres régionaux ainsi que d'autres acteurs locaux de plus petite taille. La diversité des outils, des données de références et des données à traiter requiert l'accès à des ressources pour le calcul intensif (supercalculateurs) mais aussi pour le traitement à haut débit (clouds, grilles).

RÉFÉRENCES

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31:533-538.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol 11:R106.

Brent MR. 2008. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. Nat Rev Genet 9:62-73.

Caporaso JG, *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335-336.

Chiang GT, Clapham P, Qi G, Sale K, Coates G. 2011. Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute. BMC Bioinformatics 12:361.

Chivian D, Dehal PS, Keller K, Arkin AP. 2013. MetaMicrobesOnline: phylogenomic analysis of microbial communities. Nucleic Acids Res 41:D648-654.

Dillies MA, *et al.* 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform in press.

Edwards RA, Rohwer F. 2005. Viral metagenomics. Nat Rev Microbiol 3:504-510.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. Mol Ecol Resour 11:759-769.

Logares R, Haverkamp TH, Kumar S, Lanzen A, Nederbragt AJ, Quince C, Kauserud H. 2012. Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. J Microbiol Methods 91:106-113.

Namiki T, Hachiya T, Tanaka H, Sakakibara Y. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 40: e155.

Peyretaillade E, *et al.* 2012. Annotation of microsporidian genomes using transcriptional signals. Nat Commun 3:1137.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12:38.

Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F. 2011. Metavir: a web server dedicated to virome analysis. Bioinformatics 27:3074-3075.

Soneson C, Delorenzi M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91.

Temperton B, Giovannoni SJ. 2012. Metagenomics: microbial diversity through a scratched lens. Curr Opin Microbiol 15:605-612.

Wandelt S, Rheinländer A, Bux M, Thalheim L, Haldemann B, Leser U. 2012. Data management challenges in next generation sequencing, Datenbank-Spektrum 12:161-171.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ. 2012. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci 6:427-439.

Yarza P, *et al.* 2013. Sequencing orphan species initiative (SOS): Filling the gaps in the 16S rRNA gene sequence database for all species with validly published names. Syst Appl Microbiol 36:69-73.

SITES INTERNET

e-Infrastructures pour la Génomique et la Biologie à Grande Echelle : http://www.france-grilles.fr/IMG/pdf/E-InfraGeno-Bio-rapport_final.pdf.
Greengenes : http://greengenes.lbl.gov/
RDP : http://rdp.cme.msu.edu/
SILVA : http://www.arb-silva.de/

VII STRUCTURE ET DYNAMIQUE DE LA BIODIVERSITÉ

Coordinateurs : Line Le Gall et Guillaume Lecointre
Contributeurs : Eric Bapteste , Régis Debruyne, Nicolas Puillandre, Jean-François Silvain

La biodiversité est un concept permettant la prise en compte de la structure et de la dynamique du vivant non seulement à l'échelle des espèces, mais également aux échelles infra-spécifique et supra-spécifique. La richesse du concept va presque de soi, quand on sait qu'il y a cinq à dix fois plus de divergence génétique entre deux chimpanzés de l'espèce *Pan troglodytes* qu'entre deux humains les plus distants (figure 7A), et que la variation génétique comprise au sein d'une seule espèce d'angiospermes est en moyenne bien supérieure aux variations génétiques connues au sein d'une espèce de coléoptère. Par ailleurs, les interactions et associations d'espèces en coexistence locale réalisent des communautés qui se structurent elles-mêmes à leur échelle. Bref, compter les « espèces » n'est que la première étape d'une étude de la biodiversité, qui se poursuit par l'analyse des liens, y compris historiques, et des interactions.

Figure 7A. Variabilité génétique intraspécifique chez le chimpanzé et l'homme.

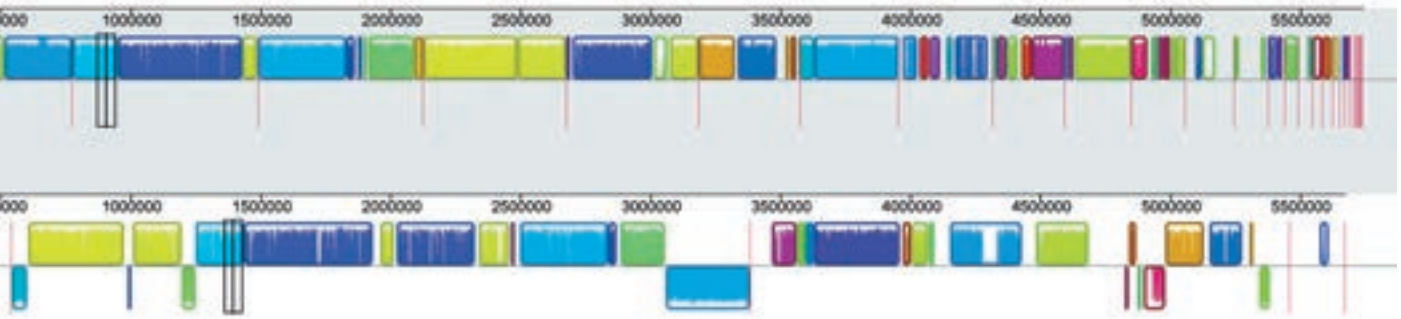
La demande est forte pour que la biodiversité soit d'abord appréhendée en termes de fonctions réalisées dans sa dynamique actuelle au sein des écosystèmes, jusqu'à l'estimation des « services écosystémiques » auxquels elle contribue et qui permettent de communiquer dans un cadre économique, juridique et social. Cette approche est légitime, puisque la première des demandes de la société est de pouvoir prédire comment les changements apportés par les activités humaines, en perturbant les dynamiques du vivant, vont modifier la fourniture des services à l'homme par les écosystèmes. Derrière les fonctions et les services il y a des espèces et si la tendance récente a été de privilégier le rôle écologique de la biodiversité dite commune, éventuellement réductible à un nombre limité de groupes fonctionnels, des travaux récents soulignent au contraire **l'importance des espèces rares**, aux traits de vie remarquables et fonctionnellement vulnérables, dans le fonctionnement des écosystèmes, même les plus diversifiés (voir Focus 11.4 Mouillot *et al.* 2013). Il ne faut donc pas négliger la dimension historique sans laquelle on estime mal l'originalité des structures et des traits des organismes, en termes d'héritage. Dans une savane africaine, un pangolin et un oryctérope sont tous deux des prédateurs d'insectes sociaux, mais en termes de divergence historique et anatomique les pangolins sont plus originaux que les oryctéropes, dont il existe dans le même milieu de proches cousins (damans, éléphants, rats à trompe). L'ordre écologique (Barbault 2006) ne reflète pas toujours l'ordre historique.

La taxinomie est précieuse pour les systématiciens, mais aussi pour ses utilisateurs comme les écologues, les gestionnaires d'espaces protégés et de multiples acteurs industriels dans des domaines aussi variés que les cosmétiques, la pharmacie ou l'agro-alimentaire. Relier la présence, l'absence ou l'abondance d'espèces actuelles correctement identifiées, et des communautés qu'elles constituent, à des paramètres environnementaux est désormais possible grâce aux développements de modèles de diversification et de niches. Il est également possible d'appliquer ces études aux fossiles et aux paléo-environnements et d'en tirer des enseignements pour **l'estimation de la dynamique future de la biodiversité** face au retour de conditions environnementales similaires (Condamine *et al.* 2013, Wappler *et al.* 2012)

La description de la biodiversité a longtemps été considérée comme une tâche quasi-irréalisable, d'autant que l'horizon d'une extinction rapide et massive se rapproche de nous (« sixième extinction »). Or, une évaluation récente (Costello *et al.* 2013) de la magnitude de la biodiversité (à environ 5 à 8 millions d'espèces d'eucaryotes) rend cette tâche de description plus accessible, dans un contexte de nouvelles technologies et de réorganisation des métiers et des pratiques taxinomiques. Parallèlement de **nouveaux compartiments de la biodiversité deviennent accessibles**, tels que les microorganismes non cultivables dont on peut assembler les génomes à partir d'approches métagénomiques (Chap XI - Iverson *et al.* 2012). L'enjeu majeur est de **faire dialoguer la systématique et la génomique environnementale**. La génomique environnementale utilise des numéros de séquences tandis que les pratiques taxinomiques gèrent des noms en lien avec des spécimens et des connaissances associées (y compris des séquences moléculaires). Les deux approches pourraient se nourrir réciproquement de manière à **maximiser la connaissance** sur chaque entité taxinomique détectée.

Des moyens dévolus à cet enjeu dépend aussi le traitement de nombreuses problématiques nécessaires à l'étude de la structure et de la dynamique de la biodiversité, notamment : 1) l'exploitation des arbres phylogénétiques à macro-échelle afin d'en inférer le tempo et les modalités de diversification du vivant au moyen des modèles récemment développés (modèles de diversification, modèles de niches, cf. ecophylogenetics, intégration de la phylogénie dans la démarche écologique, Mouquet *et al.* 2012) ; 2) l'adaptation du schéma théorique de l'arbre du vivant aux modalités d'héritabilité manifestées par les microorganismes (Focus 7-1). La mosaïque génétique des microorganismes et la réticulation permanente de leurs échanges d'ADN implique un renouveau dans l'interprétation des liens (réseaux, ou « forêts d'arbres ») et même des entités détectées.

La taxinomie intégrative (Focus 7-2) réalise aujourd'hui de nouvelles conditions de dialogue entre les approches « traditionnelles » de la taxinomie et les approches moléculaires (notamment barcode multi-marqueurs - Chap VIII) en organisant une boucle de tests mutuels d'hypothèses sur les délimitations d'espèces. La taxinomie intégrative est rendue efficace par l'incorporation, dans la mesure du possible, des séquences du type porte-nom et de diverses données qu'elles soient anatomiques, morphométriques, biochimiques, écologiques, comportementales, etc... Cette approche implique souvent l'obtention et la prise en compte de séquences issues d'ADN dégradés (Focus 7-3), notamment ceux de spécimens de collections ayant qualité de « types ». Les nouvelles techniques de séquençages appliquées à l'ADN ancien permettent d'obtenir des séquences à partir d'un volume extrêmement faible de matériel type.



FOCUS 7-1

Le défi des microorganismes aux systématiciens modernes

Le monde microbien (Figure 7B) incluant les bactéries, les archées et leurs éléments génétiques mobiles, tels que les virus (qu'on les considère comme vivants ou non) est à la croisée de nombreux processus évolutifs. Des processus d'introgression (recombinaisons légitime et illégitime des séquences génétiques, transfert latéral de gènes médiés par une grande diversité de mécanismes) et un processus d'hérédité comme la descendance depuis un ancêtre commun sont conjointement responsables de la grande diversité génétique et génomique de ces entités. L'ampleur des processus d'introgression implique que de nombreuses adaptations dont les évolutionnistes ambitionnent d'expliquer l'origine ont plusieurs sources plutôt qu'une origine commune dans un génome ancestral unique. Le phénotype des organismes est donc en partie découplé de leur hérédité verticale : des organismes proches peuvent présenter des caractères différents quand ils n'ont pas acquis latéralement les mêmes gènes ; des organismes phylogénétiquement éloignés peuvent présenter des caractères communs quand ils ont reçu le même matériel génétique par des processus d'introgression. Ces événements d'introgression impliquent que les génomes de très nombreuses lignées de microorganismes sont mosaïques ou composites sur le plan phylogénétique : tous les gènes d'un génome ne proviennent pas d'un seul et même dernier ancêtre commun ; les composants des génomes proviennent en proportions différentes de différentes sources. De telles entités composites sont mal modélisées par un arbre, dont le formalisme suggère qu'un organisme dont les parties ont différentes histoires provient d'un seul ancêtre plutôt que de plusieurs.

Une conséquence remarquable de cette fluidité génomique est que les membres d'une même espèce bactérienne ne présentent pas tous les mêmes familles de gènes. Seul un ensemble de ces familles (environ 6 % chez les bactéries *Escherichia coli*) est commun à tous les membres de cette espèce ; la somme totale des familles de gènes de cette espèce dépasse la taille du génome individuel moyen d'une *E. coli*. Ce répertoire de gènes utilisés par certains membres de l'espèce, mais pas tous, constitue le pangénome. Plus il est grand, moins on peut généraliser au sujet des caractéristiques communes de l'espèce. De plus, l'idée selon laquelle les microbes vivent isolés a vécu : les microorganismes ont une vie sociale. Les échanges de gènes favorisent l'émergence de partenariats et de phénotypes qui ne peuvent être expliqués que dans le cadre de ces collectifs. Cette observation permet une prise de conscience plus générale : à tous les niveaux d'organisation biologique, des associations de matériel génétique causées par les phénomènes d'introgression engendrent une diversité d'interacteurs dans le monde microbien (associations de bases, de gènes, d'individus…) qui peuvent faire l'objet de la sélection naturelle.

Les approches d'inspiration réductionniste utilisant une « partie » pour faire des inférences sur un « tout » plus grand que ces parties (par exemple le séquençage d'un gène pour décrire l'histoire évolutive d'une espèce) semblent donc trop naïves, certes plus simples mais pas plus justes, pour décrire les origines de la diversité des interacteurs du monde microbien selon un modèle unique. Tous les gènes n'ont pas la même histoire, et l'histoire d'un gène donné n'est pas celle des espèces. Des développements conceptuels (quels objets/interacteurs devraient être étudiés ?) et méthodologiques semblent donc indispensables pour réaliser une ou des systématiques moins abstraites. Les méthodes de NGS et de « single-cell » génomique (voir Focus 11.3) appliquées aux communautés environnementales pourraient ainsi servir à mieux décrire la collection des gènes dans les communautés microbiennes pour inférer les transferts, les partenariats et les adaptations dans le monde microbien.

Figure 7B. L'infiniment petit caché derrière le visible. La biodiversité macroscopique n'est que la partie émergée de l'iceberg des entités capables d'échanger de l'ADN. Un monde de microorganismes mosaïques dans leur composition génétique nous oblige à revoir ce que nous appelons des « espèces » au sein des microorganismes tels que les archées et bactéries, ainsi que le schéma théorique général de « l'arbre du vivant » les concernant.

FOCUS 7-2

Dans le cadre de la taxinomie intégrative, quelle stratégie d'échantillonnage pour étudier la structure et la dynamique de la biodiversité ?

Si l'approche intégrative est maintenant largement adoptée par la communauté des taxinomistes pour délimiter les taxons, son application reste encore souvent problématique. Le débat n'est plus sur la nécessité d'intégrer plusieurs types de caractères (moléculaires, morphologiques, écologiques…) et d'appliquer plusieurs critères de délimitations d'espèces (phénétiques, biologiques, phylogénétiques), mais plutôt sur le poids relatif de ces différents caractères et critères (Figure 7C). Cependant, quelque soit le type de caractère, le critère appliqué, la méthode d'analyse employée, et l'ordre dans lequel ils sont étudiés, la qualité de l'échantillonnage restera critique pour la qualité des hypothèses taxinomiques et des inférences sur la structure et la dynamique de la biodiversité qui seront proposées par la suite.

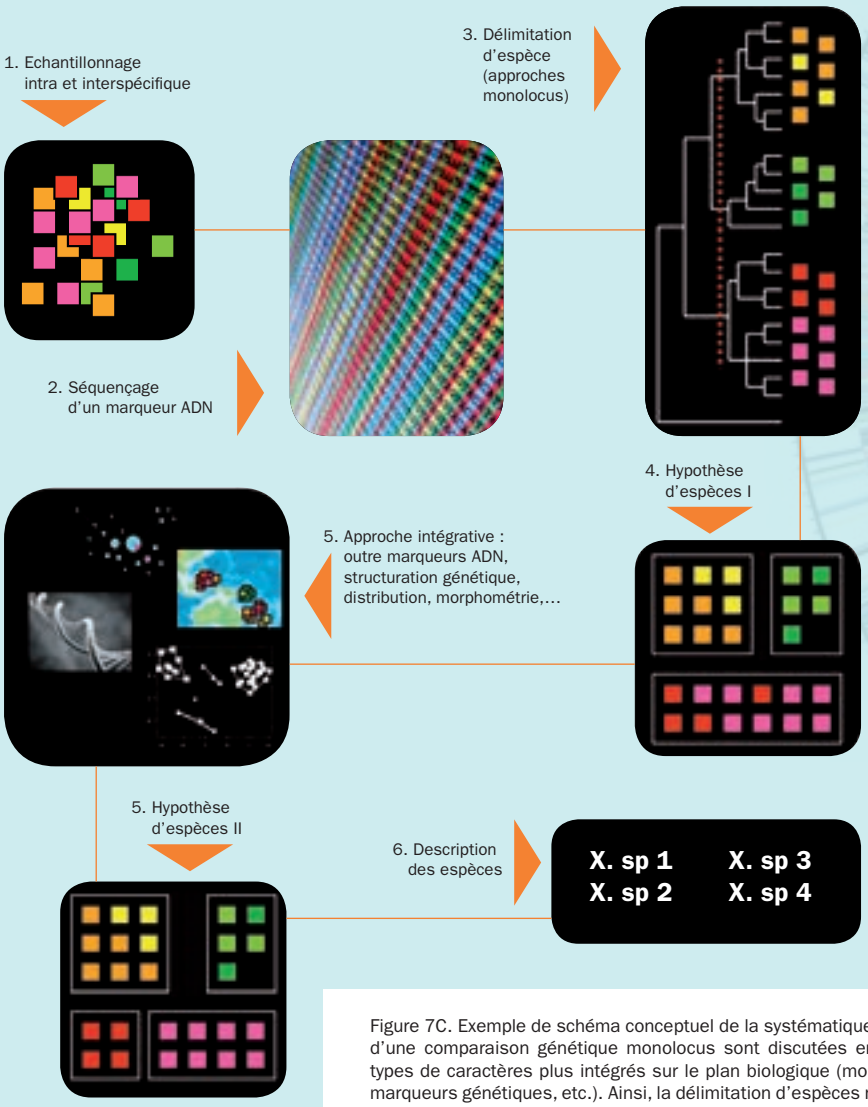


Figure 7C. Exemple de schéma conceptuel de la systématique intégrative. Des hypothèses d'espèces issues d'une comparaison génétique monolocus sont discutées en tenant compte d'autres critères et d'autres types de caractères plus intégrés sur le plan biologique (morphométrie, distribution, comportement, autres marqueurs génétiques, etc.). Ainsi, la délimitation d'espèces n'est pas réduite ni à un seul critère phénétique au niveau génétique, ni à la description d'un seul trait anatomique, morphométrique ou comportemental. Non seulement la description d'espèces adopte des démarches comparatives formalisées et donc testables, mais elle intègre en plus des informations issues de caractères ayant différentes vitesses d'évolution et différents niveaux d'intégration biologique.

FOCUS 7-2 (suite)

Selon les critères utilisés, définir les limites entre espèces au sein du réseau généalogique consiste à identifier des différences phénétiques entre groupes d'organismes, à déterminer si ces groupes constituent des lignées évolutives distinctes et à tester l'absence d'échange de gènes entre ces lignées. Chacun de ces critères repose sur une estimation des diversités intra et interspécifiques basée sur un échantillonnage d'un grand nombre de spécimens au sein de chaque espèce et d'un grand nombre d'espèces au sein de chaque ensemble étudié (taxon, communauté…). Même s'il est difficile de définir un nombre de spécimens à échantillonner par espèce qui assurerait une estimation correcte de ces diversités (l'échantillonnage d'une espèce de vertébré de grande taille endémique d'une île océanique ne nécessitera pas le même effort qu'un mollusque benthique présent dans l'ensemble de l'Indo-Pacifique), la taille des jeux de données constitués pour les approches de taxinomie intégrative se rapproche de plus en plus de celle des jeux de données de type « génétique des populations ». Au niveau interspécifique, la tendance est la même : au sein du groupe étudié, plus le nombre d'espèces intégrées à l'analyse est grand, plus l'estimation de la diversité interspécifique sera représentative, et plus les hypothèses taxinomiques seront robustes.

Dans ce contexte, les analyses moléculaires sont maintenant régulièrement intégrées aux approches de taxinomie, notamment via l'essor des projets de types « barcoding ». Comme pour les spécimens, l'échantillonnage des marqueurs génétiques au sein des génomes n'est pas à négliger : un des critères principaux de délimitation d'espèces reste la capacité (ou l'incapacité) à se reproduire, et donc à échanger (ou non) des gènes, critère qui ne peut être testé que si plusieurs marqueurs génétiques non liés sont analysés. Chez de nombreux organismes, l'identification et la caractérisation de tels marqueurs (facilement accessible et variable au niveau spécifique) est souvent problématique, de part notre manque de connaissances sur les génomes de nombreux organismes.

Les approches NGS offrent potentiellement des solutions à ces problèmes car elles permettent de séquencer un grand nombre de spécimens pour un grand nombre de marqueurs génétiques. Elles ne nécessitent pas, pour certaines, l'identification des gènes-cibles et donc de connaissances a priori du génome, un avantage évident pour les taxons dont les génomes, la diversité et la taxinomie restent peu connus. De plus, les méthodes de type RAD-seq* assurent une couverture* du génome largement supérieure aux approches « gène-centré » et permettent ainsi de s'affranchir des problèmes liés au séquençage d'un ou de quelques marqueurs dans le cadre des approches de taxinomie. Des adaptations méthodologiques sont encore nécessaires afin d'accompagner la transition du séquençage de type Sanger vers les NGS.

Pour les eucaryotes, les approches de type barcode requièrent la mise en place d'une **bibliothèque de séquences de référence** en lien avec des spécimens préservés dans des **collections d'Histoire Naturelle** (voir aussi chap XI pour les microorganismes). Les NGS posent encore des problèmes de traçabilité dans une logique clairement orientée sur des spécimens. Ainsi, la technologie de séquençage par la méthode de Sanger reste encore aujourd'hui celle qui permet de garantir une traçabilité spécimen/séquence. Des efforts méthodologiques sont à entreprendre pour accompagner la transition NGS.

Les NGS sont très prometteuses dans l'analyse de la composition des communautés, en particulier celles qui sont les plus menacées à court terme par les changements globaux, mais elles risquent, de par leurs contraintes techniques, de conduire à une **taxinomie des OTU*** (voir focus 11.4) **désincarnée** et découplée des noms de taxons et des connaissances associées. Cela pourrait paraître acceptable pour certaines évaluations quantitatives de la biodiversité, face notamment à des urgences environnementales (déforestation, changement d'usage des terres par exemple). Il faut dans les autres cas, et en

FOCUS 7-3

Les ADN dégradés dans l'analyse de la biodiversité

La dégradation de l'ADN définit un continuum d'échantillons où le temps n'est qu'un facteur, et pas toujours le plus déterminant. Ainsi, le contenu moléculaire des os d'un mammouth pris dans les glaces de Sibérie depuis plusieurs dizaines de milliers d'années est souvent en bien meilleur état que celui d'un squelette d'éléphant conservé dans les collections d'un musée d'Histoire Naturelle depuis une dizaine d'années. Une vision holistique des ADNs dégradés convie aujourd'hui aussi bien le paléogénéticien que le généticien des populations, le systématicien néontologiste ou le biologiste de la conservation à travers l'exploitation des ADNs des échantillons non invasifs (fèces, phanères, mucus) réalisés à partir d'organismes modernes tels que des échantillons environnementaux (sols, eau douce ou marine), mais aussi des spécimens de collection de musées. Ces ADNs présentent en effet des caractéristiques fondamentales communes avec les ADNs dits anciens : faible concentration moléculaire, dégradation par hydrolyse et oxydation, contamination exogènes, inhibition, etc.
Même si elles n'ont pas initialement été développées dans le but de satisfaire à ces besoins, les NGS offrent de nouvelles opportunités pour l'analyse de ces ADNs dégradés : elles offrent l'accès au séquençage massif de molécules d'ADN de taille réduite (moins de 400 nucléotides pour la majorité des plateformes actuelles), naturellement prédominantes dans les échantillons dégradés, tout en s'abstrayant de l'étape fastidieuse et peu productive d'amplification ciblée par PCR. Elles ont ainsi permis l'accès au séquençage du génome nucléaire largement prévalent (en quantité pure d'ADN) chez les restes eucaryotes dégradés mais quasi-inexploitables par PCR ciblée du fait de la dilution du matériel moléculaire (en nombre de copies génomiques) dans ces échantillons. Enfin, elles permettent l'immortalisation de banques d'ADNs dégradés réalisées à partir d'échantillons de taille limitée difficiles voire impossibles à reproduire.

L'adéquation technique au matériel dégradé ne constitue que la face émergée de l'iceberg tant il est vrai que les développements conceptuels sur les ADNs dégradés accompagnent les innovations techniques continuelles associées aux NGS. Les enseignements à tirer de l'analyse des ADNs dégradés au sens large sur la structure et la dynamique de la biodiversité recouvrent plusieurs dimensions temporelles. L'analyse de la dynamique phylogéographique d'espèces en danger d'extinction ou au contraire de risques d'invasions est largement facilitée par l'accès à un matériel dégradé non invasif ou environnemental ne nécessitant pas même l'échantillonnage direct des taxons d'intérêt souvent élusifs. Ancrée dans un passé plus distant, la biogéographie diachronique des espèces modernes ou éteintes éclaire d'un jour nouveau leur dynamique actuelle. Entre ces deux extrêmes de temps, les richesses des collections des musées d'Histoire Naturelle, patiemment constituées à travers les décennies, fournissent un matériau de premier choix pour documenter au plus près la structure historique de la biodiversité pourvu que l'on s'attèle à leur offrir une nouvelle jeunesse par le biais d'une analyse moléculaire dont on observe aujourd'hui les enthousiasmantes prémices (Figure 7D).

Figure 7D. Le grand Herbier rénové du Muséum national d'Histoire Naturelle. Les collections nationales d'Histoire Naturelle peuvent être considérées comme une extraordinaire ressource de séquences d'ADN… moyennant des technologies de séquençage adaptées aux ADNs dégradés.

particulier si l'on veut comprendre l'origine et la dynamique de la biodiversité d'un biotope ou d'un écosystème particulier (Figure 7E), aller vers un métabarcoding éclairé, c'est-à-dire assurant le lien entre séquences et connaissances associées au nom via le « voucher specimen ». Ceci ne peut être réalisé qu'en maintenant parallèlement les compétences et les recrutements sur des approches dites « traditionnelles », à savoir :

- les compétences en anatomie comparée
- les compétences en taxinomie
- les compétences en bio-informatique (volume considérable de données à traiter, à stocker et à partager, y compris les données issues des nouvelles technologies d'imagerie)
- les capacités logistiques et les compétences nécessaires à l'exploration de la biodiversité sur le terrain
- les capacités logistiques et les compétences nécessaires à la conservation dans les collections d'Histoire Naturelle considérées comme de grandes infrastructures de recherche.

Figure 7E. Atelier d'inventaire marin « Guadeloupe » au cours de la mission KARHUBENTHOS. Aujourd'hui, les opérations d'exploration et d'inventaire permettent la caractérisation de la biodiversité d'un écosystème tout en garantissant la traçabilité du lien entre chaque spécimen (qui sera mis en collections nationales) et ses séquences d'ADN.

RÉFÉRENCES

Barbault R. 2006. Comme un éléphant dans un jeu de quilles. Ed Seuil.

Condamine F, Rolland J, Morlon H. 2013. Macroevolutionary perspectives to environmental change. Ecol Lett 16:72-85.

Costello M, May R, Stork NE. 2013. Can we name Earth's species before they go extinct? Science 339:413-416.

Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust EV. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. Science 335:587-590.

Mouillot D, *et al.* 2013. Rare species support vulnerable functions in high-diversity ecosystems. PloS One 11:e1001569.

Mouquet N, *et al.* 2012. Ecophylogenetics: advances and perspectives. Biol Rev 87:769-785.

Wappler T, Labandeira CC, Rust J, Frankenhäuser H, Wilde V. 2012. Testing for the effects and consequences of mid paleogene climate change on insect herbivory. PloS One 7:e40744.

VIII CARACTÉRISER LA DIVERSITÉ DU VIVANT

Coordinateurs : François Pompanon et Sarah Samadi
Contributeurs : Régis Debruyne, Frédéric Delsuc, Catherine Hänni , Sébastien Lavergne, Morgane Ollivier, Eric Pante, Nicolas Puillandre, Jean-Yves Rasplus, Pierre Taberlet

Des grands projets exploratoires fondés sur le séquençage d'ADN ont démarré dans les années 1990 (e.g. génome humain) suivi une dizaine d'années plus tard par le projet Barcode of Life (BoL). Ce projet visait à fournir un outil de diagnostic universel de la diversité spécifique utilisable dans différents domaines (e.g. écologie, agronomie, réglementation douanière, etc…) mais aussi à accélérer la description de la biodiversité encore inconnue. Le DNA-barcoding repose sur l'obtention de données génétiques standardisées (i.e. des code-barres ADN) à partir de spécimens référencés dans des collections et identifiés par des taxinomistes, ce qui assure l'application des nomenclatures biologiques (Puillandre *et al.* 2011). Le métabarcoding est un prolongement de l'approche DNA-barcoding : son but est de capter la biodiversité d'un échantillon environnemental (sol, eau, contenu digestif…), il utilise fréquemment d'autres standards génétiques (Taberlet *et al.* 2012).

La première limitation de ces deux approches est la complétion en termes de **couverture taxinomique des banques de données de référence**. La seconde concerne **le choix de standards génétiques** offrant un niveau de résolution permettant de répondre aux différents questionnements et de couvrir la diversité du vivant. Les NGS permettent d'envisager de lever cette limitation (Figure 8A) en facilitant l'obtention d'un plus grand nombre de marqueurs, couvrant plus largement les génomes et offrant ainsi une plus grande gamme de résolution pour répondre aux différents questionnements (Focus 8-1).

Contrairement aux approches caractérisant des échantillons environnementaux (**métagénomique, métabarcoding**), le **DNA-barcoding *sensu stricto*** s'est plus tourné vers l'effort de complétion et la pertinence taxinomique que vers les développements permis par les NGS. En France, plusieurs laboratoires utilisent les NGS pour décrire la diversité du vivant dans des études systématiques, phylogénétiques et écologiques. Ces disciplines ont leurs questionnements et souvent leurs outils propres, mais il est essentiel que les résultats produits par l'une puissent être exploités par les autres. Par exemple, les bases de référence mettant en relation information génétique et taxinomique doivent pouvoir être utilisées pour caractériser la diversité des écosystèmes ; la découverte de nouveaux compartiments de biodiversité dans des études écologiques doit pouvoir être rapidement prise en compte dans les études taxinomiques et phylogénétiques. Un enjeu majeur est donc le développement de **méthodes compatibles pour la systématique et l'écologie**.

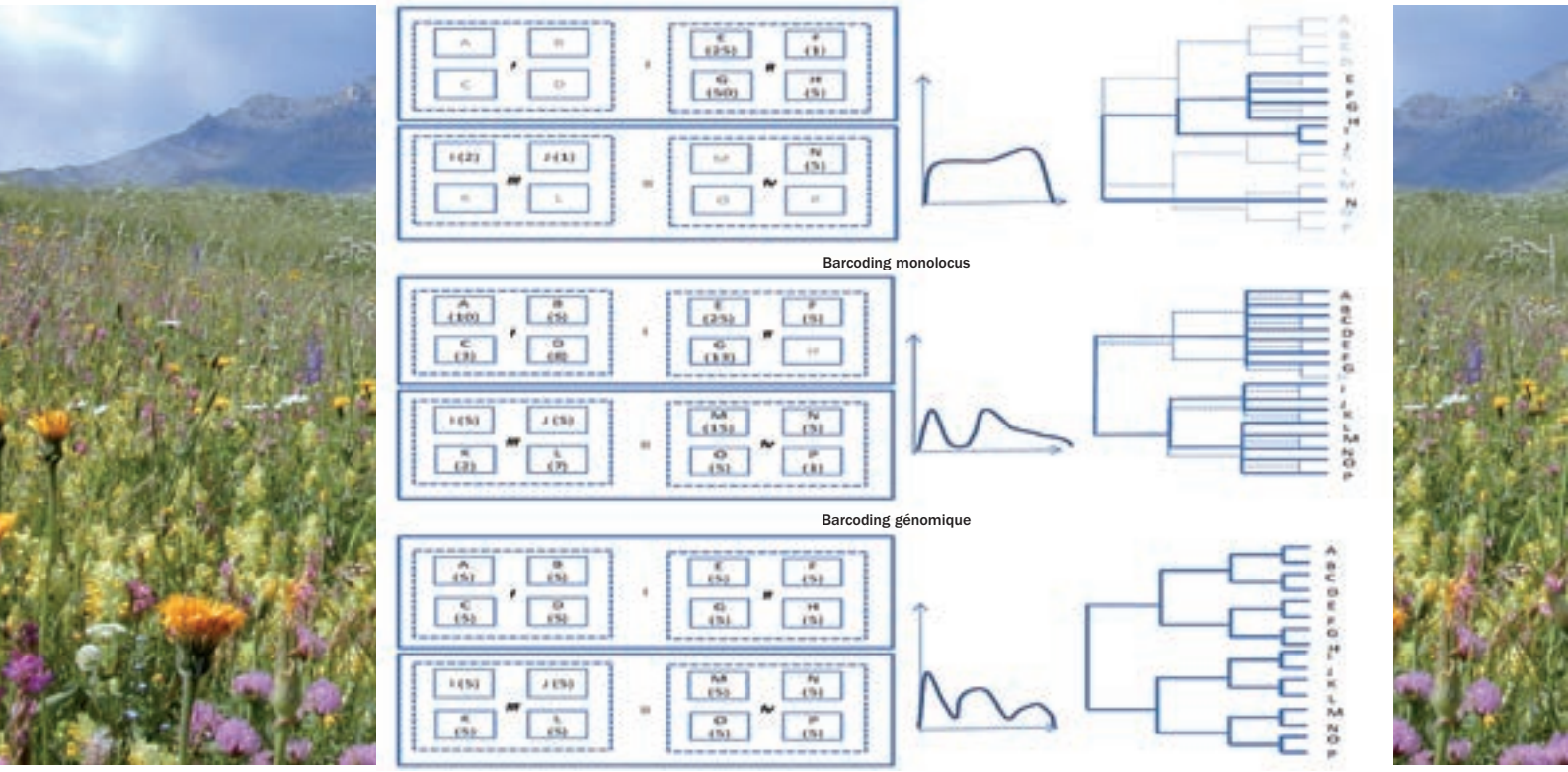


Figure 8A. Echantillonnage, barcoding et classification taxinomique à l'ère des NGS. De gauche à droite, la classification acceptée d'un taxon donné, le nombre de spécimens échantillonné pour le rang le plus inclusif est indiqué entre parenthèse (i.e. espèce), la distribution des distances obtenues en fonction du marqueur choisi et la phylogénie inférée (en pointillé la phylogénie résolue correspondant à la taxinomie acceptée). En haut : situation sans plan d'échantillonnage taxinomique et choix des marqueurs guidé par la disponibilité des données (e.g. données extraites de Genbank), les espèces sont échantillonnées en fonction d'autres études (e.g. espèces modèles, espèces emblématiques, etc ...), la distribution des distances génétiques n'est pas clairement multi-modale, les marqueurs génétiques ne sont pas résolutifs à toutes les profondeurs de l'arbre. Au milieu : situation du type DNA-barcoding classique (monolocus) : la plupart des espèces connues du taxon sont échantillonnées, généralement avec plusieurs spécimens par espèce, les marqueurs génétiques utilisées ne sont discriminants qu'à un rang donné (e.g. rang espèces) et la distribution des distances génétiques est bimodale. En bas : situation idéale (barcoding génomique) : toutes les espèces du taxon sont échantillonnées avec le même nombre de spécimens, les marqueurs génétiques utilisés couvrent largement le génome et permettent de reconstruire une phylogénie robuste à tous les rangs taxinomiques, la distribution des distances génétiques est multimodale.

La difficulté réside dans les besoins contradictoires de la standardisation des données de référence, de l'étude de la diversité des organismes et des contraintes techniques liées aux échantillons étudiés. Il est en effet difficile de concilier les propriétés nécessaires pour qu'un code-barre ADN soit à la fois performant pour les études taxinomiques et écologiques (Valentini *et al.* 2009). Les très grands nombres de séquences fournis par les NGS rendent envisageable le **développement de code-barres multi-locus** (combinaisons de nombreux fragments courts) couvrant mieux les génomes et utilisables dans différents types d'études. Les fragments courts permettent de caractériser des ADN environnementaux dégradés, le grand nombre de marqueurs fournissant suffisamment de caractères moléculaires pour avoir une bonne résolution taxinomique et un signal phylogénétique conséquent. Par ailleurs, les fragments d'ADN sont actuellement caractérisés par séquençage après amplification par PCR. Cette étape est génératrice de biais et d'erreurs, et contraint fortement la définition des code-barres ADN à des régions encadrées par des sites conservés pour permettre la fixation des amorces. Plusieurs approches basées sur les NGS permettraient de s'affranchir de ces problèmes : 1) les **méthodes de capture** (voir chapitre XI) permettent, sans recours à la PCR, de sélectionner pour le séquençage ultérieur des code-barres cibles grâce à des sondes complémentaires de régions conservées incluses dans ces fragments ; 2) les méthodes basées sur le **RAD-seq** (Focus 8-2), bien qu'inutilisables pour caractériser les ADN dégradés issus d'échantillons environnemen-

FOCUS 8-1

Les défis bioinformatiques de l'identification taxinomique

Sous le terme « identification taxinomique » sont regroupées deux approches centrales en taxinomie, mais bien distinctes et qui font appel à deux catégories de méthodes. Il faut distinguer d'un côté les approches d'assignation d'un spécimen inconnu à un groupe taxinomique donné – idéalement une espèce (« specimen identification ») et de l'autre les approches de délimitation d'espèce (« species discovery/delimitation ») ou plus généralement d'inférence phylogénétique pour les rangs taxonomiques supérieurs à l'espèce.

Dans les deux cas, et avant de s'intéresser aux aspects méthodologiques, il est nécessaire de rappeler l'importance cruciale de l'échantillonnage taxinomique. Ces approches reposent toutes en effet sur une évaluation pertinente de la variabilité intra et interspécifique (ou, plus généralement, intra et intergroupes), et toutes les méthodes actuellement disponibles sont sensibles à la qualité de l'échantillonnage. Afin de ne pas sous-estimer la diversité intra-groupe ou surestimer la diversité inter-groupes il faut assurer à la fois une couverture suffisante au sein des groupes mais également entre les groupes. Ainsi, pour le niveau spécifique, il s'agira, dans l'idéal, 1) de couvrir l'ensemble de l'aire de distribution de l'espèce, en incluant dans le jeu de données plusieurs spécimens par populations, et 2) de s'assurer que l'ensemble des espèces du genre étudié sont incluses dans le jeu de données. La conséquence directe de ce plan d'échantillonnage est la nécessité de prévoir une stratégie de séquençage NGS qui permette la prise en compte d'un (très) grand nombre de spécimens à analyser.

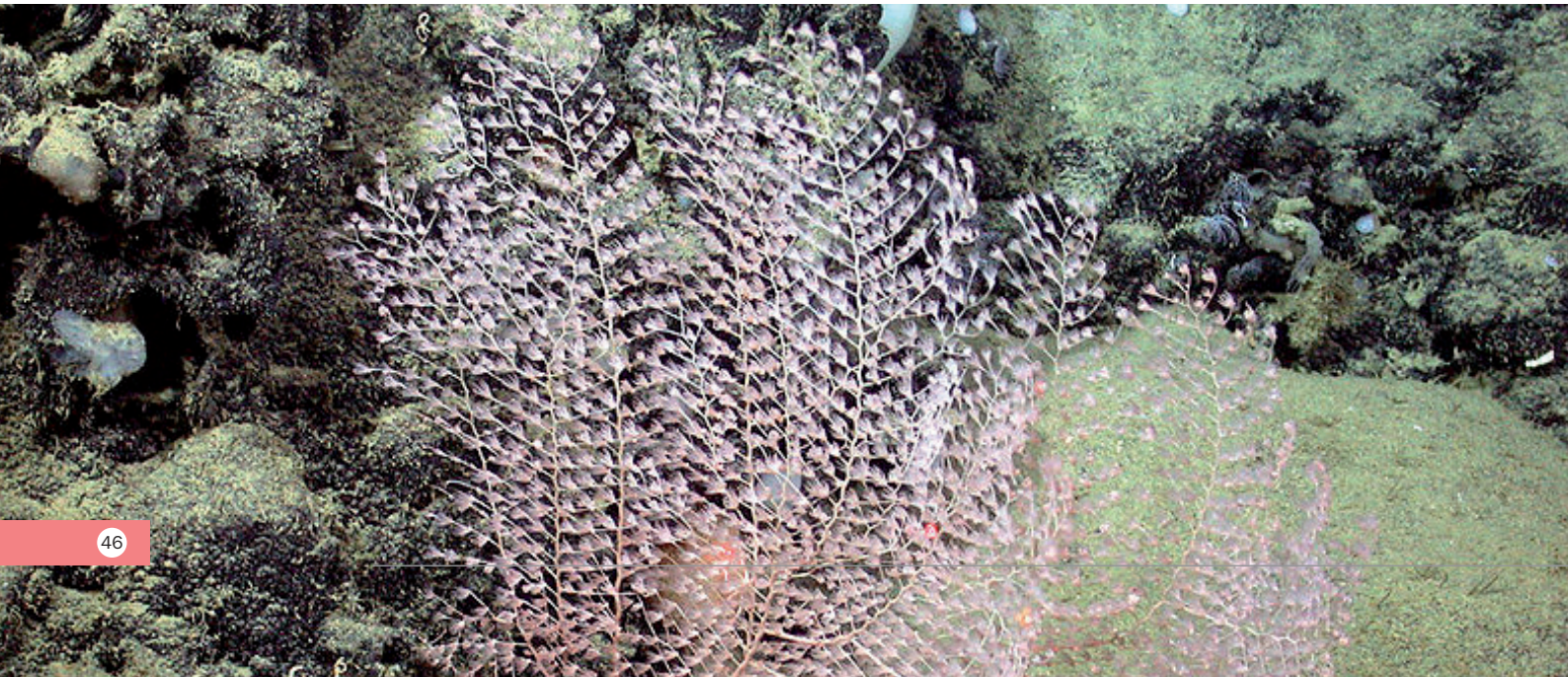
D'un point de vue méthodologique, la communauté des taxinomistes moléculaires est très active et de nouvelles méthodes d'assignation ou de délimitation d'espèces, en plus des approches de phylogénie (total evidence, supertree) devenues maintenant des classiques, sont proposées régulièrement. Elles peuvent être monolocus et sont dans ce cas adaptées pour des jeux de données incluant de nombreux spécimens, ou multilocus et sont alors limitées à un échantillonnage réduit de spécimens. Ce cadre méthodologique doit maintenant s'adapter à deux contraintes particulières liées aux approches de NGS telles qu'elles pourraient être développées dans les approches d'assignation (barcoding, metabarcoding) ou de délimitation d'espèces ou de taxons de rangs supérieurs. Tout d'abord, l'utilisation de NGS induit dans la plupart des cas un taux d'erreurs plus important qu'avec la méthode Sanger. Ce biais potentiel doit être pris en compte dans la conception de l'expérience, en assurant un équilibre entre nombre de locus séquencés, nombre de spécimens séquencés et couverture minimum pour chaque fragment séquencé. Ensuite, les méthodes actuelles de délimitation d'espèces sont basées soit sur une approche monolocus pour un grand nombre de spécimens, soit sur des approches multilocus pour un nombre de spécimens moyens et peu de marqueurs : il est maintenant nécessaire de développer des méthodes qui combinent les avantages des deux approches (beaucoup de marqueurs et beaucoup de spécimens) qui permettront d'analyser les données issues par exemple d'approche de type RAD-seq.

taux, pourraient être performantes pour la délimitation des espèces et la phylogénie (Cariou *et al.* 2013) ; 3) le séquençage massif direct permet d'assembler les génomes complets des organelles (mitochondries et chloroplastes) et des gènes présents en grand nombre de copies dans le génome (gènes ribosomiques…) qui sont utilisés pour établir les phylogénies à large échelle (Focus 8-3). Appliquée à des échantillons écologiques, cette approche métagénomique permet de caractériser simultanément les diversités taxinomique et fonctionnelle (Focus 8-4). En plus de l'information fournie par les régions ciblées, cette approche produit des millions de courtes séquences nucléaires qui contiennent aussi une information pouvant être exploitée pour des assignations fonctionnelles, taxinomiques ou des phylogénies.

Outre les difficultés techniques liées à l'utilisation des NGS (erreurs de séquençage, difficultés d'interprétation des données en termes quantitatifs à partir d'échantillons environnementaux), plusieurs limitations conceptuelles et méthodologiques existent. L'utilisation de code-barres multilocus permet une approche moins réductrice de la définition des taxons en se basant sur un plus grand nombre de caractères non liés. Cependant, des différences persisteront probablement à l'échelle de larges spectres taxinomiques notamment en raison de l'**absence de génomes complets de référence pour des compartiments entiers de l'arbre du vivant**. De plus, se pose la question de l'homologie des marqueurs entre organismes comparés et donc celle de la profondeur phylogénétique du signal obtenu. Ces approches produisent, au moins pour partie, des **données non standardisées** dont le traitement dans des approches comparatives est non trivial (e.g. quelques millions de courtes séquences représentant une fraction variable d'un génome pas toujours comparable d'un organisme à l'autre). Le test des hypothèses sur l'homologie des marqueurs ne peut être réalisé qu'avec un échantillonnage taxinomique pertinent en fonction de la profondeur phylogénétique envisagée (Focus 8-2). La question de la complétude des données de référence, et donc l'impact des données manquantes sur la qualité des inférences, qui se pose déjà quand seuls quelques marqueurs standardisés sont utilisés, est ici encore accrue.

Dans ce contexte, deux principaux défis sont à relever. Il faut d'une part **compléter les bases de référence** en augmentant leur couverture taxinomique. Cela concerne les barre-codes standards, des régions standardisées (e.g. génomes mitochondriaux complets, Dettai *et al.* 2012) mais aussi toutes les données génomiques non standardisées pour lesquelles cette couverture taxinomique est indispensable pour résoudre les questions d'homologie (génomes partiels, données de RAD-seq…). Ainsi, la réussite de ces nouvelles approches est conditionnée par le développement de **méthodes d'assignation taxinomique et d'inférence phylogénétique** prenant en compte les erreurs de séquençage, la dégradation de l'ADN notamment dans les études paléogénomiques (Focus 8-4), et exploitant au maximum l'information issue des données disponibles pour les marqueurs standards et/ou les données génomiques partielles (Coissac *et al.* 2012, Focus 8-5).

La réussite de ces développements nécessite des actions coordonnées à l'interface entre taxinomie, écologie, biologie moléculaire et bio-informatique, via le financement de **projets pilotes** ciblant par exemple le développement de protocoles et de méthodes d'analyses compatibles avec les études systématiques et écologiques. Il est également essentiel de développer des bases de **données de référence cohérentes avec les approches NGS**, faisant le lien entre données génétiques, spécimens de collections et collections d'ADN durables qui soient mobilisables dans le futur pour les nouveaux développements technologiques. Dans ce cadre, un soutien particulier devrait être accordé aux études taxinomiques ciblant les groupes d'intérêt phylogénétique ou écologique.



FOCUS 8-2

Du code-barre ADN vers les NGS : développement de marqueurs RAD-tag chez la gorgone *Chrysogorgia*

La connaissance de la diversité taxinomique des organismes est un préalable nécessaire aux projets de génomique environnementale. Ces projets requièrent que des bases de données de référence permettant de rattacher les données de génomique environnementale à des organismes nommés soient disponibles. Dans ce contexte, la systématique fait face conjointement à la magnitude de la biodiversité et à la difficulté d'appliquer des méthodes de génomique sur des organismes non modèles. L'objectif de cette étude est de tester, sur un taxon pilote, les possibilités ouvertes par la mise en œuvre des nouvelles technologies de séquençage.

Le taxon choisi est le genre *Chrysogorgia*, qui avec environ 60 espèces décrites, est l'un des genres d'octocoralliaires les plus diversifiés. C'est un genre monophylétique largement distribué géographiquement dont la variabilité morphologique coïncide avec la variation génétique observée pour le marqueur mitochondrial mtMutS. Cependant la diversité de ce gène est faible et les morphotypes peuvent ne différer que par une seule base. L'objectif est de valider l'hypothèse selon laquelle ces morpho-haplotypes sont bien du rang de l'espèce. Pour cela, le RAD-seq est utilisé afin d'obtenir un grand nombre de marqueurs variables au sein et entre les morpho-haplotypes et de définir des entités effectivement connectées par flux de gènes.

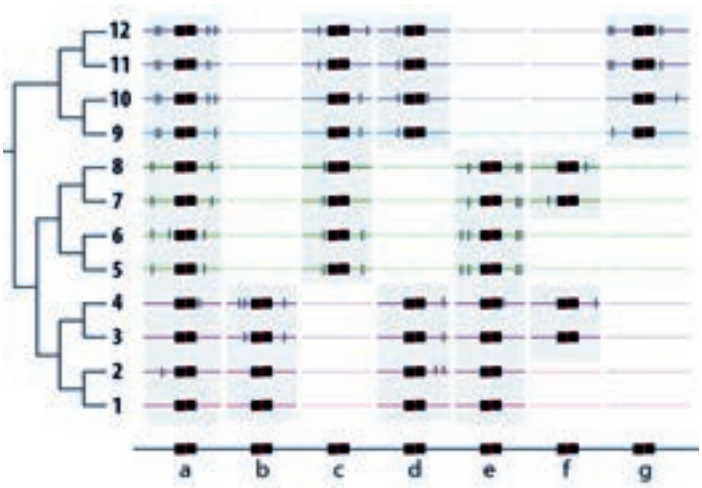


Figure 8B. La méthode RAD-Tag génère un grand nombre de marqueurs polymorphes entre individus. Le plan d'échantillonnage taxonomique permet d'inférer les groupes d'homologie et ainsi de définir l'usage qu'il peut être fait des marqueurs en fonction de la profondeur du signal phylogénétique. Le marqueur a est ainsi présent et homologue entre tous les spécimens inclus dans l'étude alors que d'autres marqueurs ne sont informatifs que pour certains clades (e.g. d et g ne sont informatifs que pour le premier clade).

Au-delà de la technique NGS utilisée, la pertinence de ce travail repose sur une stratégie d'échantillonnage permettant de tester des hypothèses taxinomiques à différentes profondeurs phylogénétiques. Cette stratégie consiste à inclure : 1) différents morpho-haplotypes échantillonnés en sympatrie afin de vérifier l'absence de flux de gènes à l'échelle génomique, 2) des spécimens d'un même morpho-haplotype échantillonnés dans des localités distantes ou différentes profondeurs afin de mettre en évidence une éventuelle diversité cryptique, 3) des spécimens de différents clades mitochondriaux afin de valider le signal phylogénétique de ce marqueur.

La méthode RAD-seq génère un grand nombre de séquences pouvant comporter des artefacts liés à la technique. L'analyse des données passe par une phase de traitement bio-informatique qui permet de trier les séquences en fonction de leur qualité, d'assembler les séquences identiques pour un même individu puis de les comparer entre individus, tout en évaluant l'homologie des marqueurs. Les mesures préliminaires de divergence génétique entre individus vont dans le sens des décisions taxinomiques prises sur la base des données mitochondriales puisque la divergence intra-haplotypique est nettement plus faible que la divergence inter-haplotypique. Un des atouts majeurs de la méthode, à condition qu'elle soit associée à un échantillonnage taxinomique pertinent, est de fournir simultanément un grand nombre de marqueurs et le cadre comparatif nécessaire pour leur interprétation. Cette méthode a de nombreuses applications telles que : la caractérisation à très fine échelle spatiale de la structure des populations, la détection de régions sous pression de sélection, ou encore la reconstruction phylogénétique.

FOCUS 8-3

PhyloAlps : séquençage nouvelle génération de la flore du biome alpin, vers des megaphylogénies haute-résolution et de nouvelles librairies* de metabarcoding

L'utilisation de données génomiques dans l'étude de la biodiversité a connu un essor sans précédent au cours de la dernière décennie, en particulier : 1) pour la reconstruction de phylogénies toujours plus grandes (mégaphylogénies), permettant de reconstruire la diversification de grands clades et l'histoire de mise en place des points chauds de diversité ; et 2) pour le développement de nouvelles approches de caractérisation des patrons de biodiversité, à partir d'ADN environnemental contenu dans les sols (technique de metabarcoding). Le projet collaboratif Phylo-Alps a pour objectif de constituer une base génomique de référence pour un biome entier, l'Arc Alpin, afin de développer des travaux de phylogénies, de génomique comparative et de metabarcoding dans cette région.

Le projet se base sur un effort d'échantillonnage systématique de l'ensemble de la flore de l'Arc Alpin en prenant pour référence la révision de Flora Alpina. L'objectif est de caractériser 4500 espèces/sous-espèces avec 1 à 2 échantillons par taxon et de réaliser un herbier de référence contenant les plantes séquencées. Le travail d'échantillonnage, débuté en 2009, a permis de récolter à ce jour des échantillons pour 85 % des taxons. L'échantillonnage de familles considérées prioritaires (Campanulaceae, Saxifragaceae, Primulaceae) est pratiquement complet sur l'ensemble de l'arc alpin (99 %). Le séquençage de nouvelle génération se base sur la technologie Illumina HiSeq, pour réaliser un séquençage très basse couverture du génome de chaque espèce récoltée (0,1 X). La phase de séquençage est en cours, après une étude pilote (20 espèces sur une ligne de HiSeq 2000) réalisée au Génoscope et financée par l'appel d'offre APEGE 2012. L'objectif est de reconstruire le génome chloroplastique et mitochondrial des espèces étudiées et d'extraire les séquences des gènes nucléaires répétés (gènes ribosomiques notamment), à l'aide de programmes d'assemblage dédiés à ce type de données qui sont en cours de développement. Les données produites permettront de réaliser des mégaphylogénies à partir de données génomiques identiques sur un grand nombre d'espèces, s'affranchissant ainsi des problèmes de données manquantes typiques des approches de supermatrices actuellement utilisées (Figure 8C).

Figure 8C. Elaboration d'une base génomique de référence pour un biome entier, l'Arc Alpin. (a) Un exemple de mégaphylogénie des 823 genres végétaux présents dans l'Arc Alpin réalisée avec la méthode mixte supe-rarbre – supermatrice (Roquet *et al.* 2013). Exemples de milieux et espèces remarquables de l'Arc Alpin échantillonnés lors du projet PhyloAlps : (b) Prairie à rhododendron ferrugineux, Parc National des Ecrins, (c) Prairie subalpine à forte diversité spécifique, col du Lautaret, (d) *Androsace helvetica*, crête du Galibier (~2800m a.s.l.), (e) *Eritrichiumnanum*, Aile froide Occidentale (~3500m a.s.l.).

FOCUS 8-4

Evolution convergente du microbiome intestinal chez les mammifères myrmécophages

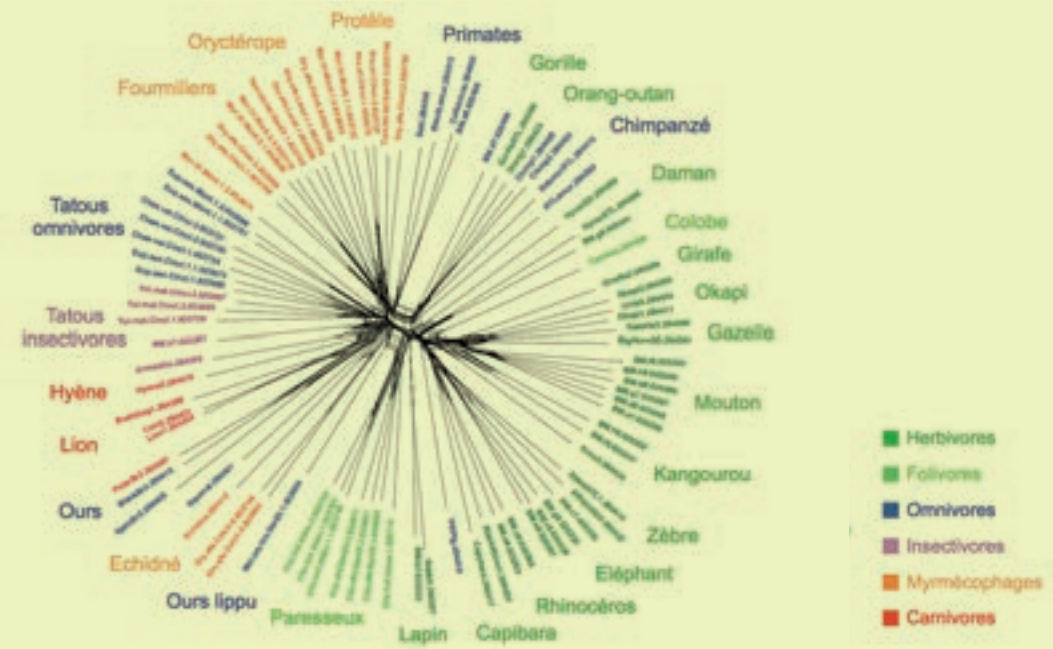


Figure 8D. Réseau phylogénétique représentant les relations entre les microbiomes intestinaux de différents mammifères sur la base de leur composition taxinomique en bactéries et archées.

L'objectif principal est de caractériser et comparer la diversité taxinomique et fonctionnelle du microbiome intestinal (i.e. l'ensemble des gènes des microorganismes) chez les mammifères dont le régime alimentaire est composé de plus de 95% de fourmis et/ou de termites. Les placentaires myrmécophages représentent un exemple classique de convergence évolutive avec l'oryctérope, les fourmiliers, les pangolins, et le protèle (Delsuc *et al.* 2002). En termes d'adaptation, le microbiome intestinal joue un rôle majeur et son évolution chez les mammifères a été influencée par la phylogénie des espèces et leur régime alimentaire (Muegge *et al.* 2011). L'étude propose de tester une hypothèse évolutive majeure sur les mécanismes de convergence du microbiome liés au régime myrmécophage, en utilisant une approche de métagénomique environnementale.

A partir d'échantillons fécaux de mammifères myrmécophages et d'espèces apparentées, il s'agit de générer des données de composition taxinomique par séquençage Illumina de code-barres d'ARNr 16S. L'assignation taxinomique des séquences et l'analyse comparative des communautés bactériennes seront ensuite effectuées. Des données métagénomiques à large échelle seront ensuite générées par séquençage Illumina de l'ADN total pour un sous-échantillon représentatif d'espèces myrmécophages. Les séquences assemblées à partir de ces données métagénomiques seront ensuite annotées fonctionnellement par des recherches de similitude contre des bases de données protéiques et fonctionnelles. Ces deux types d'analyses permettront de caractériser respectivement les compositions taxinomique et fonctionnelle du microbiome intestinal de ces espèces.

La première partie taxinomique de l'étude a permis d'analyser 93 nouveaux échantillons fécaux par code-barres d'ARNr 16S représentant la diversité des espèces myrmécophages et des espèces apparentées. Ces données ont ensuite été combinées avec celles précédemment obtenues par Muegge *et al.* (2011). La représentation en réseau phylogénétique des distances UniFrac entre les différents microbiomes montre le regroupement des échantillons à la fois par phylogénie et par régime alimentaire (Figure 8D). Ainsi, on peut distinguer les deux grands groupes d'herbivores correspondant aux espèces monogastriques (capibara, rhinocéros, éléphant et zèbre) et polygastriques (girafe, okapi, gazelle et mouton). La figure 8D regroupe également les espèces myrmécophages qui possèdent des microbiomes intestinaux similaires malgré leurs origines phylogénétiques distinctes. Notamment, le protèle, une hyène spécialisée dans la consommation exclusive de termites, est rapproché de l'oryctérope et des fourmiliers plutôt que de l'hyène tachetée qui se positionne avec les autres carnivores.

Les premiers résultats montrent que les microbiomes intestinaux ont été façonnés de façon convergente par la myrmécophagie. Reste désormais à explorer le contenu fonctionnel en gènes de ces microbiomes pour révéler les détails de cette adaptation convergente et espérer répondre aux questions suivantes : à quel point les microbiomes intestinaux des mammifères myrmécophages sont similaires en termes de contenu fonctionnel ? Les mammifères myrmécophages utilisent-ils des bactéries particulières pour digérer la chitine des fourmis et termites grâce à des chitinases ? Si oui, les mêmes bactéries ont-elles été recrutées de façon indépendante ou bien différentes espèces bactériennes fournissent les mêmes fonctions ?

FOCUS 8-5

Apport de la paléogénomique à la compréhension des effets de la domestication : origine et évolution du chien

La domestication par l'homme d'espèces animales ou végétales est un processus d'évolution rapide impliquant des modifications génétiques, phénotypiques et comportementales modulant ainsi la diversité du vivant. Le résultat de ce processus s'observe aujourd'hui par exemple au niveau des différentes races domestiques. Cependant les phénomènes de domestication anciens ne peuvent être étudiés sur la seule base des observations de la diversité génétique actuelle. Grâce à l'obtention des données génétiques/génomiques diachroniques, la paléogénomique permet d'accéder à la diversité génétique passée, à sa variation lors des différentes étapes de la domestication, et de comprendre l'adaptation des espèces dans ce contexte.

Le chien (*Canis lupus familiaris*) est le premier animal domestiqué par les peuples chasseurs-cueilleurs au Paléolithique supérieur (30.000 à 15.000 BP). Sa domestication a été possible dans les régions où son ancêtre sauvage, le loup (*Canis lupus*) était présent. L'origine et le nombre d'évènements de domestication du chien ont longtemps fait débat. Des ossements de chiens et loups anciens provenant de 41 sites archéologiques à travers l'Eurasie ont fait l'objet d'une analyse paléogénomique.

En raison des caractéristiques de l'ADN ancien, ce type d'analyse nécessite l'authentification des séquences obtenues grâce à une expertise spécifique. En effet, il faut travailler de façon à minimiser les risques de contaminations lors des différentes étapes, de l'échantillonnage au séquençage. De même, dans les séquences obtenues, les substitutions artéfactuelles (dégradations dues au temps et à l'environnement) doivent être détectées et différenciées des vraies mutations biologiques. La possibilité d'obtenir une

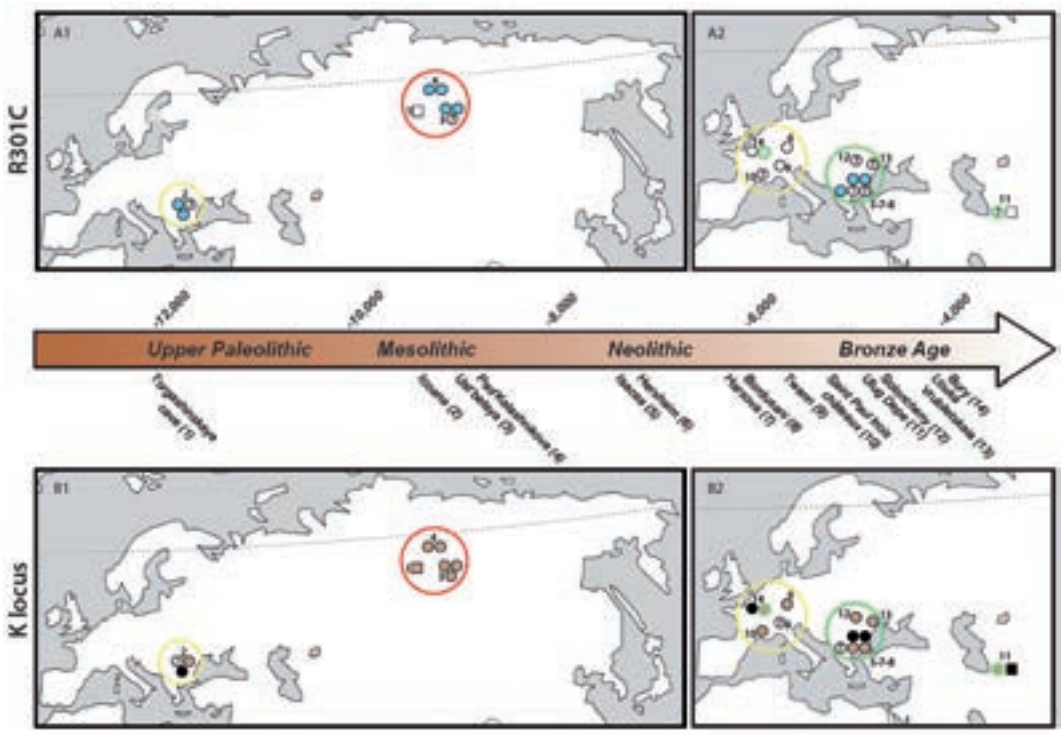


Figure 8E. Distribution des haplogroupes mitochondriaux et des allèles KB (CBD103) et R301C (MC1R) avant et après le début du Néolithique. Haplogroupes mitochondriaux A (rouge), C (jaune), D (vert). Présence de l'allèle KB (noir), absence de l'allèle KB (orange), présence de l'allèle R301C (bleu), absence de l'allèle R301C (blanc), couleur indéterminée (?). carré : loups, ronds : chiens.

FOCUS 8-5 (suite)

profondeur de séquençage importante (nombres d'amplicons*) par NGS (ici technologie 454), pour un fragment de gène donné, facilite une telle approche. Seules les séquences authentifiées doivent alors être prises en compte.

Dans un premier temps, 68 séquences authentiques d'ADN mitochondrial (D-loop) ont été obtenues et ont permis d'accéder à la diversité génétique des chiens anciens. Ces données ont été confrontées à des données morphométriques. Les résultats ont montré que la domestication du chien a eu lieu dans au moins deux régions distinctes au Paléolithique : en Asie et en Europe de l'Ouest (Figure 8E), puis, plus tard, au cours du Néolithique, au Moyen-Orient. Ceci suggère que plusieurs populations de loups sont à l'origine des chiens actuels et que les premiers chiens étaient probablement caractérisés par une importante variabilité génétique et phénotypique.

Résultant de 300 ans de sélection artificielle, la population canine est de nos jours fragmentée en 350 races phénotypiquement bien caractérisées. Cependant, cette sélection étant très récente, les données génétiques actuelles ne permettent pas :
- d'accéder aux phénotypes des premiers chiens domestiqués,
- d'accéder à la variabilité génétique sous-tendant ces phénotypes et leur diffusion à travers le temps et l'espace,
- de comprendre les relations entre la diversité génétique et phénotypique ancienne et celle des races actuelles.

Une deuxième partie de l'étude a consisté à obtenir des données sur la couleur des premiers chiens (Figure 8E) dont la variation est l'un des premiers effets de la domestication. L'analyse des variants de deux gènes codant pour la couleur du pelage, Mc1r (Melanocortin 1 Receptor) et CBD103 (canine beta-defensin), a permis de montrer une variabilité de la couleur dès le Mésolithique (Ollivier *et al.* 2013). Cette approche paléogénomique a permis de mettre en évidence non seulement la diversité issue du pool génétique de populations de loups, capturée lors du processus de domestication ; mais aussi l'apparition de nouveaux variants liés au relâchement des pressions de la sélection naturelle suite à la domestication. Ceci a été possible grâce à la comparaison des données génomiques déjà disponibles chez les chiens actuels (génome complet du boxer, séquences annotées, SNPs…) et celles obtenues chez les loups et chiens anciens.

RÉFÉRENCES

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecol Evol 3:846-852.

Coissac E, Riaz T, Puillandre N. 2012. Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21:1834–1847.

Delsuc F, Scally M, Madsen O, Stanhope MJ, de Jong WW, Catzeflis FM, Springer MS, Douzery EJP. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol Biol Evol 19:1656-1671.

Dettai A, Gallut C, Brouillet S, Pothier J, Lecointre G, Debruyne R. 2012. Conveniently pre-tagged and pre-packaged: extended molecular identification and metagenomics using complete metazoan mitochondrial genomes. PloS One 7:e51263.

Muegge B D, Kuczynski J, Knights D, Clemente JC, González A, Fontana L, Henrissat B, Knight R, Gordon JI. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332:970-974.

Ollivier M *et al.* 2013. Evidence of coat color variations sheds new ligth in ancient canids. PloS One 8:e75110.

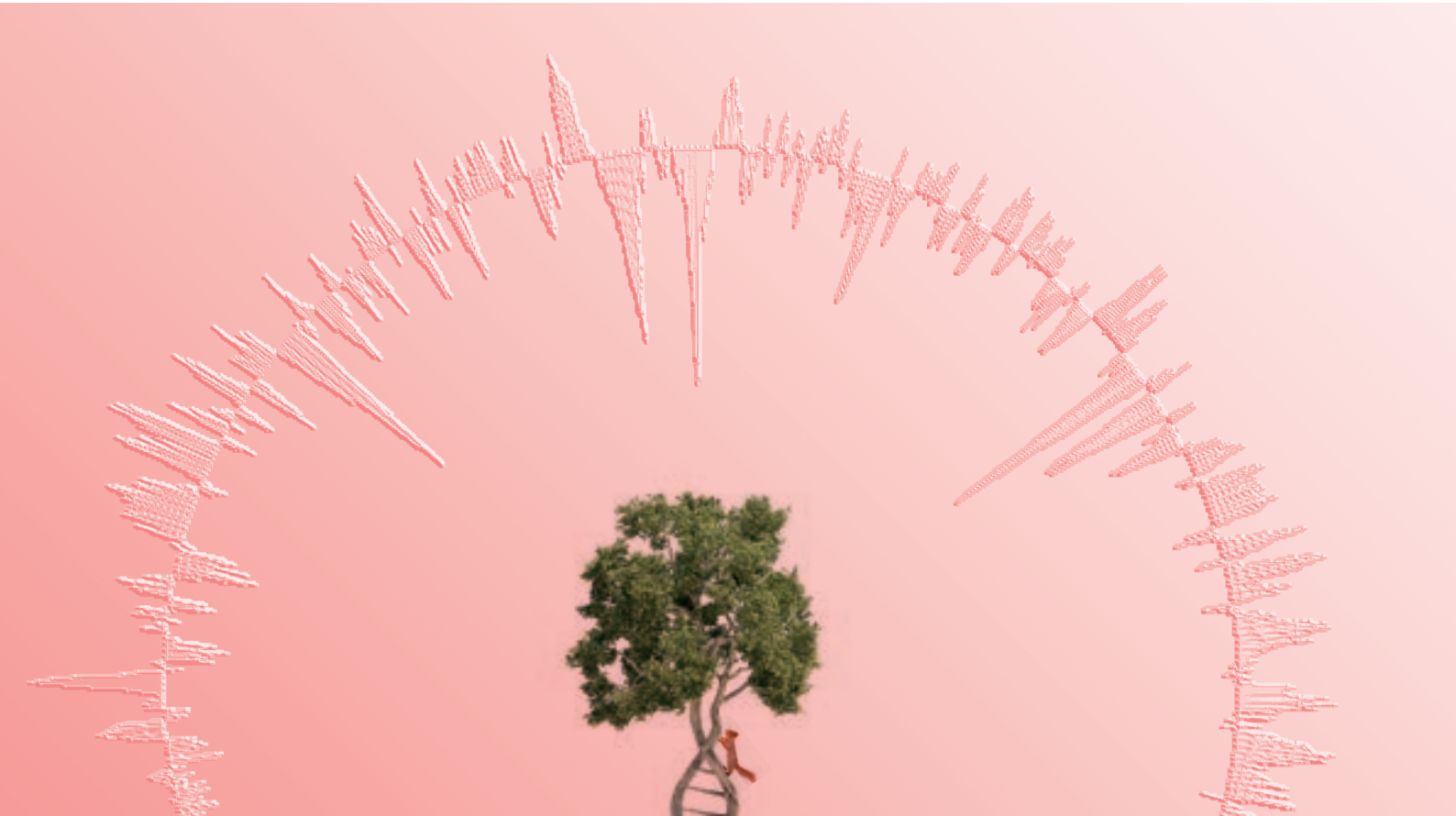
Puillandre N *et al.* 2012. New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Mol Ecol Res 12:396-402.

Puillandre N, Macpherson E, Lambourdière J, Cruaud C, Boisselier-Dubayle MC, Samadi S. 2011. Barcoding type specimens helps to identify synonyms and an unnamed new species in *Eumunida* Smith, 1883 (Decapoda: Eumunididae). Invertebrate Systematics 25:322-333.

Roquet C, Thuiller W, Lavergne S. 2013. Building megaphylogenies for macroecology: taking up the challenge. Ecography 36:13-26.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol 21:2045– 2050.

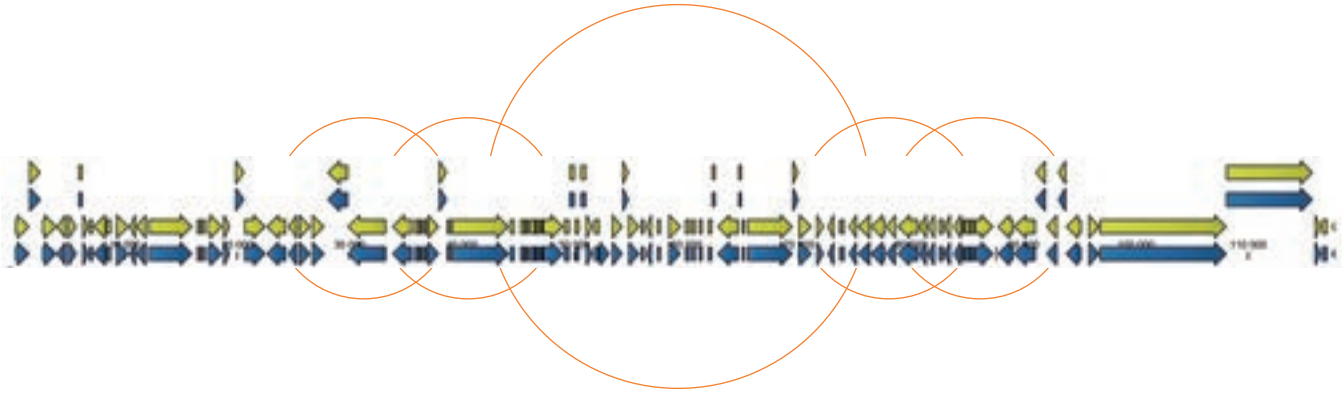
Valentini A, Pompanon F, Taberlet P. 2009. DNA barcoding for ecologists. Trends Ecol Evol 24:110–117.



IX ETUDE DE L'ÉVOLUTION ADAPTATIVE DES GÈNES ET DES GÉNOMES

Coordinateurs : Mathieu Joron et Xavier Vekemans
Contributeurs : Frantz Depaulis, David Enard, Laurence Garczarek, Sylvain Merlot, Frédéric Partensky, Carole Smadja

La compréhension fine de l'histoire évolutive des caractères nécessite la mise à jour de leurs bases génétiques. Les objectifs sont ici de découvrir les gènes ou structures génomiques impliqués dans l'évolution adaptative, c'est-à-dire en réponse à la sélection naturelle, et d'identifier les variants fonctionnels importants ainsi que les changements évolutifs ou écologiques associés, en tenant compte des effets confondants liés à l'histoire spécifique des populations concernées.



Chez les organismes non-modèles, cette démarche nécessite **l'intégration de multiples sources de signal**, issues de diverses approches allant de la génétique moléculaire des populations, à la phylogénétique et à la génétique formelle, et qui permettent par leur combinaison d'identifier les variants fonctionnels importants. Jusqu'à récemment, ces projets demandaient un fort investissement humain et technique en raison du bas débit du génotypage, de l'inaccessibilité des régions inconnues du génome, et de la limitation des méthodes statistiques multilocus. Les développements méthodologiques ont depuis révélé l'importance de la dimension multilocus dans l'étude des réponses adaptatives. Le séquençage massif en parallèle proposé par les NGS lève alors un verrou technologique permettant d'atteindre le débit nécessaire à l'approche génomique. Il permet désormais d'aborder des **questions évolutives et populationnelles très fondamentales** mais jusque-là inenvisageables, voire considérées comme relevant du champ de la biologie théorique, comme par exemple la nature des changements fonctionnels et des réseaux de gènes impliqués, la distribution des mutations et de leurs effets, la structure de la recombinaison, ou encore la nature composite des génomes soumis à des régimes de sélection complexes (Figure 9A, Focus 9-1).

La découverte des variations génétiques et génomiques impliquées dans l'évolution adaptative bénéficie de la mise en œuvre de toute la **palette des technologies NGS** actuelles, sur du matériel génomique ou transcriptomique. Bien que les NGS aboutissent à un faible coût

Figure 9A. Adaptation biologique et diversité dans l'environnement marin.

de séquençage par paire de base, chaque expérience représente malgré tout un investissement conséquent en terme de séquençage, et la taille des génomes eucaryotes impose un coût par individu qui peut être limitant. Les stratégies optimales passent souvent par un **sous-échantillonnage de la complexité des génomes** en fonction du type d'expérience voulue et du matériel accessible (reduced representation libraries, Davey *et al.* 2011).

Le RAD-seq permet un sous-échantillonnage important du génome autour des sites de restriction, associé à de fortes possibilités de multiplexage* à coût modéré. Il est utile en cartographie génétique, populationnelle, pour les scans génomiques d'association, ainsi que pour des analyses phylogénétiques ou phylogéographiques multilocus sur des clades récents (Baird *et al.* 2008, Emerson *et al.* 2010) où les méthodes classiques n'offrent souvent pas la résolution voulue. A de plus grandes profondeurs phylogénétiques, il est souvent difficile d'obtenir un nombre suffisant de marqueurs conservés. Les défis actuels concernent l'utilisation efficace des RAD-seq en l'absence de génome de référence*.

La capture de séquence (Focus 9-2) permet d'optimiser l'effort de séquençage vers des régions précises du génome (segment chromosomique, famille de gènes) « capturées » par hybridation d'après les séquences de référence. Elle est utile pour l'étude des patrons de variation de régions données (Nadeau *et al.* 2012), ou pour quantifier les niveaux de duplication de segments génomiques par des analyses de couverture. Ces approches perdent graduellement de l'intérêt à mesure que baisse le coût du séquençage lui-même par rapport aux étapes de préparation, ici assez coûteuses. Les défis actuels résident dans la capture de séquences divergentes par rapport à la référence (Nadeau *et al.* 2012), ainsi que dans l'optimisation de formes de multiplexage ou de « pooling » avant capture pour réduire les coûts de préparation.

Le transcriptome (Focus 9-3), en tant que sous-échantillonnage focalisé sur les séquences exprimées, est utile dans une optique comparative entre taxons génétiquement distants, notamment pour la comparaison du polymorphisme moléculaire dans des groupes non-modèles (Gayral *et al.* 2013). Il est également utilisé pour obtenir une référence afin d'annoter un génome (*Heliconius* Genome Consortium 2012), ou définir d'autres applications (par exemple la capture

FOCUS 9-1

Sélection récurrente dans les génomes de primates

Les mêmes locus sont-ils soumis à la sélection naturelle dans différentes lignées évolutives ? L'avènement des NGS s'accompagne d'une accélération spectaculaire du séquençage de génomes entiers de diverses espèces à partir d'un seul individu de référence. Bien qu'ils représentent un échantillonnage par espèce qui peut apparaître très limité, ces génomes uniques individuels apportent pourtant une information précieuse sur la variabilité intraspécifique (polymorphisme) à travers leur hétérozygotie. En effet, il faut considérer qu'ils reflètent chacun un grand nombre d'ancêtres, notamment du fait de la recombinaison génétique, et même à ce niveau les patrons de variabilité génétique le long du génome renseignent sur l'adaptation. En particulier, l'invasion d'un variant avantageux dans une population entraîne avec la mutation fonctionnelle toute la région voisine du génome, selon un processus de balayage sélectif bien connu, et conduit ainsi à une variabilité locale réduite le long du segment chromosomique (Enard *et al.* 2010). Un test sélectif classique (HKA, Enard et al. 2002) repose sur la détection de ces régions marginalement pauvres en diversité par rapport à la variabilité interspécifique qui reflète le taux de mutation local. Ces tests ont été adaptés aux génomes individuels entièrement séquencés. Son application sur les génomes de primates séquencés a fourni une liste de locus indiquant une possible réponse à la sélection naturelle (Hudson *et al.* 1987). Cette liste est largement congruente avec celles trouvées avec d'autres tests de neutralité, en particulier les tests focalisés sur les comparaisons interspécifiques du rapport entre variabilité non synonyme et variabilité synonyme (supposée neutre) dans les séquences codantes des gènes (Maynard Smith et Haigh 1974). Cette congruence valide ainsi l'efficacité du test. Les gènes «candidats», indiquant une réponse à la sélection, incluent plus particulièrement des gènes exprimés dans le cervelet, la rate et les testicules et impliquent des mécanismes de défense ou liés à la gamétogénèse, à la transcription et au développement du cerveau antérieur. Enfin les mêmes gènes candidats s'avèrent préférentiellement sélectionnés de façon récurrente dans les branches évolutives conduisant aux différents primates (Figure 9B). Cela remet en cause certaines interprétations adaptatives spécifiquement humaines, comme par exemple pour le gène dit du langage, FOXP2 (Yang 1997).

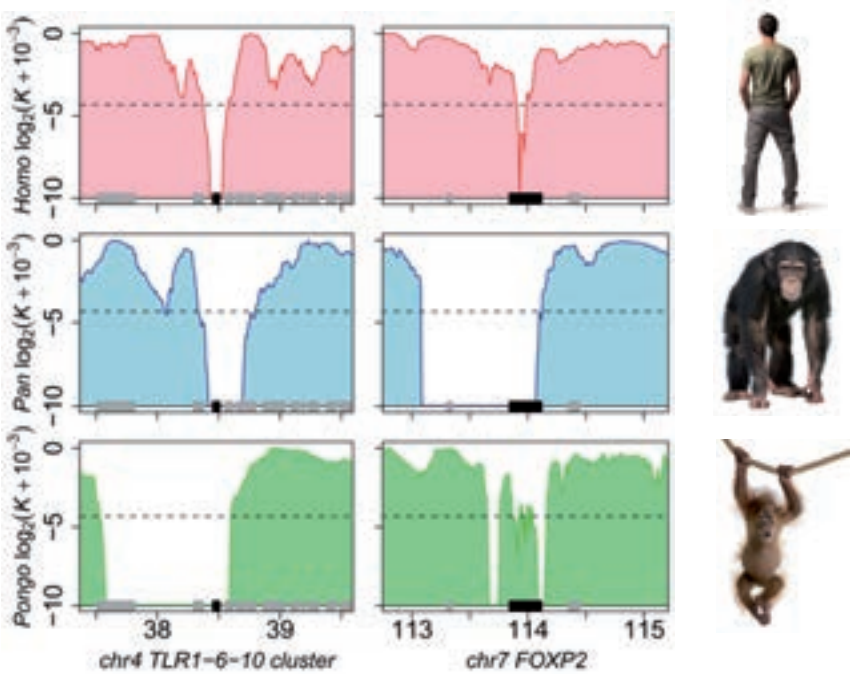


Figure 9B. Exemples de locus candidats détectés dans plusieurs lignées évolutives. Balayages sélectifs détectés aux locus des récepteurs Toll-like 1, 6 et 10 (réponse immunitaire innée ; gauche) et du gène FOXP2 (droite). Chaque figure couvre 2 mégabases (axe des abscisses). La position des gènes est indiquée sur cet axe (noir : gènes candidats à la sélection centrés sur les creux de variabilité, gris : autre gènes). Haut : homme (rose) ; milieu : chimpanzé (bleu) ; bas : orang-outang (vert).

FOCUS 9-2

Reséquençage ciblé et bases génétiques de l'adaptation et de la spéciation

Les nouvelles technologies d'enrichissement massif, appelées capture génomique ou reséquençage ciblé, permettent d'amplifier spécifiquement de larges voire de très larges portions du génome avant leur séquençage à haut débit. Basées sur un principe d'hybridation d'ADN ou d'ARN à des sondes synthétisées à partir d'une séquence de référence, ces méthodes sont une alternative au reséquençage de génomes entiers ou au génotypage à haut débit, en permettant une approche à la fois ciblée mais à large échelle. Un avantage majeur de cette stratégie expérimentale est de permettre l'obtention de couvertures de séquençage importantes pour les régions génomiques cibles dont la taille ou la complexité ne permet pas l'amplification pour un séquençage traditionnel. Pour cette raison, le reséquençage ciblé a souvent pour applications clés la recherche de variants nucléotidiques mais également de variants structuraux de type CNV (copy number variation) qui peuvent être détectés notamment grâce à des méthodes basées sur la profondeur de couverture.

Figure 9C. Puceron du pois

Exemple 1 : Bases génétiques de la spécialisation à la plante hôte et de la spéciation chez le puceron du pois (Figure 9C), *Acyrthosiphon pisum* par capture de gènes candidats.

Différentes races d'hôtes du puceron du pois coexistent en sympatrie en étant hautement spécialisées sur différentes légumineuses et cette spécialisation écologique est un élément clé de la réduction du flux de gènes entre biotypes. Le choix de la plante hôte faisant intervenir des mécanismes de reconnaissance gustative et olfactive, les gènes chimiosensoriels, organisés en larges familles multigéniques dans le génome du puceron, sont de bons candidats pour l'adaptation et la spéciation écologique chez cet insecte. Des approches de reséquençage ciblant ces familles géniques ont été développées pour rechercher les changements génétiques à l'origine de la divergence entre races d'hôte. Deux résultats majeurs ressortent de ces études (Smadja *et al.* 2012) :

- l'identification d'une dizaine de gènes des récepteurs olfactifs et gustatifs comme candidats à l'adaptation à la plante hôte car présentant la signature d'une divergence sous sélection entre biotypes ;
- la mise en évidence que certaines familles géniques candidates, et en particulier les récepteurs olfactifs et les Odorant Binding Proteins, présentent une divergence du nombre de copies entre races d'hôtes, suggérant ainsi un rôle de duplications segmentaires dans la diversification adaptative.

Figure 9D. Souris domestique

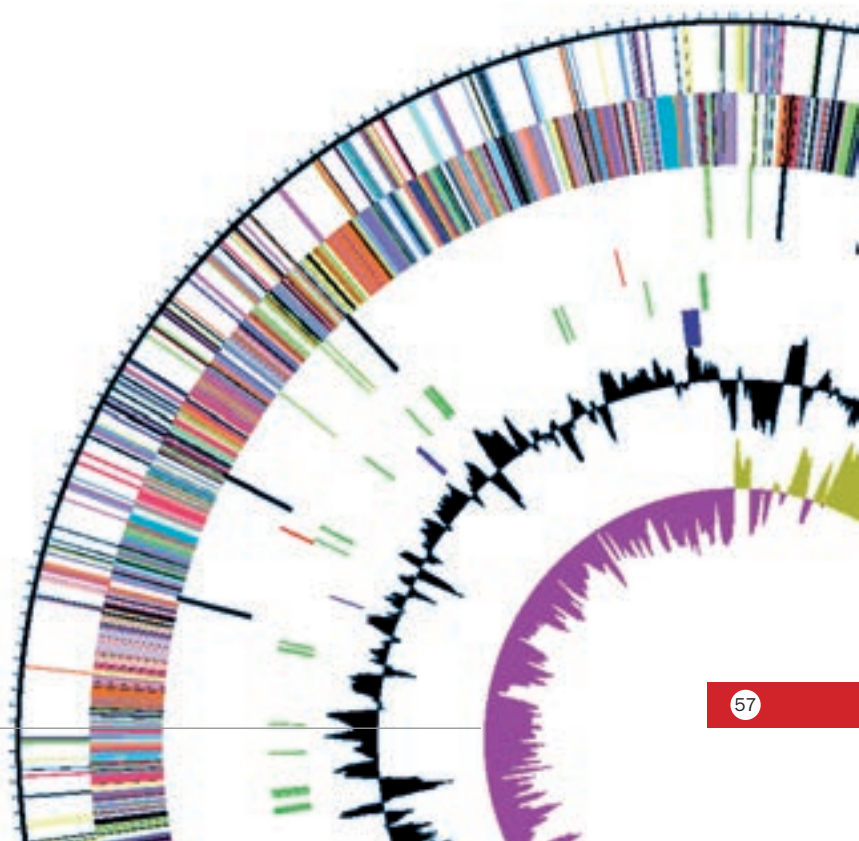
Exemple 2 : Bases génétiques de l'isolement sexuel et du renforcement chez la souris domestique *Mus musculus* (Figure 9D) par reséquençage d'exomes entiers.

Les deux sous-espèces européennes de la souris domestique, *M. m. musculus* et *M. m. domesticus*, ont divergé en allopatrie pendant 0.5 million d'années avant de se rencontrer en Europe et former une zone hybride où des barrières comportementales au flux génique semblent avoir évolué en réponse à un désavantage sélectif des hybrides (renforcement). Une stratégie de reséquençage de l'exome entier de souris a été mise en œuvre pour découvrir les régions génomiques impliquées dans cet isolement sexuel. La comparaison des degrés de polymorphisme entre populations allopatriques et populations en contact permet dans un premier temps d'identifier les gènes soumis aux pressions sélectives liées au phénomène de renforcement dans la zone hybride. Par ailleurs, l'analyse des niveaux de couverture dans les différentes régions de l'exome permet d'identifier d'éventuelles variations entre échantillons qui pourraient affecter certains gènes candidats comme les récepteurs voméronasaux, les récepteurs olfactifs, les gènes codant pour les Major Urinary Proteins ou le complexe majeur d'histocompatibilité. L'ensemble des facteurs génétiques pourraient être ainsi à la base de la divergence comportementale adaptative chez la souris.

de séquence). Mais le séquençage étant quantitatif, l'utilisation majeure (par exemple le RNA-seq) est l'étude d'expression et la recherche de séquences (gènes, ARN, transcrits) montrant une variation d'expression ou d'épissage (Colombo *et al.* 2013). Les données RNAseq sont ainsi couramment utilisées pour comparer la variation transcriptionnelle entre variants phénotypiques, et les pipelines* pour l'analyse sont nombreux. Les défis actuels concernent l'assemblage des transcriptomes et la complexité des jeux de données (Gayral *et al.* 2013), nécessitant temps et capacités de calcul importants dès que le nombre de réplicats biologiques ou techniques augmente.

Le reséquençage (WGS) permet d'accéder à la variation génétique totale et devient abordable pour les espèces à génome petit ou moyen. La fraction séquencée ne reposant pas sur une référence, les données obtenues sont indépendantes de la distance génétique avec les taxons connus, ce qui est utile pour certaines approches comparatives. Par ailleurs, la préparation des librairies ne nécessite pas d'étapes supplémentaires et leur passage sur machine est l'opération de routine des centres de séquençage, assurant un minimum de complications et un gain de temps souvent conséquent. Les défis actuels concernent l'optimisation du rapport entre profondeur de séquençage et qualité du génotypage, certaines applications pouvant se satisfaire d'un séquençage plus superficiel (Davey *et al.* 2011, Buerkle et Gompert 2013) comme par exemple la cartographie impliquant une diversité allélique réduite. Une alternative au reséquençage complet de génomes individuels visant à obtenir des données de polymorphisme intrapopulationnel consiste à séquencer en aveugle un mélange d'individus issus d'une même population (Pool-seq). Cette stratégie permet de diminuer le coût d'obtention de données de polymorphisme, et sous certaines conditions (nombre élevé d'individus assemblés), de minimiser la variance des estimateurs de fréquences alléliques. Cette approche semble particulièrement performante pour identifier des régions génomiques fortement différenciées entre populations issues de milieux contrastés (Boitard *et al.* 2012). Une autre alternative au reséquençage total concerne un séquençage ciblé qui permet d'augmenter la couverture sur des régions d'intérêt (Focus 9-1).

La disposition d'un **génome de référence** (Focus 9-4) est un atout majeur dans la plupart des approches précédentes. Pour documenter la variation structurale entre génomes, la détection de points de rupture de synténie peut passer par l'analyse des discordances positionnelles des séquences par rapport à une référence. Les défis actuels concernent la disponibilité de génomes assemblés, et leur qualité d'assemblage qui affecte la puissance de ces approches, notamment pour écarter les faux positifs. Pour des approches microévolutives comparatives, un génome de référence augmente considérablement la puissance des tests, permettant des approches par fenêtre glissante exploitant toute la continuité génomique le long des fragments conservés (scaffolds ou superscaffolds). Les NGS permettent aujourd'hui de générer des génomes eucaryotes complets de quelques centaines de mégabases, moyennant le séquençage profond (100X ou plus par Illumina 100bp) et l'emploi de librairies de fragments de tailles différentes pour optimiser l'assemblage (Zhan *et al.* 2011). L'assemblage est par ailleurs facilité par la disponibilité de ressources génomiques annexes et de matériel de départ homozygote. Outre la nécessité d'une couverture profonde, et donc un coût conséquent, les défis actuels pour les organismes non-modèles résident dans l'optimisation de l'assemblage, notamment la gestion bioinformatique de l'hétérozygotie et de la variation haplotypique lors de l'assemblage.



L'utilisation des données de polymorphisme moléculaire pour l'inférence des processus adaptatifs dans les analyses NGS (Focus 9-5) nécessite de prendre en compte les interférences liées à l'histoire démographique des populations. Les outils actuels d'inférence, faisant appel notamment aux simulations numériques intensives de modèles basés sur la coalescence, sont gourmands en temps de calcul et donc inadaptés aux données massives issues des NGS. L'exploitation optimale de ces jeux de données passe par le développement de méthodes innovantes permettant à la fois une grande flexibilité des modèles sous-jacents, et une importante efficacité numérique (Gompert et Buerkle 2011, Fariello *et al.* 2013). La communauté scientifique de théoriciens en génétique et génomique des populations en France dispose d'atouts importants pour porter l'innovation dans ce domaine (Boitard *et al.* 2012, Fariello *et al.* 2013).

Une **forte limitation pour l'utilisation des NGS** dans les recherches portant sur l'évolution adaptative est liée à l'accès aux plateformes de séquençage dans un cadre flexible et à un coût raisonnable. Contrairement à d'autres parties du monde où existent des plateformes universitaires performantes assurant au tissu scientifique local un accès direct à ces technologies, les équipes françaises doivent souvent faire appel à des **prestataires extérieurs** qui proposent un service commercial généralement standardisé, moins à la pointe des développements ou utilisations de ces technologies, et manquant de flexibilité et de proximité par rapport à la diversité des questions et approches souhaitées. Une autre limitation forte réside dans la présence limitée, au sein des laboratoires, de bioinformaticiens sensibilisés aux défis inhérents à la biologie évolutive, c'est-à-dire la variabilité naturelle rencontrée dans ces domaines et le jonglage nécessaire entre les différentes approches et niveaux d'analyse. Il ressort qu'une véritable intégration entre la formulation des questions pertinentes d'un point de vue scientifique d'une part et les forces de propositions technologiques ou méthodologiques/bioinformatiques d'autre part apporterait la synergie nécessaire aux innovations dans ce domaine.

Parmi les **défis scientifiques et techniques** liés à l'utilisation des données NGS pour l'étude de l'histoire évolutive des gènes et des génomes, on peut en citer 3 principaux pour lesquels la communauté française a un atout majeur à jouer dans ce domaine :

1- La détection des signatures de sélection à partir de données NGS populationnelles, les outils actuels étant en partie inopérants sur l'ensemble du génome, en raison du volume de la variation et de la complexité de sa structure au niveau génomique.

2- La variabilité naturelle intra et interspécifique qui affecte par exemple *in silico* l'alignement* à la référence, ou *in vitro* l'efficacité des captures de séquences (Nadeau *et al.* 2012) ; les problèmes concernent l'interprétation des estimateurs impliquant des comparaisons inter espèces, ainsi que la qualité du multiplexage. Une meilleure documentation de cette variation et une meilleure compréhension de ses effets sont cruciales.

3- Le développement d'outils bioinformatiques performants pour l'assemblage de novo pour les différents types de jeux de données, dépendant de la disponibilité des génomes de référence et/ou affectant la possibilité de s'en affranchir. Les enjeux concernent la complexité des jeux de données ainsi que la gestion de l'hétérozygotie, la polyploïdie, et les variations structurales de plus ou moins grande ampleur (indel, réarrangements, Gayral *et al.* 2013, *Heliconius* Genome Consortium 2012).

En conclusion, les études combinant de multiples niveaux d'analyse impliquant les NGS et aboutissant à la découverte de variants génétiques permettent de tester directement les scénarios adaptatifs jusque-là spéculatifs. Cependant, le débit de production des données moléculaires excédant souvent la rapidité de développement de techniques d'analyse adaptées, une grande partie des études phare est associée à des innovations dans la production, l'analyse ou la combinaison des données génomiques (Baird *et al.* 2008). Ces innovations reposent souvent sur l'association de trois piliers : expertise en biologie des populations, plateformes de séquençage, expertise et infrastructure bioinformatique. Elles constituent pour les institutions qui les développent des tremplins les plaçant au devant de la communauté, renforçant leur position internationale. La recherche française doit combler ce retard en permettant une meilleure synergie, à l'échelle locale, entre la formulation de questions fondamentales sur les processus adaptatifs, et les capacités de production et d'analyse des données.

FOCUS 9-3

Evometonicks : les technologies NGS ouvrent la voie des recherches moléculaires sur les plantes hyperaccumulatrices de nickel

Il existe environ 500 espèces végétales (0,2% des angiospermes) capables d'accumuler des concentrations importantes de métaux dans leurs parties aériennes. De manière remarquable, 400 de ces espèces réparties dans une quarantaine de familles, accumulent plus de 0,1% (masse sèche) de nickel dans leur partie aérienne (Kramer 2010). Ces espèces hyperaccumulatrices vivent dans des sols ultramafiques ou serpentiniques riches en nickel, principalement dans des zones tropicales (par ex. Cuba, Nouvelle-Calédonie), mais également dans des régions tempérées (Europe).

Figure 9E. *Psychotria semperflorens* (gauche) et *P. gabriellae* (droite) observées en sympatrie en forêt humide sur sol ultramafique en Nouvelle-Calédonie. Dans ces conditions, *P. gabriellae* accumule plus de 1% de nickel dans ses feuilles, conduisant à une coloration rose en présence de dimethylglyoxime (insert), alors que *P. semperflorens* en accumule 100 fois moins. Les transcriptomes de ces espèces phylogénétiquement proches peuvent être comparés afin d'identifier les gènes dont l'expression est liée à l'accumulation de nickel.

Ces plantes représentent aujourd'hui un atout important pour le développement de biotechnologies durables comme la phytoremédiation des sols pollués, le phytomining ou la chimie verte (Losfeld *et al.* 2012). Cependant, les mécanismes moléculaires impliqués dans l'accumulation du nickel sont encore très peu connus, car les espèces hyperaccumulatrices n'ont pas bénéficié des recherches et des outils développés pour les espèces modèles ou d'intérêt agronomique. Afin d'identifier les gènes impliqués dans l'hyperaccumulation du nickel chez les plantes, les NGS sont utilisées pour connaître les séquences génomiques de ces plantes ainsi que pour étudier l'expression des gènes.

Le transcriptome de *Psychotria gabriellae* (aussi connu sous le nom de *P. douarrei*), une Rubiacée endémique de Nouvelle-Calédonie qui accumule jusqu'à 4% de nickel dans ses feuilles a d'abord été séquencé (Verbruggen *et al.* 2009). L'assemblage *de novo* des lectures obtenues par GS-FLX (Roche 454) a fourni 30.000 contigs et ainsi la première base de données de séquence de gènes exprimés pour cette espèce. Ont ensuite été identifiés et clonés des gènes codant pour des protéines membranaires capables de transporter des métaux (Kramer 2010). Un des transporteurs identifiés, homologue de la ferroportin qui transporte le fer dans les cellules de mammifères, est capable de transporter le nickel chez les plantes. De plus, l'expression de ce transporteur est moins importante chez l'espèce proche *Psychotria semperflorens* qui, bien que vivant en sympatrie avec *P. gabriellae* (Figure 9E), n'accumule pas le nickel. Ces résultats suggèrent que ce transporteur de la famille de la ferroportine participe à l'hyperaccumulation du nickel chez *P. gabriellae*.

Ce type d'étude valorise le potentiel de la biodiversité d'espèces non-modèles. Dans la suite des recherches qui seront menées en collaboration avec des partenaires en Nouvelle-Calédonie et aux Pays-Bas, des approches seront mises en œuvre pour comparer de façon globale et quantitative le transcriptome d'espèces hyperaccumulatrices de nickel appartenant à des familles éloignées (*P. gabriellae* chez les Rubiaceae, *Noccaea caerulescens* chez les Brassicaceae), avec celui d'espèces proches non accumulatrices (*P. semperflorens*, *Microthlaspi perfoliatum*, respectivement). Sera ainsi identifiée et étudiée l'évolution des mécanismes moléculaires impliqués dans l'accumulation du nickel chez différentes familles d'angiospermes.

FOCUS 9-4

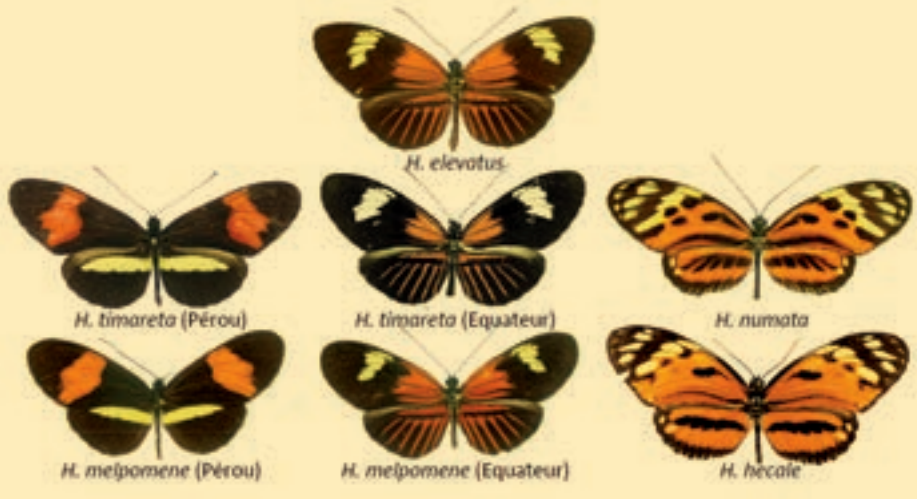
Génome de référence et reséquençage populationnel : origines de la convergence évolutive chez les papillons *Heliconius*

Le genre néotropical *Heliconius* montre une radiation-diversification spectaculaire liée au mimétisme (Joron *et al.* 2006, Figure 9F) ; un nombre croissant d'équipes s'intéresse aux déterminants génétiques de leur convergence. Afin d'obtenir un génome complet de référence, ces équipes se sont liées de manière informelle en un consortium international (www.heliconius.org) pour profiter de l'essor des NGS dès ses débuts et séquencer le génome de l'espèce *H. melpomene* (génome haploïde de 295Mb, *Heliconius* Genome Consortium 2012).

Le pyroséquençage* 454 Roche, au débit faible face à Illumina mais aux lectures longues de 400bp, fut adopté initialement pour obtenir un premier assemblage (12M de lectures sur un individu consanguin). Le séquençage Illumina par paires, et notamment le séquençage dit « mate-pair » par circularisation de fragments longs (3kb et 5kb), fut ensuite utilisé pour associer en scaffolds les contigs séparés par des séquences répétées (8M de paires de lectures). Ensuite, le développement de lectures de 100bp par Illumina HiSeq a permis d'obtenir 42M de paires de lectures supplémentaires, permettant d'améliorer la couverture moyenne et corriger les erreurs inhérentes au pyroséquençage. L'assemblage final est alors de ~269Mb (91% du génome, 3800 scaffolds, N50=277Kb). Enfin, les individus issus d'un croisement Mendélien entre la lignée consanguine et une lignée divergente ont été génotypés par Illumina RAD-seq pour réaliser une cartographie génétique, permettant d'ordonner et d'orienter sur chaque chromosome les scaffolds génomiques contenant des marqueurs RAD (superassemblage de 83% du génome, 1273 superscaffolds, N50=400Kb) (www.butterflygenome.org).

Les technologies et les méthodes analytiques, en plein développement, ont changé drastiquement au cours du déroulement du projet (4 ans). La stratégie de séquençage fut continuellement adaptée à la puissance croissante des technologies, ainsi qu'aux problèmes d'assemblage rencontrés. Ces problèmes concernent principalement l'hétérozygotie du génome qui rend l'assemblage difficile, et sa structure haplotypique qui produit une duplication artefactuelle des segments fortement hétérozygotes. Le séquençage de banques génomiques de tailles variées et l'implémentation de stratégies bioinformatiques *ad hoc* (grâce aux postdoctorants issus de laboratoires spécialisés) ont permis de résoudre en partie ces problèmes (*Heliconius* Genome Consortium 2012). Le génome obtenu est très utile en l'état pour une grande gamme d'applications, mais restera continuellement en phase d'amélioration.

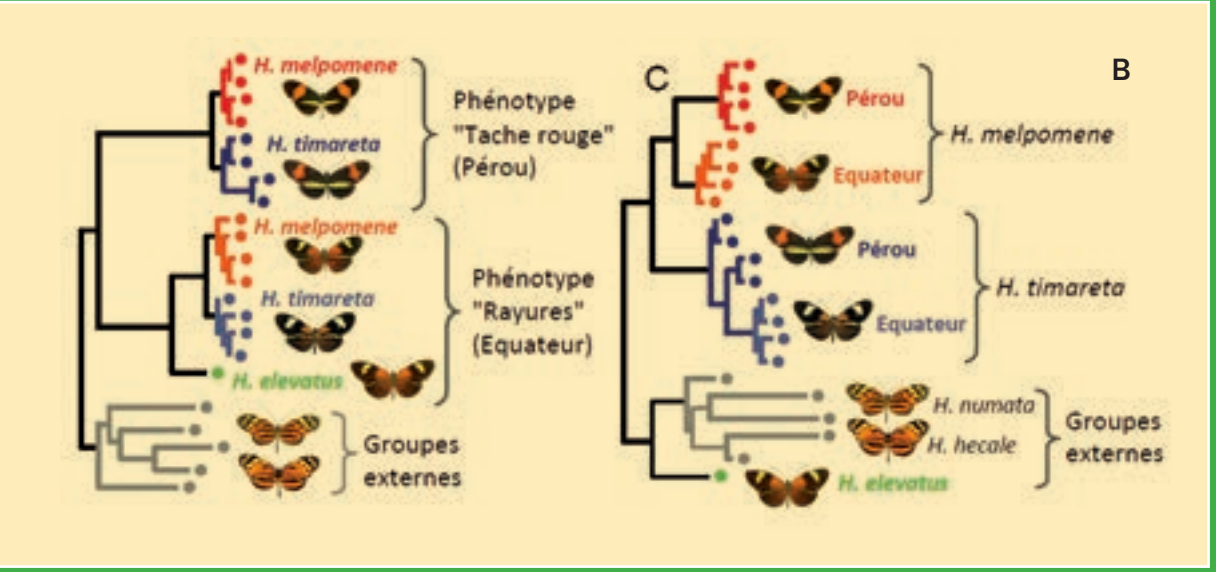
A



Afin de comprendre la structuration de la variation génétique entre taxons et entre groupes de convergence mimétique (Figure 9F), le génome assemblé fut utilisé comme référence pour aligner les lectures de reséquençage populationnel (Illumina HiSeq) sur les scaffolds génomiques positionnés, en particulier les locus contrôlant la coloration. La continuité et le positionnement génomique des scaffolds permet des analyses par fenêtres glissantes (10 à 50kb) qui révèlent la variation des patrons de polymorphisme sur de grandes quantités de sites variables, et procurent une puissance d'analyse conséquente même entre taxons peu différenciés. Ici, les intervalles génomiques contrôlant la couleur présentent une topologie phylogénétique nettement structurée par phénotype (Figure 9FB), alors que les régions immédiatement adjacentes ainsi que le génome dans son ensemble sont structurés par taxons et par la géographie (Figure 9FC, *Heliconius* Genome Consortium 2012). D'autres analyses basées sur le partage de SNPs dérivés (ABBA-BABA, *Heliconius* Genome Consortium 2012, Durand *et al.* 2012) ont révélé un signal d'introgression entre taxons mimétiques spécifiquement associé aux locus de la couleur. Ces locus montrent ainsi une histoire très différente de celle des populations les contenant. Notamment, la convergence mimétique a ici évolué à la faveur de l'échange des allèles de coloration entre espèces (permettant une ressemblance «instantanée» entre elles), et non, comme on le croyait, par l'évolution en parallèle de phénotypes semblables chez chacun des taxons (*Heliconius* Genome Consortium 2012).

L'assemblage de novo d'un génome, associé au reséquençage populationnel, a ainsi permis d'exploiter toute la variation génomique et de révéler l'introgression adaptative des allèles de la coloration, permettant de proposer un scenario inattendu pour expliquer la convergence phénotypique entre espèces.

Figure 9F. Origine génétique de la convergence mimétique chez *Heliconius*. A. Trois groupes de mimétisme distincts, montrant la convergence entre *H. melpomene* et *H. timareta* dans une vallée andine du Pérou (gauche, phénotype « tache rouge »), entre d'autres races de ces mêmes espèces ainsi que *H. elevatus* en Amazonie équatorienne (colonne du milieu, phénotype « rayures »), ainsi qu'entre *H. numata* et *H. hecale* dans ces mêmes régions (phénotype « tigré », Joron *et al.* 2006). B. « Arbre du mimétisme », topologie phylogénétique correspondant précisément à l'intervalle génomique de 50kb contenant le locus contrôlant la variation mimétique, où les individus se groupent par phénotype et non par espèce. C. « Arbre des espèces », topologie phylogénétique correspondant à l'essentiel du génome, où les taxons se groupent par espèces. B et C d'après *Heliconius* Genome Consortium et al. 2012. Barre=1% de divergence.



FOCUS 9-5

Génomique comparative et métagénomique des picocyanobactéries marines

Prochlorococcus et *Synechococcus* sont les deux organismes photosynthétiques les plus abondants de l'océan et jouent de ce fait un rôle primordial dans la production primaire globale (Flombaum *et al.* 2013). L'un des principaux objectifs de recherche est de mieux comprendre le lien entre la diversité génomique au sein de ces genres et la capacité des différents écotypes à s'adapter à des niches écologiques particulières de l'écosystème marin.

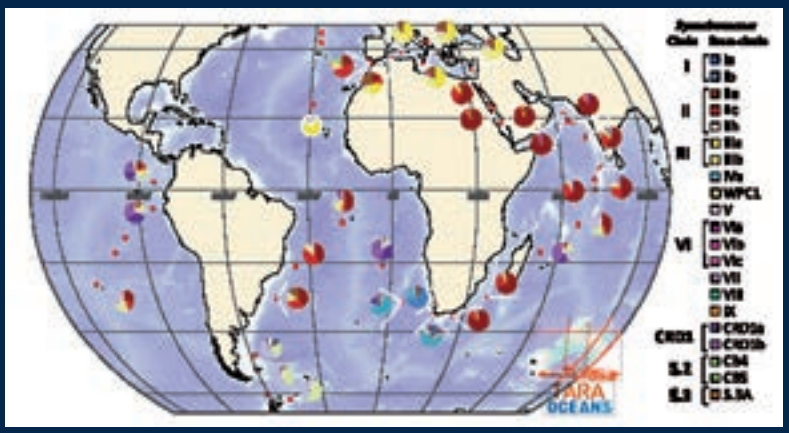


Figure 9G. Analyse préliminaire de la distribution des groupes phylogénétiques de *Synechococcus* marins à 32 stations le long du transect de la campagne océanographique TARA-Oceans. L'assignation des séquences à un groupe taxonomique particulier (sous-cluster, clade ou sous-clade, a été effectué par recrutement des séquences environnementales sur les 40 génomes de référence de *Synechococcus*.

En collaboration avec le Génoscope, 25 génomes de *Synechococcus* ont été récemment séquencés, élevant à 57 le nombre de génomes de picocyanobactéries marines actuellement disponibles (40 *Synechococcus*, 14 *Prochlorococcus*, 3 *Cyanobium*). Grâce à une analyse comparative génomique les séquences orthologues ont été regroupées en utilisant l'algorithme OrthoMCL puis intégrées au système d'information Cyanorak v2 (http//sb-roscoff.fr/cyanorak/) dédié à l'annotation et l'analyse de ces génomes. Afin de réaliser l'analyse métagénomique, la diversité génétique des picocyanobactéries a été analysée sur 32 stations de la campagne TARA-Oceans, prélevées à 2 profondeurs. Après assemblage et nettoyage, les séquences de 175 bp obtenues par la technologie Illumina ont été alignées par Blastn sur les 57 génomes de référence.

Une étude comparative des 14 premiers génomes de picocyanobactéries (Dufresne *et al.* 2008) avait permis dès 2008 d'avoir une première estimation des répertoires de gènes communs, accessoires et uniques. L'ajout de 25 génomes de *Synechococcus* a, dans le cadre de cette étude, considérablement amélioré la couverture de la diversité génétique au sein de ce genre, avec plusieurs souches par clade phylogénétique. Une analyse comparative préliminaire de ces génomes a d'ores et déjà permis de mettre en évidence la présence d'îlots génomiques spécialisés dans l'adaptation à des niches écologiques particulières, notamment un îlot de 4 à 6 gènes impliqué dans l'acclimatation à différentes qualités de lumière.

Ces génomes ont également été utilisés comme références pour l'analyse de données de métagénomique de la campagne TARA-Océans (voir focus 3.2) par recrutement des séquences de populations naturelles de cyanobactéries sur le génome le plus proche, permettant leur assignation fonctionnelle et taxonomique. Ces analyses ont notamment permis de mettre en évidence : 1) la prédominance dans certains écosystèmes de clades qui avaient jusqu'à présent été considérés comme minoritaires, tels que CRD1 ou WPC1, et 2) des changements brusques dans la composition des communautés entre différents bassins océaniques, notamment entre la Méditerranée et la Mer Rouge ou de part et d'autre du cap de Bonne Espérance (Figure 9G). La disponibilité d'un tel set de génomes de référence a également permis de mettre en évidence des distributions distinctes entre populations appartenant à un même clade, mais à des sous-clades différents. Ainsi, cette approche permet d'atteindre une échelle taxonomique beaucoup plus fine que celle à laquelle donne accès le gène de l'ARN 16S, et devrait nous permettre de mieux définir la notion d'écotype (voire d'espèce) au sein des picocyanobactéries marines. Parmi nos autres objectifs, l'analyse de la distribution et de l'expression in situ des gènes identifiés par comparaison génomique comme étant spécifiques d'écotypes devrait considérablement contribuer à une meilleure compréhension du rôle de ces gènes dans l'adaptation des cellules à des conditions environnementales particulières.

RÉFÉRENCES

Baird NA et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS One 3:e3376.

Boitard S, Schlotterer C, Nolte V, Pandey RV, Futschik A. 2012. Detecting selective sweeps from pooled next-generation sequencing samples. Mol Biol Evol 29:2177-2186.

Buerkle CA, Gompert Z. 2013. Population genomics based on low coverage sequencing: how low should we go? Mol Ecol 22:3028-3035.

Colombo M *et al.* 2013. The ecological and genetic basis of convergent thick-lipped phenotypes in cichlid fishes. Mol Ecol 22:670-684.

Davey JW. *et al.* 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Rev Genet 12:499-510.

Dufresne *et al.* 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine *cyanobacteria*. Genome Biol 9:R90.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol 28:2239-2252.

Emerson KJ *et al.* 2010. Resolving postglacial phylogeography using high-throughput sequencing. Proc Natl Acad Sci USA 107:16196-16200.

Enard D, Depaulis F, Roest Crollius H. 2010. Human and non human primate genomes share hotspots of positive selection. PloS Genet 6:e1000840.

Enard W, Przeworski M, Fisher SE, Lai CS, Wiebe V, Kitano T, Monaco AP, Pääbo S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. Nature 418:869-872.

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193:929-941.

Flombaum P, *et al.* 2013. Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proc Natl Acad Sci USA 110:9824-9829.

Gayral P *et al.* 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PloS Genet 9:e1003457.

Gompert Z, Buerkle CA. 2011. A hierarchical bayesian model for next-generation population genomics. Genetics 187:903-917.

Heliconius Genome Consortium 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Joron M, Jiggins CD, Papanicolaou A and McMillan WO. 2006. *Heliconius* wing patterns: an evo-devo model for understanding phenotypic diversity. Heredity 97:157-167.

Kramer U. 2010. Metal hyperaccumulation in Plants. Ann Rev Plant Biology 61: 517-534.

Losfeld G, Escande V, Jaffre T, L'Huillier L, Grison C. 2012. The chemical exploitation of nickel phytoextraction: an environmental, ecologic and economic opportunity for New Caledonia. Chemosphere 89:907-910.

Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res 23:23-35.

Nadeau NJ, *et al.* 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. Phil Trans R Soc B-Biological Sciences 367:343-353.

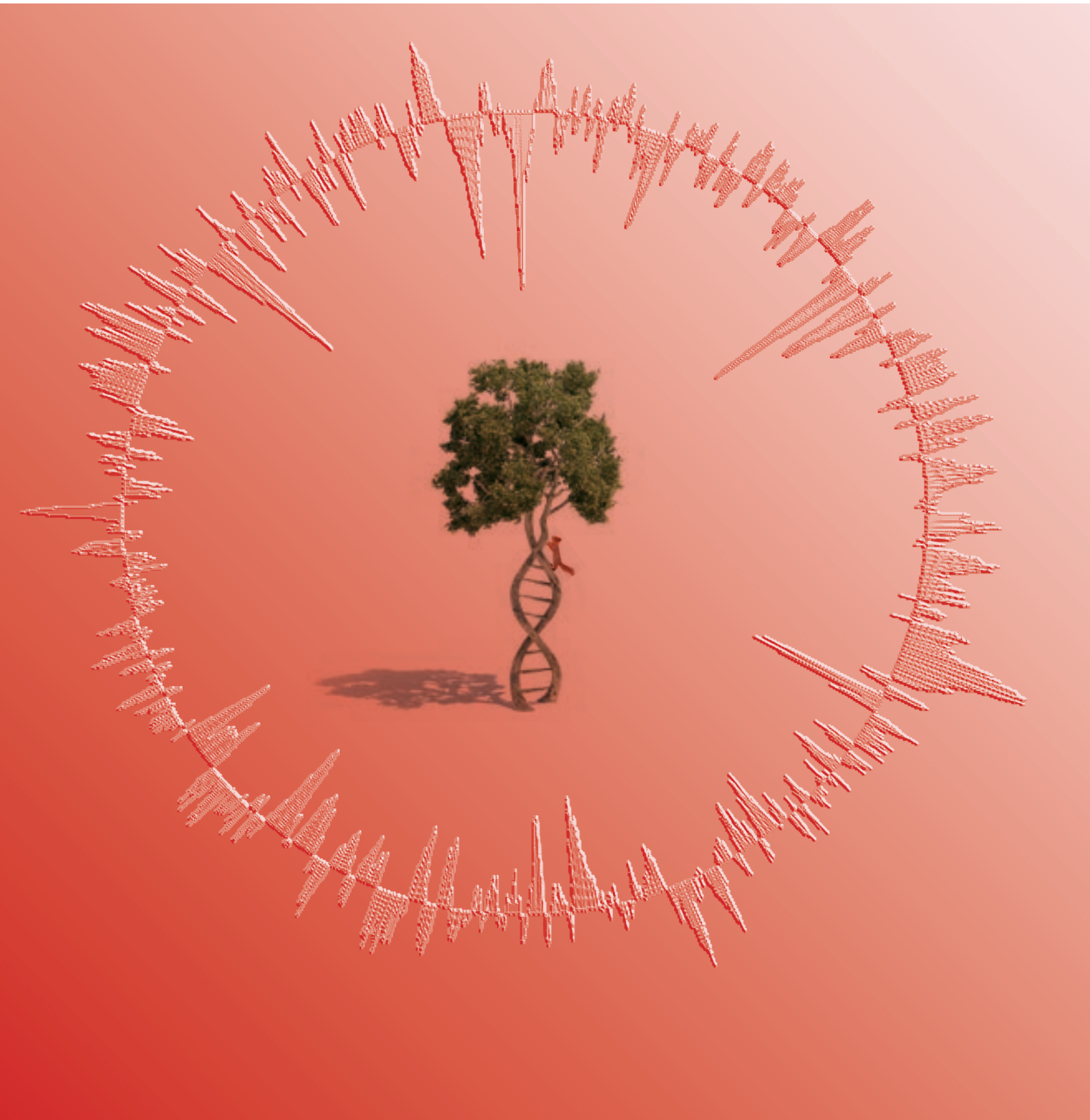
Smadja CM, Canbäck B, Vitalis R, Gautier M, Ferrari J, Zhou JJ, Butlin RK. 2012. Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialisation and speciation in the pea aphid. Evolution 66:2723-2738.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nature Genet 42:260-263.

Verbruggen N, Hermans C, Schat H. 2009. Molecular mechanisms of metal hyperaccumulation in plants. New Phytol 181:759-776.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Appl Biosciences 13:555–556.

Zhan S, Merlin C, Boore JL, Reppert SM. 2011. The monarch butterfly genome yields insights into long-distance migration. Cell 147:1171-1185.



X ECOLOGIE FONCTIONNELLE ET GÉNOMIQUE DES POPULATIONS

Coordinateurs : Denis Faure et Francis Martin
Contributeurs : Didier Bogusz, Annegret Kohler, Xavier Nesme, Philippe Normand, Aurélie Tasiemski

L'association de l'écologie fonctionnelle et de la génomique des populations permet de confronter la connaissance des traits liés à des fonctions centrales des organismes aux connaissances acquises sur la structure et l'expression des génomes. Les informations collectées par différentes approches omiques (génomique, transcriptomique, protéomique*, métabolomique*) éclairent directement les traits fonctionnels des populations au plus près de leurs interactions avec l'environnement et ses composantes biotiques et abiotiques. Les NGS soutiennent l'émergence de nouveaux champs scientifiques appelés Ecologie génomique et Ecologie inverse. Au-delà de l'écologie fonctionnelle, les attendus de ces champs de recherche intéressent la biologie évolutive, la taxinomie, l'écologie des communautés et l'écologie chimique, mais aussi la génomique, la physiologie, la biologie intégrative, ainsi que l'écotoxicologie et l'ingénierie écologique.

La **connaissance des propriétés fonctionnelles** (ou traits fonctionnels) des organismes (Eucaryotes, Procaryotes, Virus) constitue un des **enjeux majeurs de l'écologie fonctionnelle**. Cette connaissance des traits fonctionnels s'inscrit dans la compréhension des interactions des organismes avec leur environnement entendu dans ses composantes biologique, physicochimique, spatiale et temporelle. La démarche d'étude de ces traits est intégrative puisqu'elle aborde leurs déterminants moléculaires, leur variabilité potentielle et exprimée, ainsi que leurs rôles dans le cycle de vie et d'interactions des organismes qu'ils soient considérés au niveau individuel ou populationnel.

Les NGS ont bouleversé ce champ de recherche. D'une part, les NGS ont permis d'accéder aux génomes d'organismes étudiés pour leurs traits fonctionnels par les écologues, mais qui n'étaient pas considérés comme modèles en biologie comme le sont l'homme, la souris, l'arabette, *C. elegans*, ou *E. coli*… De cette interface entre écologie fonctionnelle et génomique s'est développé le champ de l'**écologie génomique** qui « étudie la structure et le fonctionnement du génome dans l'objectif de comprendre les relations entre un organisme et son environnement biotique et abiotique » (van Straalen et Roelofs 2012). Cette approche d'écologie génomique est illustrée par le Focus 10-1. D'autre part, les NGS donnent accès aux données génomiques et transcriptomiques issues de plusieurs individus permettant l'intégration de la dimension populationnelle dans l'étude des traits fonctionnels. Ainsi est née la démarche d'écologie inverse (Li *et al.* 2008) en analogie avec la génétique inverse qui permet de comparer des individus portant un allèle différent d'un gène. **L'écologie inverse** permet de comparer les génomes d'individus issus d'au moins deux populations afin d'en identifier les traits génétiques distinctifs, et ainsi prédire leurs traits fonctionnels distinctifs. Ce paradigme peut être étendu aux comparaisons de transcriptomes. Cette approche est illustrée par le Focus 10-2.

Les connaissances acquises sur les **déterminants moléculaires supports des traits fonctionnels** connus ou prédits doivent être repositionnées au sein des réseaux de régulation génétique et épigénétique des individus et populations soumises aux contraintes environnementales, y compris les interactions entre organismes. Les traits prédits doivent être caractérisés aux niveaux cellulaire et physiologique. Ce questionnement est d'autant plus crucial que des gènes et protéines associés à des traits prédits sont souvent sans fonction connue dans les bases de données. Ces approches moléculaires ne pourront être portées que par des interactions fortes entre écologie, génomique, biochimie et biologie en combinant différentes omiques et approches prédictives ou expérimentales. Ainsi, la caractérisation intégrée des traits fonctionnels constitue un enjeu qui doit mobiliser des expertises au-delà de la communauté d'écologie fonctionnelle stricto sensu (voir chap. XI et XII).

L'analyse des traits fonctionnels des populations au sein des **réseaux complexes des interactions biologiques des écosystèmes** est un autre enjeu porté par l'écologie génomique et l'écologie inverse. L'écologie des populations doit pouvoir appréhender la complexité des écosystèmes dans lesquels vivent les populations dont les traits sont étudiés (Focus 10-3). Ce questionnement stimule des interactions fortes entre écologie des populations et écologie des communautés, notamment, lorsque les organismes et traits fonctionnels s'inscrivent dans des réseaux trophiques, des cycles géochimiques, des relations durables entre organismes (mutualisme, parasitisme, transmission d'hôtes ou de réservoirs) ou encore lorsque les organismes sont étudiés pour leur trait d'effet qui sont estimés comme déterminants pour le fonctionnement d'un écosystème. Ainsi, le positionnement des populations au sein des systèmes biologiques complexes est aussi un moteur d'agrégations d'expertises.

Des données primaires (génomes) ou secondaires (identification et caractérisation des traits fonctionnels) issues de l'écologie génomique et écologie inverse intéressent directement certains champs de la **biologie évolutive, la taxinomie, et l'étude de la biodiversité**. Par exemple, les communautés de biologie évolutive et d'écologie fonctionnelle peuvent s'intéresser à des mêmes déterminants moléculaires étudiés pour leur fonction ou leur valeur adaptative, et entreprendre ainsi leur étude sur une même population ; une approche d'écologie inverse peut également permettre la caractérisation de traits fonctionnels caractéristiques d'un taxon et ainsi contribuer à sa définition (Focus 10-4) ; enfin, la caractérisation des traits et des populations contribue directement au décryptage de la biodiversité.

Les NGS apportent de nouvelles connaissances et outils au service d'**enjeux sociétaux** portés par l'écologie fonctionnelle. Il peut s'agir d'évaluer ou prédire l'effet des changements climatiques ou activités anthropiques sur les populations et leurs propriétés fonctionnelles, de proposer des outils de diagnostic (populations ou activités biologiques bioindicatrices) de ces changements ou activités humaines, y compris dans une approche écotoxicologique, ou bien d'utiliser les connaissances sur les traits fonctionnels des populations dans des approches d'ingénierie écologique ou d'agro-écologie.

FOCUS 10-1

Recherche des déterminants symbiotiques de la bactérie *Frankia* et des plantes actinorhiziennes (*Alnus* et *Casuarina*) par génomique et transcriptomique

Figure 10A. Racines d'*Alnus* (Aulne) montrant une nodosité où se déroulent l'interaction avec la bactérie *Frankia* et la fixation de l'azote atmosphérique.

Il s'agit d'identifier les déterminants symbiotiques de bactéries du genre *Frankia* et ceux des plantes hôtes *Alnus* et *Casuarina* qui sont responsables d'environ 15% des entrées biologiques d'azote sur terre (Figure 10A).

La stratégie développée associe 1) l'obtention et l'analyse des génomes bactériens, 2) la fabrication de banques de marqueurs d'expression des gènes des plantes hôtes à partir de données de séquençage de gènes et transcrits, 3) la fabrication de puces et l'hybridation des transcrits d'origines végétale et bactérienne, et analyse des données transcriptomiques, et 4) la génétique inverse ou additive pour l'analyse de gènes et fonctions d'intérêt avec la construction de mutants de plantes et de mutants bactériens.

Ce travail a révélé que les génomes bactériens de *Frankia* ne contiennent pas les gènes canoniques de nodulation, leurs déterminants symbiotiques sont donc distincts de ceux des bactéries *Rhizobium* qui forment des nodosités avec les légumineuses (Alloisio *et al.* 2010, Hocher *et al.* 2011, Normand *et al.* 2007). Par contre, les plantes actinorhiziennes possèdent presque tous les homologues de la cascade symbiotique des légumineuses avec les *Rhizobium*, dont SymRK qui est essentiel à la nodulation des plantes actinorhiziennes par *Frankia*. Parmi les gènes de plantes surexprimés au cours de la mise en place de la symbiose, il y a plusieurs homologues des défensines codant des peptides anti-microbiens qui sont potentiellement impliquées dans le contrôle du développement de la bactérie symbiotique *Frankia*. Un premier modèle de l'interaction *Frankia*-Aulne est proposé (Figure 10B).

Frankia semble avoir inventé des effecteurs nouveaux pour s'adapter à une voie symbiotique de l'hôte qui apparaît conservée dans ses grandes lignes dans l'ensemble du règne végétal. L'expression de fragments du génome de *Frankia* dans la bactérie *Streptomyces* permettra d'identifier les gènes impliqués dans l'interaction *Frankia-Alnus/Casuarina*. Cette approche aidera à proposer un modèle détaillé du fonctionnement des plantes au cours de la mise en place de la symbiose.

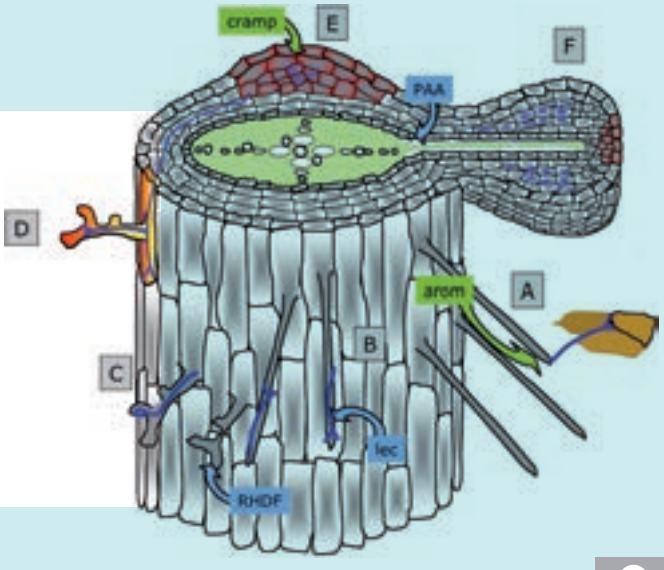


Figure 10B. Schéma des interactions entre *Frankia* et ses plantes symbiotiques. Des composés aromatiques (arom) sont synthétisés (A) par la plante-hôte et sont détectés par la bactérie qui s'attache par des protéines appelées lectines (lec) aux poils racinaires (B). La bactérie augmente la synthèse d'un composé appelé RHDF qui déforme les poils racinaires (C). Les poils racinaires déformés permettent l'internalisation de la bactérie qui se multiplie dans les tissus du cortex. Un effecteur encore inconnu est alors synthétisé et reconnu par des kinases de la plante (dont la kinase SymRK), ce qui déclenche alors une pulsation calcique (D) qui induit le programme de synthèse de protéines de la plante-hôte et la formation d'un prénodule (E). La plante surexprime alors un grand nombre de gènes dont les défensines appelées CRAMP. En parallèle, la bactérie synthétise alors une auxine (PAA) qui déclenche une multiplication cellulaire des cellules du péricycle et l'émergence d'une racine secondaire modifiée, la nodosité (F), siège de la fixation de l'azote moléculaire.

FOCUS 10-2

Dix ans de génomique comparative et de recherche des traits de vie des champignons forestiers : du Sanger aux NGS

Ce programme de génomique comparative a pour objectif d'identifier les liens éventuels entre traits fonctionnels et patrons génétiques chez différents groupes de champignons (saprotrophes, symbiontes, pathogènes) contrôlant le cycle du carbone dans les écosystèmes forestiers.

Figure 10C : *Laccaria amethystina*, de son nom vernaculaire le laccaire améthyste, est un champignon symbiotique ectomycorhizien de la famille des Hydnangiacées (Agaricales, Basidiomycotina). Il est l'un des champignons mycorhiziens séquencés dans le cadre du projet Mycorrhizal Genomics Initiative (http://mycor.nancy.inra.fr/IMGC/MycoGenomes/).

Ces travaux sont développés dans le cadre du programme de génomique fongique, MycoCosm, piloté par le Joint Genome Institute du Département de l'énergie américain (http://genome.jgi-psf.org/programs/fungi). Ce programme a entrepris le séquençage de 1 000 génomes d'espèces fongiques (exemple Figure 10C) afin d'explorer leur diversité génomique et identifier les mécanismes moléculaires déterminant l'évolution de ces microorganismes et en particulier, de leurs traits fonctionnels. MycoCosm concerne les sciences de l'énergie et de l'environnement. Il cible les génomes fongiques dans trois domaines : la santé des végétaux, la bioraffinerie et la diversité fongique.

Le séquençage et l'analyse du génome des symbiotes ectomycorhiziens *Laccaria bicolor* et *Tuber melanosporum*, du champignon de Paris (*Agaricus bisporus*), de l'agent de la rouille du Peuplier et d'une douzaine de pourritures brunes et blanches ont été réalisés (Duplessis *et al.* 2011, Martin *et al.* 2008, Martin *et al.* 2010, Morin *et al.* 2012). La comparaison des répertoires de gènes de ces champignons a mis en évidence le rôle clé des gènes codant les enzymes de dégradation des polysaccharides de la paroi végétal (CAZymes) (Floudas *et al.* 2012). Leur présence et leur distribution sont des facteurs primaires dans l'acquisition du carbone et de ce fait, déterminent la niche écologique de l'espèce. Cette étude a également permis d'établir un scénario décrivant l'évolution des champignons forestiers : l'espèce ancestrale était vraisemblablement une pourriture blanche qui joua un rôle décisif dans la décomposition massive de la lignocellulose issue des premiers arbres il y a 300 M d'années ; ce groupe fonctionnel s'est considérablement diversifié avant de donner naissance aux pourritures brunes et aux décomposeurs de litière. Les champignons ectomycorhiziens seraient issus de ces derniers groupes fonctionnels par acquisition d'effecteurs protéiques contrôlant l'immunité et le développement racinaire de la plante-hôte.

L'analyse en cours du génome d'une cinquantaine de saprotrophes et symbiontes devrait permettre de confirmer le scénario évolutif proposé et d'explorer le continuum entre saprotrophisme et symbiose. L'analyse fonctionnelle des effecteurs symbiotiques, couplée à l'étude de leur évolution par génomique populationnelle, devrait apporter des informations nouvelles sur les mécanismes développés par les champignons ectomycorhiziens afin de contrôler les réactions de défense immunitaire de leur plante-hôte.

A moyen terme, voici quelques pistes sur l'évolution de l'écologie fonctionnelle portée par les NGS :
- Le nombre et la diversité des organismes dont les données génomiques et/ou transcriptomiques seront accessibles vont se multiplier ; les données produites ne seront plus limitées à un seul individu mais concerneront plusieurs individus d'une population d'intérêt : les études des traits fonctionnels se baseront à la fois sur **l'étude d'organismes modèles et de populations modèles.**
- L'écologie fonctionnelle sera questionnée afin de développer des approches intégratives sur la connaissance des traits des organismes au regard de leurs interactions avec l'environnement ; les interfaces entre écologie fonctionnelle et biologie, biologie évolutive, taxinomie, écologie des communautés vont s'intensifier, voire devenir un passage obligé pour la valorisation des données acquises ;
- L'écologie fonctionnelle sera sollicitée pour répondre à des enjeux sociétaux : écotoxicologie, ingénierie écologique et restauration, biodiversité, introduction d'espèces, contrôle des espèces invasives, agro-écologie …

FOCUS 10-3

Transcriptomique du système immunitaire des vers côtiers *Capitella*

Figure 10D : Vers *Capitella capitella* (échelle 10 cm)

La défaillance du système immunitaire, favorisée par les changements environnementaux et l'accroissement de souches pathogènes exotiques plus ou moins résistantes à de nombreux antibiotiques, constitue l'une des menaces les plus sérieuses d'extinction d'espèces dans le domaine marin.
Cette étude porte sur l'impact de l'environnement sur l'immunité d'organismes marins simples (petite taille et anatomie peu complexe), tolérants à la pollution et pour lesquels il est possible de travailler sur un nombre significatif d'individus : les vers Annélides. Ceux des zones marines sont particulièrement adaptés car soumis à des contraintes environnementales extrêmes, thermiques ou chimiques : les estuaires et les habitats hydrothermaux profonds des dorsales océaniques. La forte variabilité de ces environnements sur de courtes échelles spatiales et temporelles permet d'appréhender directement par l'expérience, et en conditions naturelles, les adaptations physiologiques et génétiques ainsi que les interactions organisme-environnement. Les données de transcriptomique obtenues grâce à un financement APEGE 2012 sur *Capitella* (Figure 10D), un annélide côtier de distribution très ubiquiste, modèle en écotoxicologie du fait de sa grande tolérance à la pollution, sont en cours d'exploitation.

Le génome de ce polychète a été séquencé et est disponible, ce qui fait de cet animal un modèle de choix pour des analyses fonctionnelles en transcriptomique. De manière intéressante, il a été observé que certaines *Capitella* prélevées en zone polluée présentaient un développement microbien au niveau tégumentaire. Aucun microorganisme n'a été observé sur le corps des *Capitella* vivant dans un environnement non pollué. Cette colonisation microbienne en milieu pollué traduit soit l'acquisition d'une épibiose (effet bénéfique) soit un état d'infection microbienne (effet délétère). Dans les deux cas, il implique une modification du système immunitaire devenu soit tolérant à la mise en place d'une symbiose microbienne, soit sensible au développement d'un pathogène qu'il ne parvient plus à éliminer. L'approche transcriptomique en RNAseq, vise à identifier et à quantifier les gènes différentiellement exprimés 1) au sein de populations de *Capitella* exposées ou non aux polluants et 2) au sein de populations de *Capitella* exposées aux polluants, colonisées ou non par des microorganismes. Cette étude permettra d'obtenir des informations quant à l'impact in situ de la pollution sur l'immunité antimicrobienne d'organismes marins.

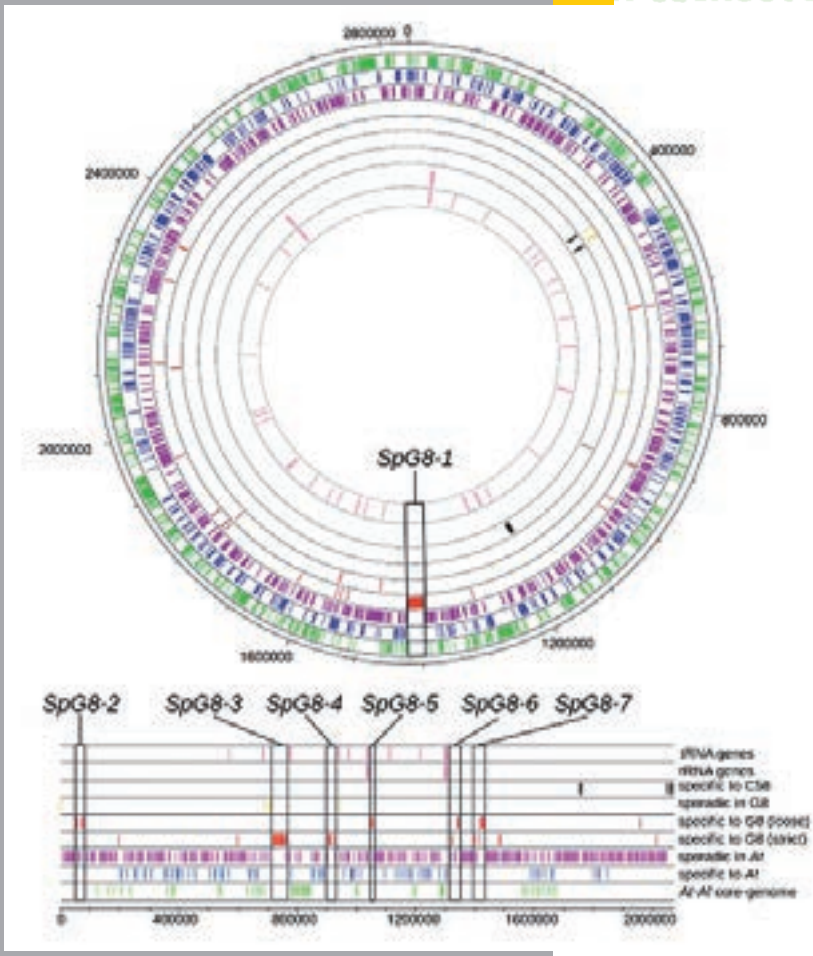
FOCUS 10-4

Découvrir les adaptations spécifiques des espèces bactériennes qui ont conduit à leur spéciation et leur permettent d'échapper à la compétition inter-spécifique

Suite aux hypothèses de Cohan (2006) sur l'effet de la sélection sur l'homogénéité des populations bactériennes, on suppose que les espèces bactériennes définies sur la base de leur proximité génomique (Stackebrandt *et al.* 2002) sont en fait des espèces écologiques adaptées à des niches particulières. Découvrir ces niches spécifiques potentielles permet de comprendre les mécanismes de spéciation tout en pointant les niches écologiques primaires qui permettent aux espèces bactériennes d'échapper à la compétition inter-spécifique. Dans ce projet, il s'agit de tester cette hypothèse en découvrant et caractérisant ces adaptations « spécifiques d'espèce » chez les bactéries du genre *Agrobacterium* constitué de nombreuses espèces sympatriques, génétiquement bien différenciées mais étroitement apparentées.

Une première étape a consisté à identifier les gènes spécifiques de l'espèce *Agrobacterium fabrum* en comparant les contenus géniques d'espèces apparentées dans une approche d'écologie inverse. Après avoir identifié les régions génomiques spécifiques d'espèce, la deuxième étape a consisté à tester expérimentalement les prédictions fonctionnelles de l'annotation experte par génétique inverse afin d'inférer les niches écologiques spécifiques potentielles des espèces.

Les premiers résultats ont montré qu'il y avait 196 gènes spécifiques d'espèce chez *A. fabrum*, groupés dans sept régions génomiques localisées sur les chromosomes circulaires et linéaires. Ces régions codent des voies métaboliques particulières à partir desquelles il a été possible d'inférer une niche spécifique potentielle hypothétique (Lassalle *et al.* 2011). Cette niche se situerait dans la rhizosphère des plantes où *A. fabrum* pourrait utiliser des composés phénoliques et des sucres particuliers, surmonter la carence en fer, résister à des composés toxiques et percevoir des signaux émis par la plante (Figure 10E). Les travaux de génétique fonctionnelle confirment ces prédictions et démontrent que les gènes spécifiques d'*A. fabrum* permettent effectivement à ses membres d'échapper à la compétition avec les espèces apparentées.



Afin de généraliser l'approche, une base de données de famille de tous les gènes orthologues d'*Agrobacterium*, Agrogenom (http://phylariane.univ-lyon1.fr/) a été construite à partir des séquences génomiques de toutes les espèces connues à ce jour chez *Agrobacterium* plus des représentants des autres genres de la famille des Rhizobiacées (Shams *et al.* 2013). Cette base de données

FOCUS 10-4 (suite)

permet de reconstituer l'histoire évolutive des gènes et d'inférer les génomes ancestraux à chaque nœud de la phylogénie*. Elle permet de déterminer les niches écologiques spécifiques potentielles effectives de toutes les espèces d'*Agrobacterium* mais aussi les niches écologiques spécifiques de genre chez les Rhizobiacées.

En conclusion, l'approche d'écologie inverse alliant génomique comparative et génétiques inverse et fonctionnelle a permis de démontrer que l'espèce bactérienne «génomique» est bien une espèce écologique et que la spéciation, au moins chez *Agrobacterium*, est de type parapatrique. Cette approche révèle en outre la nature des niches écologiques primaires — le plus souvent méconnues — qui permettent aux espèces bactériennes d'échapper à la compétition inter-spécifique. Il semble utile et fructueux de généraliser cette approche aux différentes espèces bactériennes afin de tenir compte de leur rôle écologique dans les microbiomes. In fine, il semble indispensable que l'International Committee on Systematic of Prokaryotes (ICSP) demande le séquençage de plusieurs génomes des espèces nouvellement décrites et encourage ce travail pour les espèces plus anciennes. Ceci permettrait d'associer une annotation écologique à chaque espèce bactérienne qui serait de grande utilité pour explorer les potentialités fonctionnelles des métagénomes.

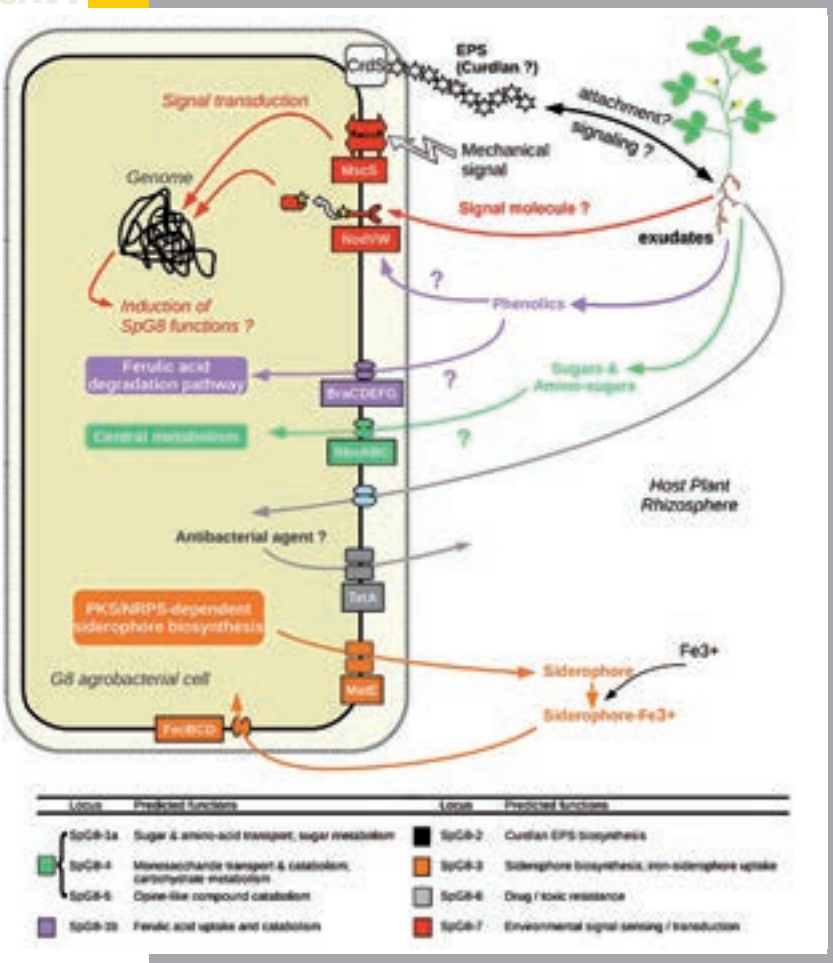


Figure 10E. Cartographie génomique et traits fonctionnels. A gauche, position des régions du génome (SpG8-1 à -8) identifiées comme spécifiques de l'espèce *A. fabrum* ; à droite, traits fonctionnels codés par ces régions spécifiques et leurs rôles dans la niche spécifique hypothétique d'*A. fabrum* montrant une adaptation à des échanges étroits avec une plante (adapté de Lassalle *et al.* 2011).

Les principales limitations aux développements des connaissances dans ce domaine sont les suivantes :

- Si l'acquisition et les outils de comparaison des structures des génomes sont aujourd'hui accessibles, celles des données transcriptomiques sont encore en développement (préparations des ARNm, traitement des données NGS). De plus, l'**acquisition des données transcriptomiques** est entièrement dépendante des conditions environnementales dans lesquelles vivent les populations étudiées, donc d'autant plus difficiles à maîtriser (reproductibilité) et interpréter à mesure que la complexité de l'environnement s'accroît. La confrontation des données des différentes omiques est encore limitée, souvent pour des raisons de coûts, de contraintes techniques, dont l'interfaçage bioinformatique des bases de données, et de manque de soutien à la fédération d'expertises différentes.
- Une forte contrainte, qui peut devenir un défi majeur en écologie fonctionnelle, est l'accumulation dans les bases de données de **gènes et protéines de fonctions totalement inconnues**. Il s'agit d'un frein réel à l'analyse des traits fonctionnels des individus et populations, mais aussi dans le cadre de l'étude des métagénomes et métatranscriptomes en écologie des communautés. Le décodage des fonctions inconnues est un travail à haut risque et consommateur de temps et de moyens. Il requiert de multiples expertises et ne peut pas être entrepris dans le cadre de projets de courtes durées.

Voici quelques propositions de stratégies pour résoudre les défis scientifiques et techniques :

- **Développement et soutien de réseaux** (nationaux et internationaux) dédiés à l'étude de populations d'intérêt (de la collecte d'échantillons à la mise à disposition des données moléculaires et environnementales) permettant l'association d'expertises différentes et la mise en place et l'entretien de bases de données génomiques et fonctionnelles. Cette action peut être conduite en renforçant les liens et le partage de données entre communautés scientifiques différentes travaillant des objets communs.
- Mise en place de chaînes d'expertises pluridisciplinaires permettant l'étude de gènes et protéines de fonctions inconnues, et qui sont identifiés comme liés à des traits fonctionnels clés.
- Soutenir la **formation permanente des personnels** afin d'accéder à un socle commun de connaissance sur l'utilisation des données NGS, et favoriser les échanges d'expertises et savoir-faire entre personnels scientifiques, notamment dans un cadre intergénérationnel.

RÉFÉRENCES

Alloisio N, Queiroux C, Fournier P, Pujic P, Normand P, Vallenet D, Médigue C, Yamaura M, Kakoi K, Kucho KI. 2010. The *Frankia alni* symbiotic transcriptome. Mol Plant Microbe Interact 23:593-607.

Cohan FM. 2006. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. Trans R Soc Lond B Biol Sci 361:1985-1996.

Duplessis S *et al.* 2011. Obligate biotrophy features unraveled by the genomic analysis of the rust fungi, *Melampsora larici-populina* and *Puccinia graminis f. sp. tritici*, Proc Natl Acad Sci USA 108: 1966-1972.

Floudas D *et al.* 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science 336: 1715.

Hocher V *et al.* 2011. Transcriptomics of actinorhizal symbioses reveals homologs of the whole common symbiotic signaling cascade. Plant Physiol 156:1-12.

Lassalle F *et al.* 2011. Genomic species are ecological species as revealed by comparative genomics in *Agrobacterium tumefaciens*. Genome Biol Evol 3:762-781.

Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse ecology" and the power of population genomics. Evolution 62:2984–2994.

Martin F *et al.* 2008. Symbiosis insights from the genome of the mycorrhizal basidiomycete *Laccaria bicolor*. Nature 452: 88-92.

Martin F *et al.* 2010. Périgord Black Truffle genome uncovers evolutionary origins and mechanisms of symbiosis. Nature 464 : 1033-1038.

Morin E *et al.* 2012. The genome sequence of the Button Mushroom *Agaricus bisporus* reveals mechanisms governing adaptation to a humic-rich ecological niche. Proc Natl Acad Sci USA 109 : 17501-17506.

Normand P *et al.* 2007. Genome characteristics of facultatively symbiotic *Frankia sp.* strains reflect host range and host plant biogeography. Genome Research 17:7-15

Shams M, Vial L, Chapulliot D, Nesme X, Lavire C. 2013. Rapid and accurate species and genomic species identification and exhaustive population diversity assessment of *Agrobacterium spp.* using recA-based PCR. Syst Appl Microbiol 36:351-358.

Stackebrandt E *et al.* 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int J Syst Evol Microbiol. 52:1043-1047.

Van Straalen M, Roelofs D. 2012. An Introduction to Ecological Genomics, OUP Oxford Editor, second edition.

XI FONCTIONNEMENT DES ÉCOSYSTÈMES : MÉTAGÉNOMIQUE ET INTÉGRATION DES OMIQUES

Coordinateurs : Philippe Bertin et Télesphore Sime-Ngando
Contributeurs : Didier Debroas, Denis Le Paslier, Roland Marmeisse, Sébastien Monchy, Frédéric Plewniak

A l'interface entre la biologie moléculaire et l'écologie, ont émergé des techniques révolutionnaires basées sur le séquençage génomique, allant de la description d'une cellule à celle de communautés biologiques complexes prélevées dans leur environnement. Associées à des approches fonctionnelles globales (métatranscriptomique, métaprotéomique, métabolomique, Figure 11A), ces techniques permettent aujourd'hui d'accroître la compréhension du fonctionnement des écosystèmes (Focus 11-1), notamment par la prise en compte d'organismes récalcitrants à l'isolement en culture (single-cell genomics, Focus 11-3). Or, ceux-ci peuvent former l'essentiel de la communauté microbienne dont les fonctions sont à la base de la pérennité de notre environnement et des services écosystémiques associés (Bertin *et al.* 2011). Ainsi, par l'intégration de différents niveaux d'organisation de la biodiversité spécifique et fonctionnelle (molécules, gènes, populations, communautés…), la génomique environnementale permet, grâce à une synergie entre biologistes, géochimistes et bioinformaticiens, de mieux comprendre les relations entre biodiversité et fonctions au sein d'un réseau trophique, en lien avec les facteurs biotiques et abiotiques de l'environnement (Sime-Ngando et Niquil 2011).

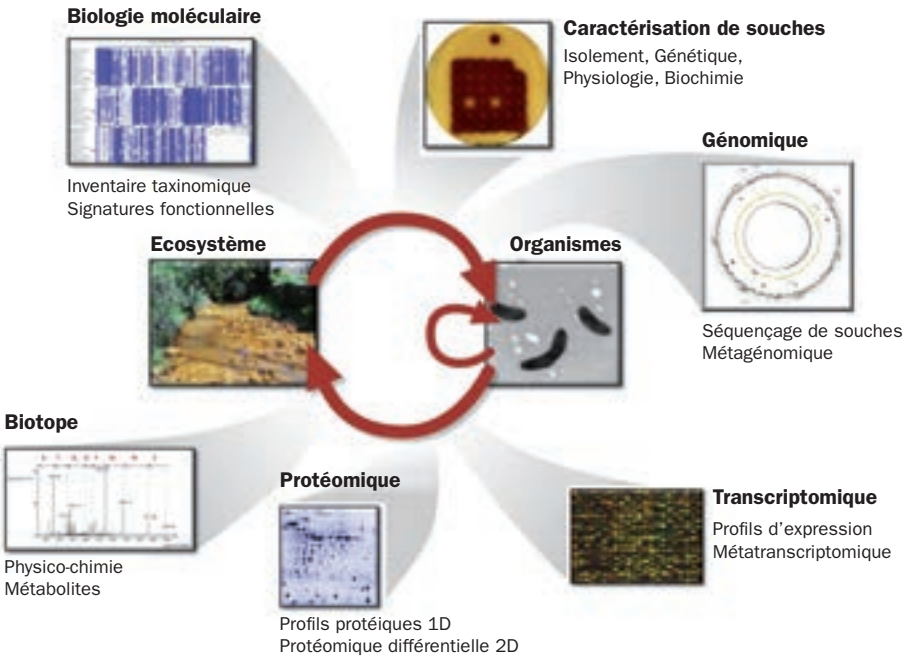


Figure 11A. Différentes facettes de la génomique appliquées aux microorganismes et aux écosystèmes. Ces approches combinées permettent d'obtenir une image intégrée de leur structure et de leur fonctionnement.

FOCUS 11-1

La métatranscriptomique « à la rescousse » des microorganismes eucaryotes

L'approche métatranscriptomique représente l'approche de choix pour l'étude des communautés microbiennes eucaryotes : « protistes » prototrophes ou auxotrophes, champignons, mais aussi mésofaune (nématodes, acariens, collemboles…). Ces organismes jouent des rôles majeurs aussi bien dans le fonctionnement des milieux aquatiques (photosynthèse, consommateurs primaires) que terrestres (dégradation de la biomasse végétale, pathogènes et symbiontes des végétaux et des animaux). Les grandes tailles et la complexité structurale des génomes eucaryotes rendent plus difficile l'annotation des séquences d'ADN environnemental d'origine eucaryote caractérisées par la présence d'introns et l'abondance de séquences répétées, parmi lesquelles des transposons. Les ARNm eucaryotes qui présentent la propriété unique d'être polyadénylés peuvent être toutefois spécifiquement isolés et analysés séparément des autres ARN environnementaux (ARN ribosomiques majoritaires et ARNm bactériens).

Le séquençage systématique « haut débit » des ADNc est l'approche privilégiée pour appréhender la complexité des métatranscriptomes. Les techniques actuelles qui génèrent des séquences courtes (<500pb, voir seulement 100-150pb) rendent difficile leur assemblage pour obtenir des séquences géniques complètes. Ceci peut être envisagé pour des communautés microbiennes dominées par un nombre limité de taxa (environnements très pollués) ou celles dont l'étude rassemble une communauté scientifique importante qui s'investit dans le séquençage de nombreux génomes de référence (ex : le microbiome humain). Le séquençage de métatranscriptomes conduit aujourd'hui à dresser un inventaire de gènes auxquels peuvent être associées des fonctions, mais auxquels il apparaît hasardeux d'associer une origine taxonomique. Au-delà du perfectionnement des méthodes d'analyses bioinformatiques, il est nécessaire de soutenir les initiatives de séquençage systématique de génomes d'espèces représentatives de la diversité du vivant pour permettre d'associer les fonctions observées à des groupes taxonomiques.

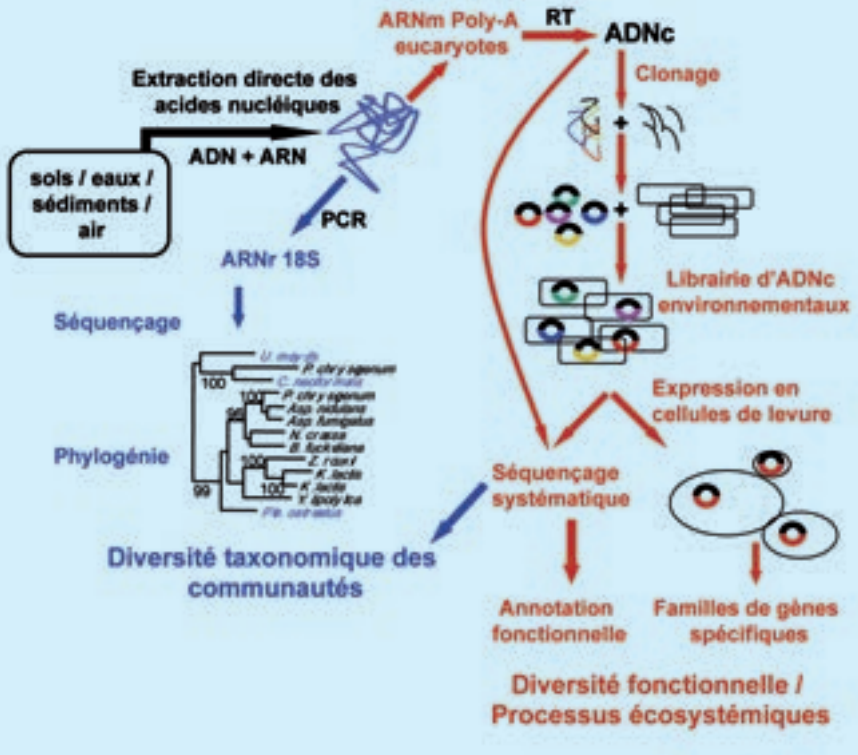


Figure 11B. Les différentes facettes de l'approche métatranscriptomique : des ARNm d'origine eucaryote extraits d'un échantillon environnemental à l'expression des ADNc environnementaux dans une levure.

FOCUS 11-1 (Suite)

Quelle fiabilité peut-on accorder aux annotations in silico, et qu'en est-il des très nombreuses séquences sans homologues dans les bases de données ? Au-delà du séquençage systématique, il est nécessaire de promouvoir des analyses expérimentales des séquences environnementales susceptibles de révéler de nouveaux processus biologiques réalisés au sein des écosystèmes (Figure 11B). Ainsi, en exprimant des ADNc eucaryotes « de sols » dans des souches de levure, une nouvelle famille de transporteurs de peptides pouvant servir à l'acquisition de l'azote organique par les champignons du sol a été décrite (Damon *et al.* 2011). L'analyse in silico de ces séquences prédisait pourtant qu'il s'agissait de transporteurs d'acides aminés. La même approche a aussi permis de caractériser de nouveaux gènes de résistance aux métaux lourds importants dans l'adaptation des communautés microbiennes à des pollutions de l'environnement (Lehembre *et al.* 2013).

La métatranscriptomique, tout comme la métagénomique, ne doit ainsi pas être considérée comme la science du « tout séquençage » ; cette discipline doit combiner différentes approches complémentaires pour pleinement rendre compte de la complexité des processus biologiques à l'œuvre dans les écosystèmes. La métatranscriptomique n'est pas non plus une discipline à elle seule, cette approche doit s'ajouter à d'autres approches d'écologie fonctionnelle telles que des mesures de flux et de stock de nutriments.

En raison du haut débit des outils de la génomique, leur utilisation dans le domaine environnemental permet d'étendre considérablement les échelles d'étude, comparativement aux techniques classiques de la biologie des organismes et des populations. En conjonction avec une application a priori aisée à tous les types d'écosystèmes de la biosphère et la prise en compte de l'ensemble de leurs acteurs biologiques, y compris les virus, ces outils devraient permettre d'aborder des questions de changements d'échelle biologique (du gène à la communauté) et écosystémiques (variations spatio-temporelles), tout en produisant des données intégrant les temps de générations, même les plus courts. Cet enjeu de la **prise en compte des changements d'échelle,** mais aussi de leurs interfaces, est fondamental pour la microbiologie environnementale. Cette discipline est en effet confrontée, aujourd'hui, à deux carences majeures dans le contexte global des sciences de l'environnement, qui sont l'absence 1) de théories écologiques microbiennes recoupant ou pas les grandes théories écologiques et 2) de patterns biogéographiques et temporels recoupant ou pas ceux connus chez les animaux et les plantes. Ces carences ont des conséquences considérables. Par exemple, elles font que les microorganismes, et plus encore les virus, ne sont pas réellement pris en compte dans les problématiques du changement global (CO_2 et autres gaz à effet de serre, climat, acidification, ozone, eutrophisation, hypoxie, invasion, dispersion de xénobiotiques, …), en raison de la méconnaissance de patrons microbiens à différentes échelles de temps, d'espace et d'organisation. Dans ce cadre scientifique général de la **biogéographie des organismes**, la mise en cohérence des séquences moléculaires avec les phénotypes microscopiques se développant dans l'environnement représente un enjeu technologique et scientifique majeur (Focus 11-2).

Outre des outils permettant d'établir des empreintes génétiques (« barcoding » chap VIII, puces à ADN), la profondeur de séquençage qu'offrent les nouvelles technologies de la génomique permet de prendre en compte les **espèces cryptiques** au sein de complexes spécifiques (diversité infraspécifique) et les espèces moins abondantes et moins représentées dans l'écosystème (**biosphère rare**). Chez les microorganismes par exemple, ces espèces peuvent constituer un **réservoir de nouveaux gènes** dont le potentiel évolutif et fonctionnel peut assurer le maintien de la structure et du fonctionnement de l'écosystème et jouer le rôle d'une véritable **assurance écologique** pour la résilience, la stabilité et la pérennité de notre environnement (Focus 11-4). Conjuguées à une indispensable expérimentation en conditions de laboratoire et in situ, de telles approches requièrent aussi le

développement, en amont, de méthodes d'échantillonnage et d'extraction et, en aval, de bases de données et de méthodes permettant non seulement le stockage et l'échange de données sous des formats standardisés, mais aussi – et surtout – leur analyse intégrée (assemblage, binning, annotation, comparaison…), en particulier dans le cas des lectures de faible longueur obtenues à partir d'environnements complexes.

Ainsi, la génomique environnementale appliquée au fonctionnement des écosystèmes devrait permettre, bien au-delà de l'établissement de simples inventaires d'objets biologiques (espèces, gènes…), 1) de prendre en compte les associations symbiotiques caractéristiques des organismes de l'environnement (interactomique, notamment dans le dialogue moléculaire des systèmes procaryote-eucaryote, Focus 11-5), 2) d'appréhender les processus adaptatifs et le rôle éventuel de la plasticité génomique des microorganismes, 3) de relier la dynamique de ces écosystèmes au développement social, économique et culturel associé à cette dynamique, et 4) d'anticiper leur évolution en réponse aux contraintes, y compris celles d'origine anthropique, à l'aide de modèles prédictifs élaborés à partir de la compréhension de leurs propriétés.

FOCUS 11-2

Structure des communautés

Une connaissance approfondie des espèces présentes et de leurs potentiels métaboliques est indispensable pour comprendre le fonctionnement d'un écosystème. Pendant longtemps, les études de diversité étaient réalisées par identification au microscope ou en cultivant les espèces. C'est seulement à partir des années 80 que la biologie moléculaire, via le clonage/séquençage couplé à des analyses phylogénétiques, a permis l'identification des espèces sans devoir préalablement les cultiver. Cette approche est longue, fastidieuse, coûteuse, et surtout ne permet d'identifier que la partie émergée de l'iceberg « biodiversité » en fonction du nombre de clones isolés. Récemment, les études de diversité ont bénéficié du développement de technologies de séquençage à haut débit qui permettent d'accéder directement à la diversité structurale et/ou fonctionnelle des communautés microbiennes. Toutefois, dans sa compréhension du fonctionnement d'un écosystème, l'écologue moléculaire est confronté à deux défis 1) relier données de séquence et données morphologiques 2) s'assurer que les approches moléculaires et traditionnelles donnent des résultats qualitativement et quantitativement comparables.

La structure des communautés eucaryotes microbiennes a été analysée en milieu aquatique marin par le biais d'observation au microscope et d'approches moléculaires. L'ADN génomique total a été extrait d'échantillons marins, puis amplifié par PCR avec des amorces eucaryotes universelles ciblant les régions hypervariables du gène codant l'ARNr 18S (Sogin *et al.* 2006). Cette méthode de barcoding (voir chap. VIII), couplée à une analyse avec des outils bioinformatiques performants, a permis d'identifier de manière exhaustive la composition des communautés microbiennes. Cette diversité moléculaire a ensuite été comparée aux données morphologiques obtenues au microscope.

L'approche moléculaire a permis de capturer le changement de fréquences relatives des taxons les plus abondants en cohérence avec les observations au microscope (Figure 11C). Les deux approches sont toutefois complémentaires dans leur estimation de la diversité puisque certaines espèces n'ont été identifiées que par l'une ou l'autre des méthodes. L'observation au microscope permet d'identifier les espèces dont le gène ARNr n'est pas séquencé ou mal amplifié par le couple d'amorces choisi. Inversement, la méthode moléculaire a permis de révéler les espèces pour lesquelles la description morphologique était inexistante ou pas suffisamment discriminante d'un point de vue taxonomique. Enfin, les espèces rares ou celles nécessitant des préparations particulières pour leur observation au microscope n'ont pu être détectées que via l'approche moléculaire (Monchy *et al.* 2012).

L'utilisation conjointe d'approches classiques et moléculaires récentes permet une identification précise de la diversité structurelle d'un écosystème. Toutefois, ces approches ne renseignent pas sur le rôle des microorganismes dans la communauté. Une approche métagénomique est essentielle pour mettre en évidence la diversité fonctionnelle d'une communauté et sa variabilité saisonnière. Les défis seront d'optimiser et de comparer les données obtenues en biologie moléculaire (métagénomique, barcoding,…) et en microscopie afin d'estimer la composition des communautés pour affilier les données de séquences à la morphologie microscopique (voir aussi chap. VII). Cette étape est cruciale dans la compréhension des réseaux trophiques et du fonctionnement global d'un écosystème.

Figure 11C. Composition de communautés planctoniques. Les analyses ont été réalisées avant (clair) et après (foncé) le bloom printanier de *Phaeocystis* soit par pyroséquençage (à gauche), soit par observation microscopique (à droite). D'après Monchy *et al.* 2012.

FOCUS 11-3

Single Cell Genomics

La compréhension de la diversité génétique des microorganismes procaryotes est limitée aux génomes des organismes cultivés et estimée entre 1% et 10% des espèces bactériennes connues. Les progrès récents du séquençage de l'ADN de cellules bactériennes isolées (Single Cell Genomics ou SCG) sont en train d'accélérer la découverte de microbes non encore cultivés, fournissant des assemblages génomiques pour des espèces connues uniquement par des séquences d'ARNr 16S et métagénomiques (Pelletier *et al.* 2008).

Les étapes de l'approche SCG sont les suivantes : isolement des cellules soit par micromanipulation, soit par dilution ou tri automatisé et leur lyse, amplification de l'ADN par WGA (Whole Genome Amplification) et le plus souvent par MDA (Multiple Displacement Amplification en utilisant les ADN polymérases Phi29 ou Bst DNA) afin de construire des banques pouvant être séquencées par NGS, sélection des génomes amplifiés d'après leur profil taxonomique (gène de l'ARNr 16S par exemple), séquençage et assemblage. Cette approche, en cours de développement, souffre de quelques défauts et biais. Parmi ceux-ci, citons les problèmes de lyse, les biais engendrés par l'amplification MDA (formation de chimères) et les contaminations lors de l'isolement des cellules et des réactifs utilisés lors de l'amplification (Blainey 2013). Certains de ces biais peuvent être réduits par la décontamination des réactifs, la miniaturisation (microfluidique) et donc la réduction des volumes réactionnels (de quelques nanolitres à quelques dizaines de picolitres).

Le taux de complétion des génomes obtenus est variable (0 à 100%) et dépend de nombreux facteurs comme la qualité de la lyse, les contaminations, la complexité des génomes. L'approche SCG est très importante dans le domaine de la santé car il devient possible de caractériser des pathogènes difficiles à cultiver à partir d'une cellule unique. Ceci est d'importance car dans certains cas les cultures réalisées à partir de colonies isolées peuvent conduire à l'apparition de variants non représentatifs à croissance rapide ou ayant perdu des îlots de pathogénicité (Seth-Smith *et al.* 2013). Cette méthodologie peut s'appliquer également aux eucaryotes (Yoon *et al.* 2011). Il n'existe à ce jour que deux structures proposant des services de «single cell genomics». Il s'agit du Bigelow Laboratory's Single Cell Genomics Center (bigelow.org/scgc) et du U.S. DOE Joint Genome Institute (www.jgi.doe.gov) au niveau international. Le premier plateau technique français est en cours d'installation à Orsay. Notons également que les méthodes SGC n'ont pas encore été adaptées aux organismes anaérobies.

Même si l'approche SGC n'est pas encore une méthode simple et bien rodée, elle est en plein développement et jouera certainement un rôle important dans la compréhension du fonctionnement des communautés microbiennes complexes. Ces méthodes sont également en train d'être développées pour l'étude de cellules humaines uniques et auront de nombreuses retombées dans le domaine médical.

FOCUS 11-4 : La biosphère rare microbienne

La détermination de la structure des communautés (abondance, richesse, composition) d'un écosystème est un enjeu central en écologie (voir Focus 3.2, 3.3, 3.4). Les travaux pionniers de Sogin *et al.* (2006), basés sur les nouvelles techniques de séquençage ont permis de détecter entre 1184 et 3290 OTUs en milieu marin, suggérant ainsi que la richesse spécifique détectée jusqu'alors (~200 OTUs) dans ces écosystèmes était certainement sous-estimée. En outre, la plus grande partie de la biodiversité est représentée par un grand nombre d'OTUs présents en faible quantité, constituant la biosphère rare. Les détracteurs de ce concept pensaient que ces OTUs pouvaient être uniquement des artefacts liés à la technique de séquençage, ou ne concernaient que des cellules mortes ou en transit. Or, récemment, deux études ont mis en évidence que les espèces rares pouvaient être actives montrant ainsi que l'étude de cette biodiversité est importante pour la compréhension du fonctionnement des écosystèmes.

La stratégie est basée sur du séquençage à haut-débit d'amplicons (454-Roche). Afin d'éviter certains biais dans l'évaluation de la richesse, une séquence connue de la petite sous unité de l'ADNr, absente de l'écosystème étudié, est introduite avant l'amplification à hauteur de 1 % des ADNr totaux. L'utilisation de ce standard interne, permet après pyroséquençage de déterminer les erreurs liées au séquençage et à l'amplification, mais aussi de déterminer un seuil de clusterisation et de normaliser les données. Les séquences sont affiliées par une méthode phylogénétique originale adaptée au séquençage haut-débit (Taib *et al.* 2013).

L'affiliation phylogénétique appliquée au séquençage haut débit permet de mettre en évidence des clades appartenant à la biosphère rare jamais détectés auparavant. A titre d'exemple, on peut citer parmi les Chlorophyceae, un clade affilié aux Mamiellales (Taib *et al.* 2013) ou au sein des Thaumarchaeota les clades MGI.C et MGI.D (Hugoni *et al.* 2013). L'étude à long terme de l'activité de ces microorganismes (ADNr vs ARNr) permet alors de différencier 3 fractions distinctes. Une fraction qui devient abondante et active en fonction des saisons et qui représente un pool de cellules dormantes (seed bank) indispensable au fonctionnement de l'écosystème. Une seconde fraction toujours rare et active, et une dernière, inactive, contenant des séquences ayant une faible identité avec les bases de données, constituée de ce fait d'*Archaea* étrangères à l'écosystème (Figure 11D). La répartition spatiale (i.e. biogéographie) de ces microorganismes rares montre que celle-ci dépend significativement de la distance géographique entre les écosystèmes et non des conditions environnementales.

Ces études sur la biosphère rare mettent en évidence une communauté jusqu'à présent inconnue, en partie active dans les écosystèmes et pouvant présenter une répartition géographique limitée. Ces organismes peuvent représenter un réservoir de gènes encore inconnus dont la dynamique reste à déterminer. Ces espèces rares pourraient avoir un rôle fonctionnel dans l'écosystème en tant que tel (rare actif) ou du fait de leur dynamique (transitions entre rare et dominant). Ces transitions seraient alors capitales pour prédire le fonctionnement des écosystèmes dans le cadre du changement global.

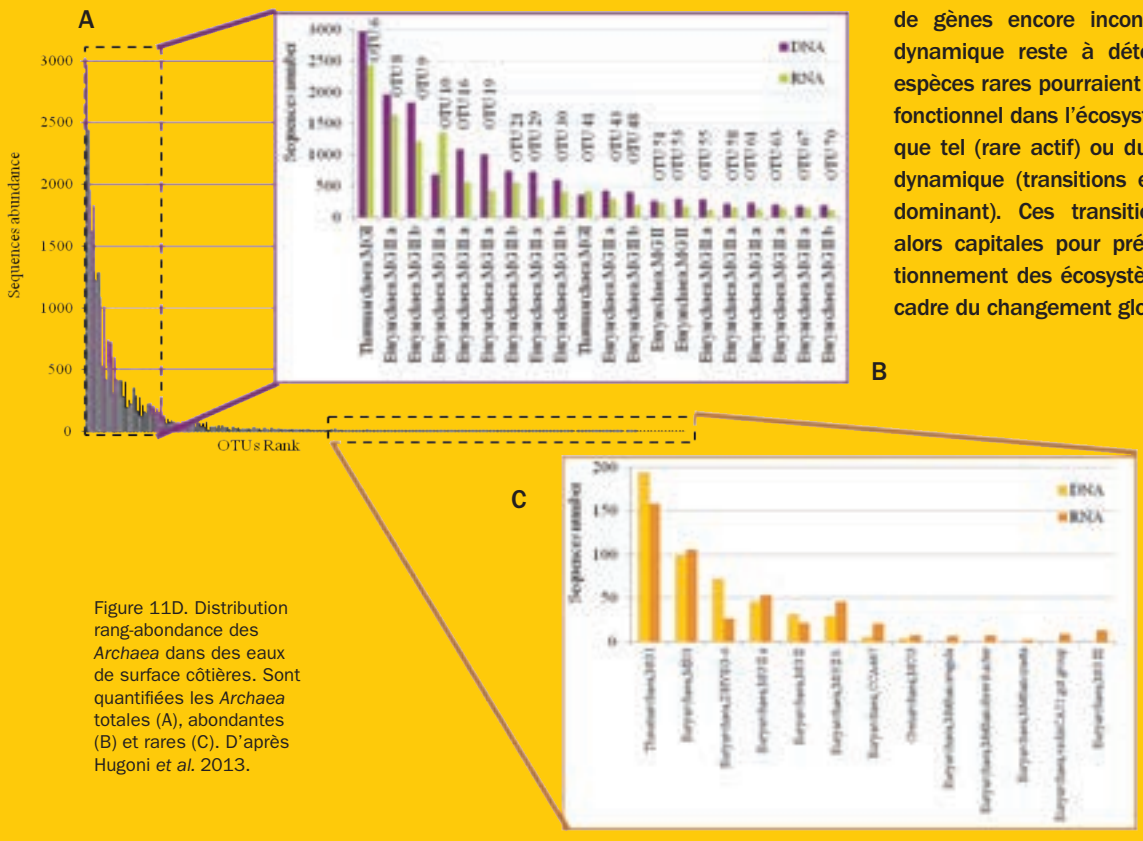


Figure 11D. Distribution rang-abondance des *Archaea* dans des eaux de surface côtières. Sont quantifiées les *Archaea* totales (A), abondantes (B) et rares (C). D'après Hugoni *et al.* 2013.

FOCUS 11-5

Des données multi-omiques au modèle métabolique de la communauté

L'objectif principal du projet « Combinaison de Modèles du Métabolisme pour l'Etude des Relations entre *Candidatus Fodinabacter communificans* et *Euglena mutabilis* » (CoMMERCE) est de déterminer, par des méthodes complémentaires expérimentales et de modélisation métabolique sous contraintes, les modalités d'interaction trophique d'une communauté microbienne constituée d'un eucaryote (*Euglena mutabilis*) isolé du drainage minier acide de Carnoulès (Gard) et d'une bactérie (*Candidatus Fodinabacter communificans*) identifiée sur le même site.

Ca. Fodinabacter communificans fait partie des 7 souches majoritaires identifiées à Carnoulès et dont le génome a été reconstitué par assemblage des séquences métagénomiques (Bertin *et al.* 2011). Des approches de protéomique et métabolomique menées sur *Euglena mutabilis* ont montré qu'elle sécrète dans le milieu des métabolites qui pourraient être utilisés par d'autres microorganismes (Halter *et al.* 2012), en particulier par *Ca. Fodinabacter communificans*. Cette dernière pourrait en retour lui fournir la cobalamine pour laquelle elle est auxotrophe.

Un modèle du réseau métabolique combiné des deux organismes et de leur environnement est en cours d'élaboration. Ce modèle est basé sur l'intégration des données multi-omiques obtenues précédemment (métagénomique, protéomique, métabolomique, transcriptomique) et des données complémentaires obtenues durant le projet (Figure 11E). Le modèle élaboré sera étudié informatiquement par analyse de flux (Oberhardt *et al.* 2009) (Flux Balance Analysis, Flux Variation Analysis et Flux Coupling Analysis) pour identifier les réactions importantes et couplées. Il s'agira de caractériser les interactions entre *E. mutabilis*, *Ca. Fodinabacter communificans* et leur environnement au-delà des mécanismes individuels de résistance ; et également de mettre en évidence les éventuels métabolites échangés, les effets potentiels du métabolisme d'un partenaire sur celui de l'autre et sur la capacité de ces deux organismes à répondre aux conditions qui règnent sur le site d'étude (acidité, concentration en arsenic). Les résultats des analyses informatiques seront validés expérimentalement par des analyses génomiques et des études physiologiques.

Le projet s'attache à modéliser pour la première fois au niveau métabolique les interactions entre des organismes de l'environnement dont le génome n'est que partiellement ou pas séquencé à partir de données de génomique fonctionnelle. Le rôle joué par des partenaires supplémentaires dans les échanges sera également pris en compte en considérant cette communauté de manière globale, ce qui n'a jamais été réalisé jusqu'à présent, les études publiées à ce jour se limitant strictement à deux partenaires (Stolyar *et al.* 2007).

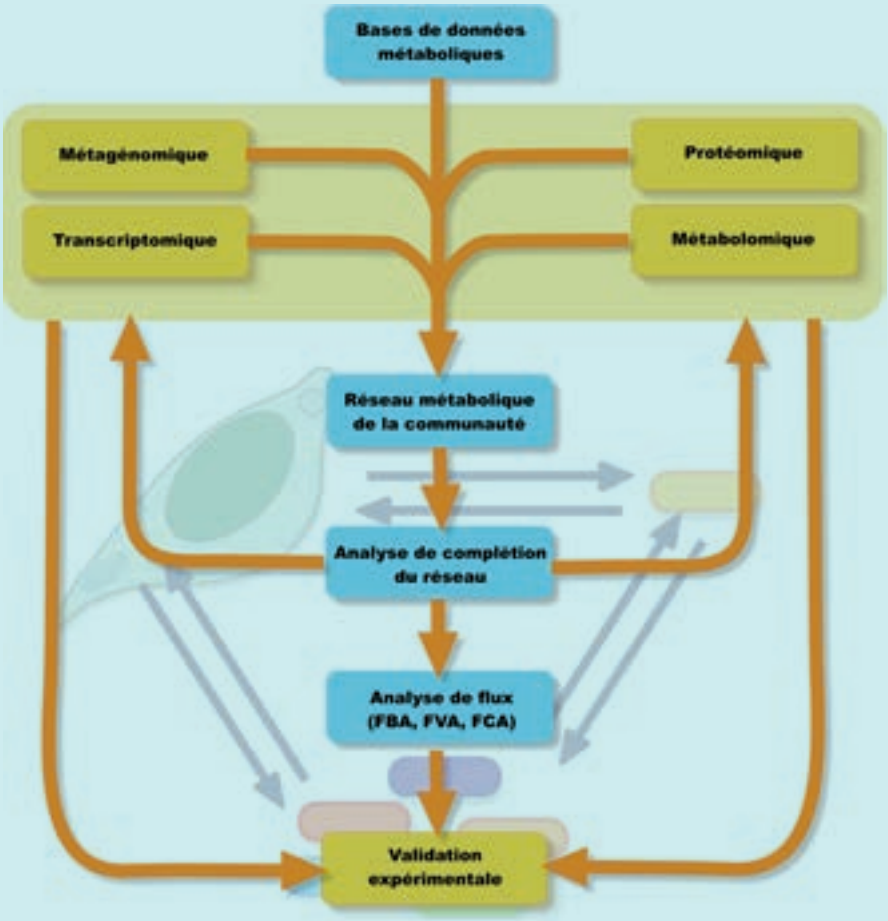


Figure 11E. Stratégie de modélisation des interactions dans une communauté à partir de données multi-omiques. Les boîtes à fond jaune représentent les aspects expérimentaux (production de données et validation), les boîtes à fond bleu correspondent aux données et analyses bioinformatiques.

RÉFÉRENCES

Bertin P *et al.* 2011. Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics. ISME J 5:1735–1747.

Blainey PC. 2013. The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev 37:407-427.

Damon C, Vallon L, Zimmermann S, Haider MZ, Galeote V, Dequin S, Luis P, Fraissinet-Tachet L, Marmeisse R. 2011. A novel fungal family of oligopeptide transporters identified by functional metatranscriptomics of soil eukaryotes. ISME J 5:1871-1880.

Halter D *et al.* 2012. In situ proteo-metabolomics reveals metabolite secretion by the acid mine drainage bio-indicator, *Euglena mutabilis*. ISME J 6:1391-1402.

Hugoni M, Taib N, Debroas D, Domaizon I, Jouan Dufournel I, Bronner G, Salter I, Agogué H, Mary I, Galand PE. 2013. Structure of the rare archaeal biosphere and seasonal dynamics of active ecotypes in surface coastal waters. Proc Natl Acad Sci USA 110:6004-6009.

Lehembre F *et al.* 2013. Soil metatranscriptomics for mining eukaryotic heavy metal resistance genes. Environ Microbiol 15:2829-2840.

Monchy S, Grattepanche JD, Breton E, Meloni D, Sanciu G, Chabé M, Delhaes L, Viscogliosi E, Sime-Ngando T, Christaki U. 2012. Microplanktonic community structure in a coastal systemrelative to a *Phaeocystis* bloom inferred from morphological and tag pyrosequencing methods. PloS One 7:e39924.

Oberhardt MA, Chavali AK, Papin JA. 2009. Flux Balance Analysis: Interrogating Genome-Scale Metabolic Networks. Maly, I. V. ed. In Systems Biology Vol.500, Humana Press: Totowa, NY.

Pelletier E *et al.* 2008. « *Candidatus Cloacamonas acidaminovorans* »: genome sequence reconstruction provides a first glimpse of a new bacterial division. J Bacteriol 190:2572-2579.

Seth-Smith HM *et al.* 2013. Whole-genome sequences of *Chlamydia trachomatis* directly from clinical samples without culture. Genome Res 23:855-866.

Sime-Ngando T, Niquil N (Editors) 2011. Disregarded microbial diversity and ecological potentials inaquatic systems. Series Developments in Hydrobiology 216, Springer, The Netherlands.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored «rare biosphere». Proc Natl Acad Sci USA 103: 12115-12120.

Stolyar S, Van Dien S, Hillesland KL, Pinel N, Lie TJ, Leigh JA, Stahl DA. 2007. Metabolic modeling of a mutualistic microbial community. Mol Syst Biol 3:92.

Taib N, Mangot JF, Domaizon I, Bronner G, Debroas D. 2013. Phylogenetic affiliation of SSU rRNA genes generated by massively parallel sequencing: new insights into the freshwater protist diversity. PLoS One 28:e58950.

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D. 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332:714-717.

XII DES DONNÉES HAUT DÉBIT À LA MODÉLISATION DES ÉCOSYSTÈMES

Coordinateurs : Damien Eveillard et Xavier Raynaud
Contributeurs : Jérémie Bourdon, Alain Franc, Frédéric Plewniak

Le fonctionnement des écosystèmes tel que nous l'observons aujourd'hui est le produit d'interactions entre différents organismes. Ces interactions, notamment celles des microorganismes, expliquent de nombreuses étapes des cycles biogéochimiques, qui ont aussi un intérêt sociétal (recyclage des nutriments, production de matière, etc.). Si l'abondance microbienne est connue depuis longtemps, les données génomiques haut débit ont mis en évidence une grande diversité microbienne jusque là encore insoupçonnée (Roesch *et al.* 2007, Hingamp *et al.* 2013). Cependant, ces nouvelles descriptions ne sont qu'une étape vers la promesse d'élucider, au niveau moléculaire, le fonctionnement des écosystèmes microbiens. Pour y parvenir, il est important d'utiliser, dans une modélisation dédiée, toute la diversité des ressources biotechnologiques (Zengler et Palsson 2012).

Les approches métaomiques permettent d'accéder à plusieurs niveaux d'informations utiles pour la modélisation (Figure 12A). La métagénomique donne accès à la diversité des acteurs présent dans le système que l'on veut modéliser, la métatranscriptomique (gènes exprimés) et la métaprotéomique (protéines présentes), à la fonctionnalité exprimée par une communauté donnée. Ainsi, il est aujourd'hui possible d'avoir une **perception complète de la biogéographie des microorganismes** à l'échelle d'un territoire (Ranjard *et al.* 2013), voir hollistique d'un écosystème (Karsenti *et al.* 2011). Cependant, cette perception reste vaine si l'effort n'est pas fait d'intégrer différentes données hétérogènes à disposition au sein d'un protocole dédié de modélisation, et ce pour raisonner sur la fonction même de l'écosystème. Ces différentes approches, complémentaires, ouvrent donc la voie vers une **nouvelle modélisation des écosystèmes**, appelé « systems ecology » (Klitgord et Segrè 2011), qui permet 1) d'améliorer la description des compartiments biologiques au sein des écosystèmes (Focus 12-1), tant en termes

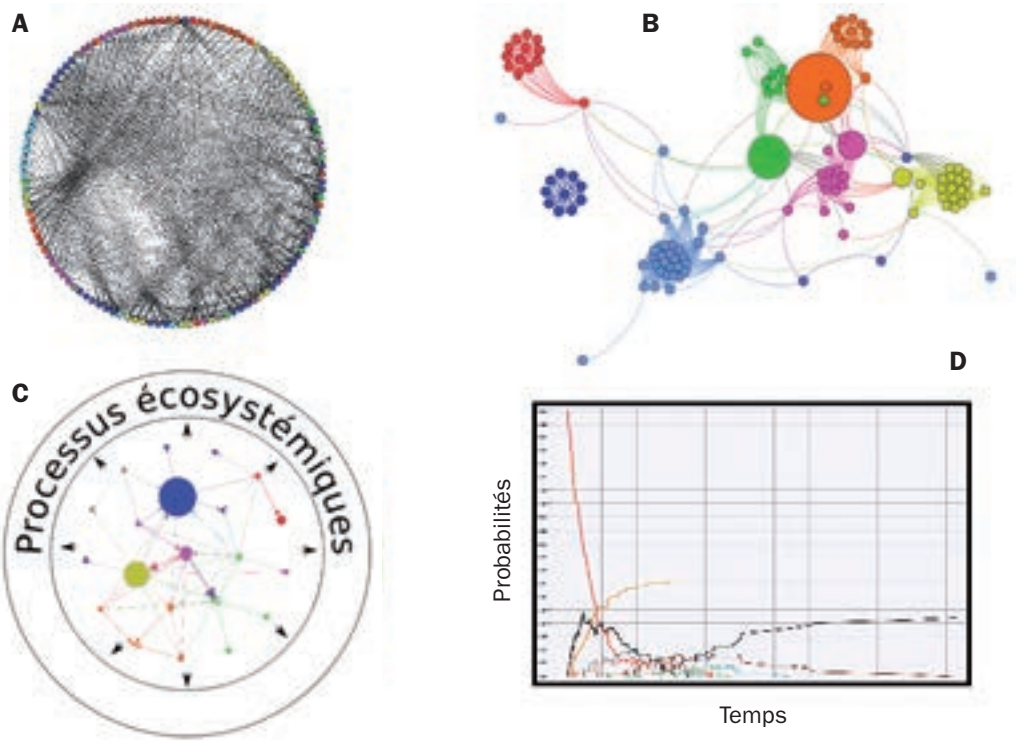


Figure 12A. Différentes étapes du processus de modélisation à partir de données haut-débit. A. Identification des corrélations entre communautés, B. Identification des communautés, C. Identification des interactions entre communautés pour D. modéliser les probabilités d'activation de communautés en fonction des inter-dépendances. Chaque point représente un domaine de recherche existant, mais la cohésion entre ces quatre points représente le principal défi à la modélisation des systèmes microbiens environnementaux.

de diversité que de fonctionnalité, 2) et d'accéder à des niveaux de complexités supérieurs. En « systems biology », le rôle de l'informatique est une fois de plus essentiel et ce, bien au-delà de ce qu'elle peut traditionnellement fournir (puissance de calcul, capacité de stockage), notamment par ses capacités à raisonner pour extraire des propriétés émergentes.

Si les données métaomiques peuvent servir à complexifier légèrement les modèles existants, la généralisation de nouveaux modèles génomiques s'appuiera sur la **structuration de l'écologie des systèmes** en tant que domaine scientifique. Elle sera, comme pour la biologie des systèmes, issue de l'informatique, des mathématiques et de l'écologie. Cette structuration reposera sur des données métagénomiques, mais également sur **des écosystèmes de référence** pouvant être utilisés comme base de tests pour de nouvelles techniques.

L'écologie des systèmes se focalisera sur deux approches complémentaires :
- Une approche bottom-up dont l'objectif sera de donner une description fonctionnelle à des données de métagénomique (par exemple, mise en évidence d'un métamétabolisme de l'écosystème – celui-ci devant converger vers les modèles de cycles biogéochimiques existants, Focus 12-2).
- Une approche top-down dont l'objectif sera de permettre une recherche ciblée dans les bases de données de génomique afin de valider les modèles existants et raisonner sur le fonctionnement des écosystèmes.

Les modèles fonctionnels résultants seront de grande complexité, décrivant à la fois les aspects diversité des communautés et de transferts de matière et d'énergie en leur sein. Cette modélisation appliquée aux écosystèmes d'intérêt sociétal permettra **l'émergence de l'écologie synthétique** pour définir les conditions optimales au pilotage des communautés comme pour promouvoir la dénitrification naturelle des sols, produire du biocarburant à partir d'écosystèmes contrôlés, etc.

FOCUS 12-1

Metabarcoding : utilisation du calcul intensif pour l'inventaire de communautés microbiennes

L'objectif est de développer un outil faisant appel au calcul intensif pour construire des inventaires automatisés de communautés microbiennes de diatomées à partir de données de séquençages d'amplicons issus de NGS. Les diatomées sont des algues unicellulaires (Figure 12B), possédant un squelette siliceux, vivant isolées ou en biofilms, en milieu marin ou dulçaquicole, à reproduction sexuée ou clonale. Les communautés de diatomées sont un indicateur de qualité du milieu où elles vivent. Il en existe environ 20 000 espèces. Actuellement, ces inventaires basés sur une identification microscopique selon leur frustule (squelette siliceux) sont très fastidieux. Les communautés de diatomées sont un excellent modèle biologique car un inventaire de diversité préalable à toute question d'écologie microbienne peut se réaliser tant de façon naturaliste et classique, que sur une base moléculaire comme dans le cadre du barcoding. Le caractère microbien de ces organismes permet également la mise en œuvre d'outils nouveaux (NGS sur amplicons) pour traiter des jeux conséquents de données. Cette démarche requiert de s'intéresser à la fois aux algorithmes, aux limites de mémoire et de temps dans leur implémentation.

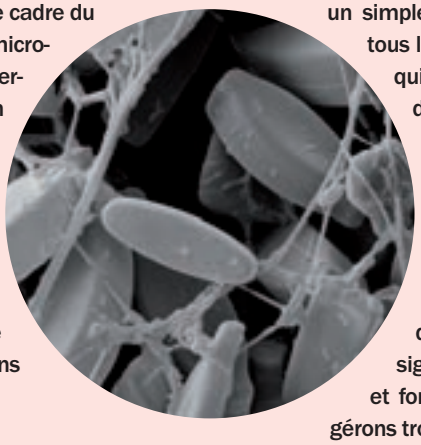


Figure 12B. Biofilm sur fond rocheux en rivière : communauté de diatomées, champignons et bactéries.

La stratégie classique est de produire des reads* de marqueurs d'intérêt taxonomique pour les diatomées (rbcL, 18S, cox1), et de les traiter par une classification, soit supervisée par un outil de type BLAST pour les comparer à une séquence complète dans une base de référence, soit non supervisée par une batterie de traitements de classification (Scloss et Westcott 2011). La base de référence est disponible dans le cadre du réseau RSYST (http://www.rsyst.inra.fr/). Dans une étude préalable, la pertinence du BLAST sur données NGS a été testée en analysant les résultats de pyroséquençage sur des communautés artificielles de composition connue, construites à partir de souches cultivées.

Ces travaux ont montré que, si le BLAST permet de reconnaitre les souches présentes, il produit également des faux positifs, à savoir la reconnaissance de souches qui ne sont pas présentes dans la communauté. Aussi, des algorithmes exacts d'alignement ont été développés, en revenant sur la base des algorithmes d'alignement global (Needleman et Wusnch) et local (Smith et Waterman).

L'utilisation des algorithmes exacts sans heuristiques d'accélération est plus gourmande en temps et mémoire, mais produit des résultats de meilleure qualité. Si l'alignement exact peut être réalisé en un temps raisonnable sur un simple ordinateur, le calcul de tous les alignements locaux requiert de fortes puissances de calcul. Comme cette question est massivement distribuable, la grille de calcul est un outil adéquat pour réaliser cette tâche. Utiliser la grille demande cependant un investissement significatif en apprentissage et formation. Aussi, nous suggérons trois directions de poursuite des recherches : 1) développer en classification supervisée des algorithmes et outils d'inventaires taxonomiques à partir d'alignements locaux exhaustifs, 2) développer les outils de classification non supervisée avec parallélisation et 3) faciliter l'accès à la grille de calcul pour la communauté des biologistes. Ces travaux qui pourront être appliqués à d'autres approches du même type requièrent une collaboration entre biologistes, mathématiciens appliqués et informaticiens autour des outils de calcul intensif.

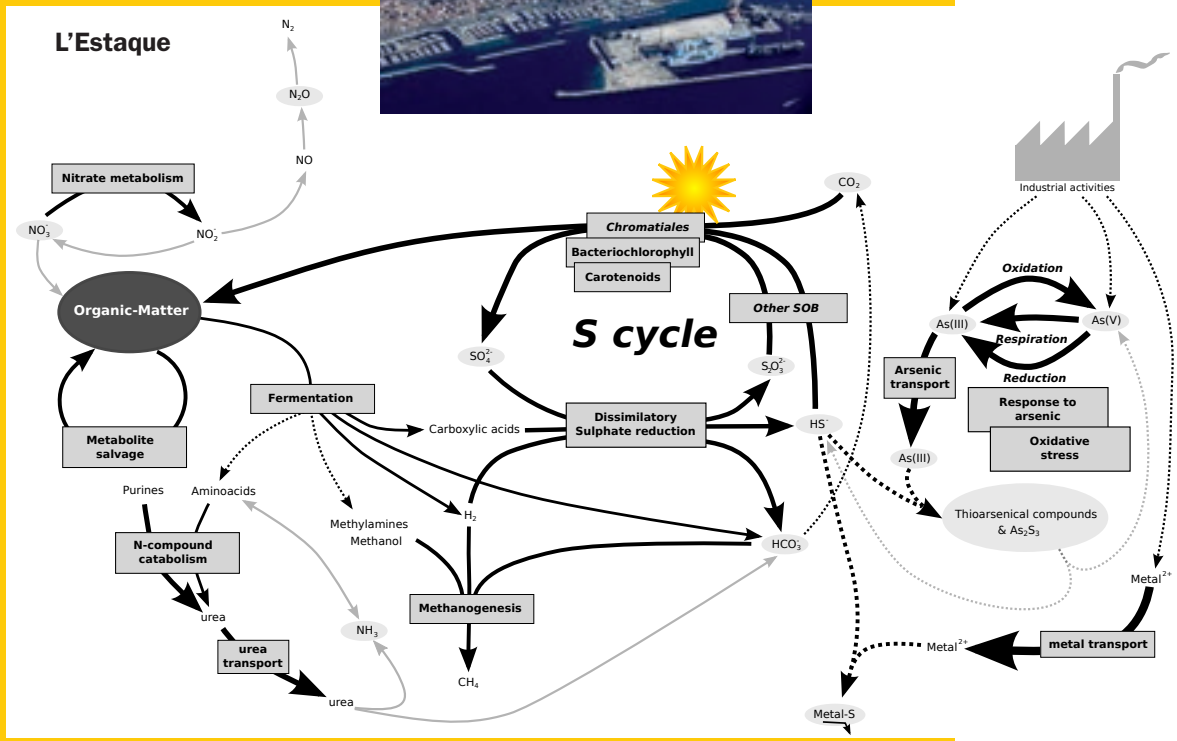
FOCUS 12-2

Modèle descriptif des cycles biogéochimiques à l'œuvre dans des sédiments marins pollués en Méditerranée

L'arsenic, responsable d'une sévère pollution dans les régions industrielles et post-industrielles présente un risque majeur de santé. Il s'agit de mettre en évidence par analyse métagénomique les facteurs biotiques susceptibles d'avoir une influence sur le cycle de l'arsenic dans des sédiments marins pollués par ce métalloïde.

Concernant la méthodologie, les métagénomes des sédiments du port de l'Estaque près de Marseille, un ancien site métallurgique pollué par l'arsenic et des métaux lourds, et de Saint-Mandrier, près de Toulon, présentant une pollution importante par les métaux mais d'une concentration en arsenic beaucoup plus faible ont été séquencés (454/Roche GS FLX Titanium).

Les séquences de ces métagénomes ont été annotées par le système RAMMCAP disponible sur le portail CAMERA (Sun *et al.* 2011) en même temps que quatre autres métagénomes contrôles. Le rôle de ces contrôles a été de mettre en évidence des processus communs aux deux sites pollués qu'une comparaison directe aurait masquée. L'enrichissement des annotations GO (Gene Ontology) dans les deux métagénomes étudiés et les informations disponibles sur les paramètres géochimiques (Mamindy-Pajany *et al.* 2013) ont permis de construire un modèle descriptif du fonctionnement des deux communautés.



FOCUS 12-2 (Suite)

Les résultats montrent qu'à l'Estaque, le cycle du soufre est central au modèle et couplé à la fermentation et au cycle de l'arsenic (Figure 12C). La fermentation fournit aux bactéries sulfato-réductrices, les donneurs d'électrons nécessaires à la réduction du sulfate. Les sulfures produits réagiraient avec l'arsenic, formant des composés thioarseniés observés expérimentalement et précipitant avec les métaux. Les complexes thioarseniés étant très solubles, ils favorisent donc sur ce site la dispersion de l'arsenic vers la colonne d'eau.

Ce projet souligne qu'au delà des résultats biologiques il est envisageable d'étudier des métagénomes complexes même avec une couverture faible et sans assemblage, à condition de se placer au niveau fonctionnel pertinent (processus biologiques). Dans ce cas, le métagénome est considéré comme un supra-organisme dont on étudie le métabolisme global. Une comparaison directe ne donnant qu'une image partielle, l'utilisation de contrôles adéquats est une nécessité pour la comparaison de métagénomes proches. Idéalement, les contrôles ne devraient contenir que les processus ubiquitaires pour ne pas masquer des processus intéressants. Il manque ici des standards, sorte de moyenne sur les grandes catégories d'environnement, permettant de se focaliser sur les données pertinentes. Il serait également intéressant de passer du modèle descriptif à un modèle plus rigoureux, qui permettrait d'étudier plus précisément les relations entre les différents cycles et les effets des différents processus sur la dispersion de l'arsenic.

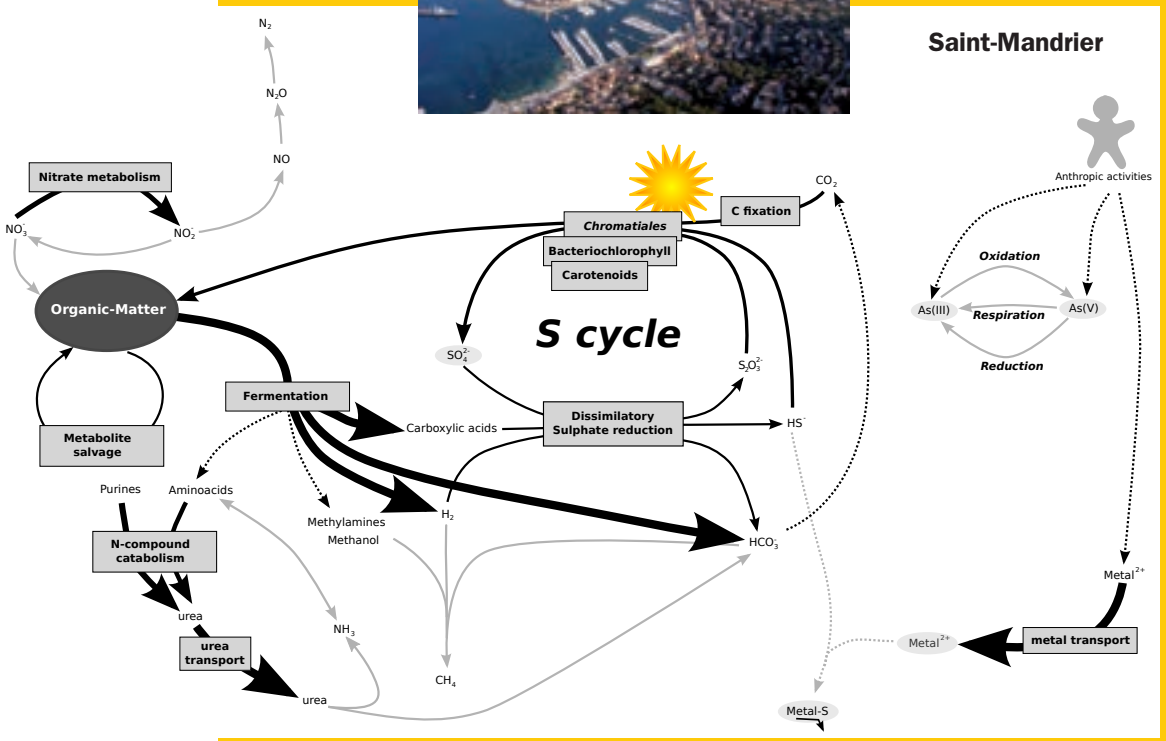


Figure 12C. Comparaison des cycles biogéochimiques par analyse métagénomique de deux sites méditerranéens pollués par les métaux lourds. A gauche, ancien site métallurgique de l'Estaque, à droite, station de Saint-Mandrier.

FOCUS 12-3

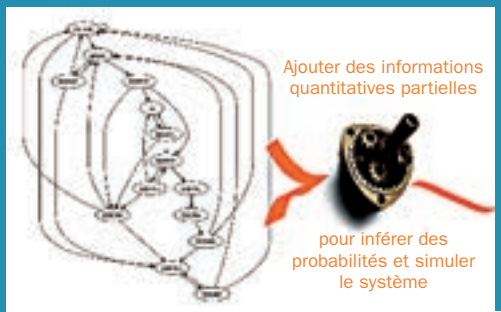
Modélisation probabiliste des interactions au sein des communautés microbiennes

Si l'intégration des données de métagénomique permet la mise en place de graphes de causalités qui décrivent qualitativement les interactions entre communautés microbiennes, elle ne permet pas de donner une dimension fonctionnelle aux écosystèmes étudiés au niveau moléculaire. Pour pallier à ce manque, il faut développer des techniques de modélisation qui intègrent d'autres types d'informations quantitatives (ou fonctionnelles) issus des expériences physicochimiques, et ce même si ces informations quantitatives restent partielles par rapport aux données métagénomiques.

Récemment, Bourdon *et al.* (2011) ont montré qu'en introduisant des aspects probabilistes dans une modélisation booléenne, il était possible d'intégrer de manière efficace les aspects qualitatifs et quantitatifs d'un système vivant. Cette modélisation ETG (Event Transition Graph) est ici appliquée à un petit réseau microbien représentant le cycle biogéochimique de l'azote afin de quantifier l'influence d'une communauté microbienne sur la fonction globale de l'écosystème. La modélisation considère 1) une description d'un événement biologique et ses interactions : ici l'occurrence d'une réaction biogéochimique qui produit un composé, connecté à d'autres réactions ayant le composé comme substrat, et 2) une information quantitative comme la variation de concentration d'un composé au cours du temps. Par un procédé d'optimisation, la modèlisation ETG consiste à identifier un jeu de probabilités associé aux interactions d'événements, de manière à ce que le graphe ainsi pondéré, puisse reproduire « en moyenne » le comportement quantitatif proposé en (2).

Les résultats montrent que le cycle biogéochimique de l'azote est principalement contrôlé par des communautés bactériennes. Le réseau correspondant est une succession de réactions biogéochimiques induites par différents microorganismes. Nous considérons par cette modélisation l'écosystème comme un système métamétabolique qui synthétise les communautés bactériennes par les réactions qu'elles contrôlent. Le graphe qualitatif correspond à 14 événements ou réactions et 32 interactions (Figure 12D). Chaque événement induit un effet quantitatif : chaque fois que l'événement est emprunté en parcourant le graphe, cela induit une augmentation active de la population associée à l'événement (i.e. augmentation de la population ammonia-oxydante quand la réaction amo est empruntée), une dégradation passive sinon. Pour estimer les probabilités de l'ETG, nous utilisons ensuite les variations de quantités d'ammoniac et de nitrates entre 2001 et 2004 dans la baie de Chesapeake (USA) (Bouskill *et al.* 2011). Les probabilités permettent de reproduire les comportements utilisés en apprentissage, mais aussi de reproduire les variations de nitrites alors que cette concentration n'était jusqu'alors pas considérée. Une fois fixées, les probabilités permettent de simuler le comportement du modèle probabiliste avec un algorithme dédié (Figure 12D), mais aussi de raisonner sur le système afin d'identifier l'importance relative d'un événement par rapport aux autres pour "contrôler" le comportement quantitatif de l'écosystème microbien.

La modélisation ETG se focalise sur le comportement quantitatif dynamique. Cet aspect de modélisation est motivé par le fait que les mesures quantitatives sont rarement obtenues à l'équilibre, et que les mesures fonctionnelles s'intéressent principalement aux communautés sous pressions adaptatives. Une extension de l'ETG est actuellement en préparation pour tenir compte de la diversité comme mesure quantitative d'apprentissage.



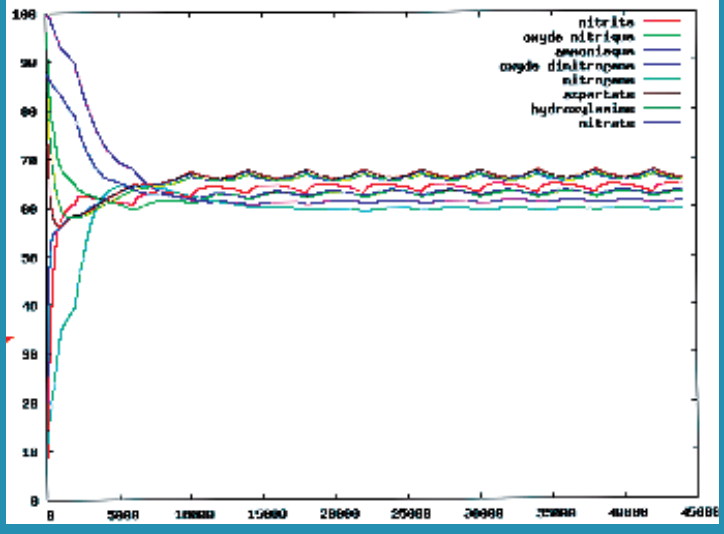


Figure 12D. Protocole de modélisation probabiliste. À partir d'un graphe d'événements (cycle biogéochimique de l'azote), via l'inférence de probabilités, il est possible de simuler la dynamique du système modélisé, comme ici les variations des métabolites d'intérêt au cours du temps.

S'il est cependant possible d'accéder à des données métaomiques et fonctionnelles (e.g. mesures de flux) pour des échelles spatiales similaires, ces échelles restent encore très supérieures aux échelles d'interactions microbiennes, souvent moléculaires. Il est aussi difficile de mettre en relation les données de diversité et les données fonctionnelles. Par ailleurs, les données existantes sont encore hétérogènes, mal normalisées, et peu accessibles à la communauté des modélisateurs.

L'obtention de modèles prédictifs (Focus 12-3) est conditionnée par l'intégration de dimensions spatiales et temporelles de la diversité des communautés de microorganismes et de leurs fonctionnements. Pour répondre à ce défi, il faut renforcer les stratégies d'échantillonnage métagénomique et les exploiter à l'échelle du microenvironnement. En effet, les connaissances de la structure spatiale (ex : celle des pores du sol ou celle des fluides autour de particules) et de la structure temporelle de l'écosystème microbien restent très rudimentaires comparées à la masse de données métaomiques accessibles. Par ailleurs, les analyses actuelles en génomique environnementale se focalisent sur la corrélation d'abondance de communautés issues de différents écosystèmes, alors que seules les informations de causalité permettent le pont vers les aspects fonctionnels des écosystèmes. Enfin, un dernier défi, plus difficile, consiste à modifier le mode de partage des données afin de faciliter leur mise à disposition via des dépôts centralisés.

Pour développer la **structuration de l'écologie des systèmes**, il est indispensable de renforcer les liens entre les microbiologistes, les modélisateurs et les bioinformaticiens. Seule cette interdisciplinarité permettra de modéliser la complexité des problèmes actuels. Cela implique de renforcer les formations universitaires notamment en proposant des spécialisations aux formations existantes de bioinformatique et de modélisation. Cela implique également de décloisonner les laboratoires en accentuant des initiatives inter-organismes de recherche.

Afin de promouvoir **l'interdisciplinarité et la standardisation des techniques de modélisation**, il faut identifier des **sites ou des écosystèmes de référence**. Ainsi, l'ensemble des données disponibles (génomiques, fonctionnelles, spatiales, temporelles) sera à disposition de manière ouverte pour stimuler la production de solutions informatiques. Par ailleurs, les stratégies d'échantillonnage devront être concertées afin de maximiser l'adéquation entre génomique et connaissances fonctionnelles et faire ainsi émerger des liens de causalité entre communautés. L'intérêt sera de susciter un **engouement international autour de problèmes ciblés**. Dès l'ouverture des données, les diverses plateformes bioinformatiques et centres de calculs pourront se fédérer pour appuyer une nécessaire démarche collaborative (notamment via la mise en oeuvre de workflow de type galaxy) à l'analyse des datavalanches de génomique environnementale.

RÉFÉRENCES

Bourdon J, Eveillard D, Siegel A. 2011. Integrating quantitative knowledge into a qualitative gene regulatory network. PloS Comput Biol 7:e1002157.

Bouskill N, Eveillard D, O'Mullan G, Jackson G, Ward B. 2011. Seasonal and annual reoccurrence in betaproteobacterial ammonia-oxidizing bacterial population structure. Environ Microbiol 13:872–886.

Hingamp P *et al.* 2013. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J 7:1678-1965.

Karsenti E *et al.* The Tara Oceans Consortium. 2011. A holistic approach to marine eco-systems biology. PloS Biol 9: e1001177.

Klitgord N, Segrè D. 2011. Ecosystems biology of microbial metabolism. Curr Opin Biotechnol 22:541–546.

Mamindy-Pajany Y, Hurel C, Géret F, Galgani F, Battaglia-Brunet F, Marmier N, Roméo M. 2013. Arsenic in marine sediments from French Mediterranean ports: geochemical partitioning, bioavailability and ecotoxicology. Chemosphere 90:2730-2736.

Roesch L, Fulthorpe R, Riva A, Casella G, Hadwin A, Kent A, Daroub S, Camargo F, Farmerie W, Triplett E. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J 1:283–290.

Ranjard L, Dequiedt S, Chemidlin Prévost-Bouré N, Thioulouse J, Saby NPA, Lelievre M, Maron P-A, Morin F, Bispo A, Jolivet C, Arrouays D, Lemanceau P. 2013. Turnover of soil bacterial diversity driven by wide-scale environmental heterogeneity. Nature commun 4:1434-1444.

Schloss PD, Westcott SL. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol 77: 219-3226.

Sun S *et al.* 2011. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Res 39:D546–D551.

Zengler K, Palsson BØ. 2012. A road map for the development of community systems (CoSy) biology. Nat Rev Microbiol 10:366–372.

XIII BILAN ET RECOMMANDATIONS

Coordinateurs : Dominique Joly et Denis Faure

La diversité des questions scientifiques et des approches méthodologiques abordées dans les différents chapitres de ce document de prospective en génomique environnementale témoigne du dynamisme et de l'engagement de la communauté française pour relever le défi de porter l'écologie globale dans un champ intégré des sciences de l'environnement. L'apport des outils omiques engendre à ce titre un renouvellement des pratiques et des questionnements dans lesquels la communauté française s'est résolument engagée. Son savoir-faire dans ce domaine plaide pour une systématisation des approches pluridisciplinaires des objets et des questions relevant de la génomique environnementale, pour un meilleur partage de la collecte et de la valorisation des données et pour une implication dans les initiatives internationales.

Au cours des différentes actions menées par le RTP-GE, un certain nombre de recommandations ont régulièrement émergé, certaines clairement reprises ici au fil du document, d'autres mentionnées dans le cahier de prospective d'Avignon de l'INEE en 2012. L'objectif de ce chapitre n'est pas de faire un inventaire de l'ensemble de ces réflexions mais plutôt d'en faire une synthèse, aussi fidèle que possible, afin d'identifier les **priorités d'actions** que ce soit au niveau des chercheurs, des laboratoires ou des organismes de recherche.

Le devenir de la génomique environnementale s'appuie sur quatre axes fondamentaux :

1) le premier axe concerne les **objets de recherche**. A ce jour, une masse importante de données existantes concerne des organismes modèles (certains mammifères, oiseaux, poissons, plantes, bactéries ou champignons) et nous sommes encore bien loin d'avoir une image exhaustive de la diversité biologique, qu'elle soit ordinaire, remarquable, rare ou cryptique. Il s'agit donc **d'accroître notre connaissance du**

vivant en s'attaquant à des phylums peu étudiés tant pour leur biodiversité que pour leur distribution, leurs rôles dans les écosystèmes, ainsi que leurs comportements et interactions avec les autres organismes qu'ils soient macroscopiques (ex : insectes, crustacés, champignons) ou microscopiques, qu'ils soient autonomes, parasitaires ou symbiotiques (virus, protozoaires, bactéries ou helminthes). Un investissement important reste donc à faire pour **coordonner et diversifier la collecte des données** sur l'arbre du vivant tout en associant aux données taxinomiques des informations biologiques, environnementales et sociétales permettant d'**intégrer les relations organismes-hommes-environnements**.

2) Le second axe est lié aux **défis techniques et aux enjeux scientifiques et sociétaux** associés à la **qualité des informations** fournies par les approches omiques. Celles-ci génèrent en effet un flux de données dont le volume de production se fait au détriment de sa qualité et sa fiabilité avec des risques importants de mauvaises interprétations scientifiques et sociétales. Il est urgent et nécessaire d'améliorer la qualité de l'information aussi bien au niveau des **outils de production des données omiques** (techniques et méthodologiques, notamment pour les analyses de communautés complexes), que de la **consolidation des différentes approches omiques** entre elles (ex : annotations automatiques, interfaçage). Il s'agit également de renforcer notre démarche explicative avec d'autres approches expérimentales ou prédictives (ex : réseaux de relations trophiques, écoinformatique, socioécologie). Enfin, augmenter la **synergie entre communautés scientifiques** omiques et non-omiques est indispensable pour être à même de mieux comprendre comment la « data-avalanche » de données bouleverse la vision du vivant et des populations humaines par les sociétés elles-mêmes (ex : commercialisation d'outils d'identification des origines ethniques des populations humaines ; quantification des impacts anthropiques sur la biodiversité).

3) Le troisième axe concerne les **outils d'analyse, de stockage et d'interfaçage** des données omiques par et pour les différentes communautés scientifiques. Le haut débit génère du **volume et de l'hétérogénéité dans les données** (ADN/ARN, protéines, chimie du vivant, géochimie, physico-chimie) et dans les sources (populations, communautés, écosystèmes). Au-delà des questions relatives à leur **manipulation, leur stockage, et leur partage** (espaces d'archivage et capacités de calculs), se pose la question première de leur exploitation qui est à ce jour très partielle. Cette exploitation se doit d'aller au delà des chercheurs ayant produit les données lesquelles doivent (pour des questions de coût) pouvoir être exploitées par d'autres chercheurs. Un enjeu majeur est donc le **développement de méthodes compatibles de partage des données.**

4) Le quatrième axe concerne la promotion de **l'interdisciplinarité et la standardisation des techniques de modélisation**. A cette fin, il faut identifier des sites ou des écosystèmes de référence pour lesquels l'ensemble des données (génomiques, fonctionnelles, spatiales, temporelles) sera à disposition de manière ouverte pour stimuler le développement de modèles intégrés.

Au niveau institutionnel, on retiendra deux réflexions sur lesquelles une priorité nationale devrait être engagée par les différents organismes. La première concerne la demande de **formation et de développement de cadres conceptuels** liés à l'émergence des nouvelles questions de recherche portées par les omiques. L'augmentation des interfaces entre disciplines, notamment pour le développement d'outils mathématiques, bioinformatiques et statistiques nécessaires à l'exploitation de ces données, devra s'accompagner d'une **programmation de formation évolu-**

tive et diversifiée y compris pour les chercheurs confirmés. Ceci permettra d'accompagner l'évolution des outils, le seuil des NGS de troisième génération étant déjà franchi.

La deuxième réflexion concerne une **capitalisation autour d'équipements et de programmes de recherche** (via notamment l'ANR ou l'accès à France Génomique) dédiés à la génomique environnementale. En effet, les priorités actuelles pour les infrastructures nationales concernent les domaines d'intérêt économique (agroalimentaire, ressources carbonées, etc) ou de santé publique (cancers, maladies neurodégénératives, maladies rares, etc) et l'ouverture à la génomique environnementale reste modeste. Le recours aux compagnies privées de séquençage ne répond que très partiellement aux besoins d'optimisation et d'implication dans les procédures et développements particuliers et le financement des programmes de recherche reste très éclaté et limité au regard de la pression. Des actions concertées, locales, nationales et inter-organismes seraient un atout certain pour le devenir de ce champ disciplinaire.

Enfin, pour conclure sur ce document de prospective (dont les principaux mots-clés sont illustrés figure 13A), une dernière recommandation porte sur la représentation française à l'international. Différents consortia traitent de thèmes proches de ceux développés par la génomique environnementale, et notamment celui porté par l'European Science Foundation, Ecological and Evolutionary Functional Genomics (EuroEEFG, voir chap. III). La France n'y est représentée que de façon anecdotique et aurait avantage à renforcer son implication.

Figure 13A. Analyse des mots les plus fréquents issus de l'intégralité du texte de ce cahier de prospective. Noter le mot « plus » qui semble tout aussi important que le mot « séquence » : plus de données, plus d'analyses, plus de gènes, plus de génomes, plus d'espèces, plus de biodiversité, plus d'interactions…. Encore plus, tel pourrait être le mot de la fin !

L'autre initiative internationale de grande ampleur concerne le réseau des observatoires génomiques, « Genomic Observatories (GO) » (voir chap. III). La France a des atouts majeurs à promouvoir dans un tel dispositif, avec notamment les outils et instruments de l'INEE, à savoir les Zones Ateliers (ZA), et les Observatoires Hommes-Milieux (OHM).

Si un certain nombre d'actions à l'international dépendent de l'investissement des chercheurs eux-mêmes, d'autres bénéficieraient de concertations plus fortes entre différents organismes de recherche, notamment dans le cadre d'une programmation européenne et internationale.

GLOSSAIRE GÉNÉRAL

Alignement (p 58) : comparaison de séquences d'ADN placées les unes sous les autres pour faire ressortir leurs éventuelles similarités et différences.

Assemblage (p 26) : action de reconstituer un génome ou un fragment de génome à partir de ses fragments séquencés. Ceci est rendu possible par les chevauchements des fragments de séquences.

Barcode (p 8) : méthode d'identification moléculaire des organismes à partir de leur ADN quand il est en relativement bon état. Des marqueurs moléculaires de l'organisme à identifier sont séquencés et comparés à des séquences dans une banque de données.

Couverture (p 39) : évaluation du nombre de fragments recouvrant la séquence d'une même région génomique. La séquence d'un génome est obtenue à partir de nombreux fragments assemblés. Plus une même région est séquencée plusieurs fois, plus sa séquence sera connue avec précision.

Ecologie fonctionnelle (p 6) : domaine de recherche ayant pour objectif l'étude du fonctionnement des écosystèmes.

Génome de référence (p 28) : génome séquencé et assemblé qui sert de référence pour d'autres projets de séquençage d'espèces proches.

Génomique/transcriptomique (p 3) : science qui s'intéresse à l'ensemble du génome ou du transcriptome (ensemble des transcrits).

Gigabases (p 23) (et autres unités de cette mesure) : la paire de bases est l'unité du génome et de son séquençage. Ce concept est ensuite décliné en fonction de la taille du fragment considéré en kilobases (mille paires de bases), mégabases (millions de paires de bases), gigabases (milliards de paires de bases)…

Métagénomique/métatranscriptomique (p 8) : c'est la génomique/transcriptomique de matériel génétique prélevé directement dans l'environnement et donc issu de communautés d'organismes.

Métabolomique (p 65) (voies métaboliques) : science qui étudie l'ensemble des métabolites (sucres, acides gras, acides aminés…) isolés d'une cellule, d'un tissu, d'un organisme ou d'un environnement.

Microorganismes (p 10) : organismes vivants microscopiques unicellulaires ou pluricellulaires (Virus, Bactéries, Archées et certains Eucaryotes).

OTU (p 39) : Operational Taxonomic Unit. Ensemble de séquences ADN regroupées entre elles sur la base de leur identité dont le seuil est fixé arbitrairement (1 ou 5 % par exemple).

Pipeline d'analyse (p 57) : suite de programmes bioinformatiques permettant l'analyse automatique de nombreuses données.

Phylogénie (p 5) : étude des relations de parenté entre différents êtres vivants ou non en vue de comprendre leur évolution.

Protéomique (p 65) : science qui étudie l'ensemble des protéines d'une cellule, d'un tissu ou d'un organisme.

Séquençage de novo (p 31) : séquençage de génomes inconnus, n'ayant pas de référence dans les bases de données.

Taxinomie moléculaire (taxon) (p 5) : science se basant sur la génétique moléculaire pour décrire les êtres vivants et les regrouper en entités nommées taxons.

GLOSSAIRE TECHNIQUE

Amplicon (p 51) : produit d'une amplication par PCR.

Contigs (p 29) : fragments d'ADN qui se recouvrent et forment ensemble une région d'intérêt.

Illumina (Hi-seq) (p 5) : méthode de séquençage à haut débit de seconde génération (NGS) avec laquelle la lecture est directe.

Lecture (read) (p 22) : détermination de l'enchaînement des bases d'un fragment d'ADN, produit des séquençages de type NGS.

Librairie (banque) (p 48) : important ensemble d'éléments physiques ou informatiques, comme des clones ou des séquences de gènes, regroupés en un même lieu ou support (par exemple congélateur ou fichier).

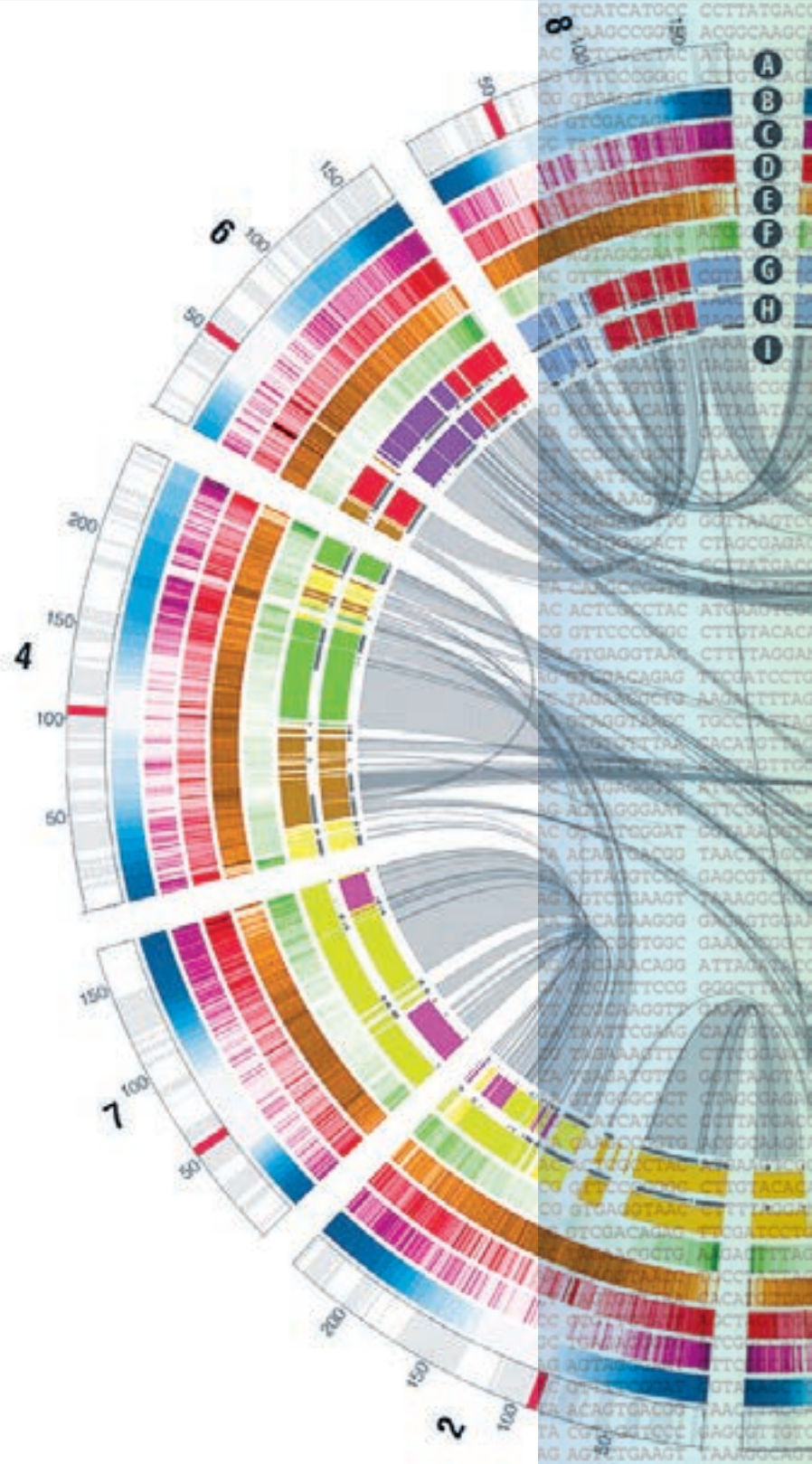
Multiplexage (p 54) : technique qui consiste à analyser plusieurs échantillons en parallèle en une seule opération de séquençage NGS ou de PCR par exemple.

Pyroséquençage 454 (p 5) : méthode de séquençage haut débit de seconde génération (NGS) dans laquelle il n'y a pas besoin de cloner le fragment d'ADN concerné. La lecture de la séquence est directe.

RAD-seq (p 39) : Restriction site Associated DNA sequencing. Séquençage partiel de génomes autour de certains sites de restriction permettant une réduction des coûts et une plus grande vitesse d'analyse comparative des génomes de grande taille.

Run (p 23) : cycle complet de fonctionnement d'un séquenceur.

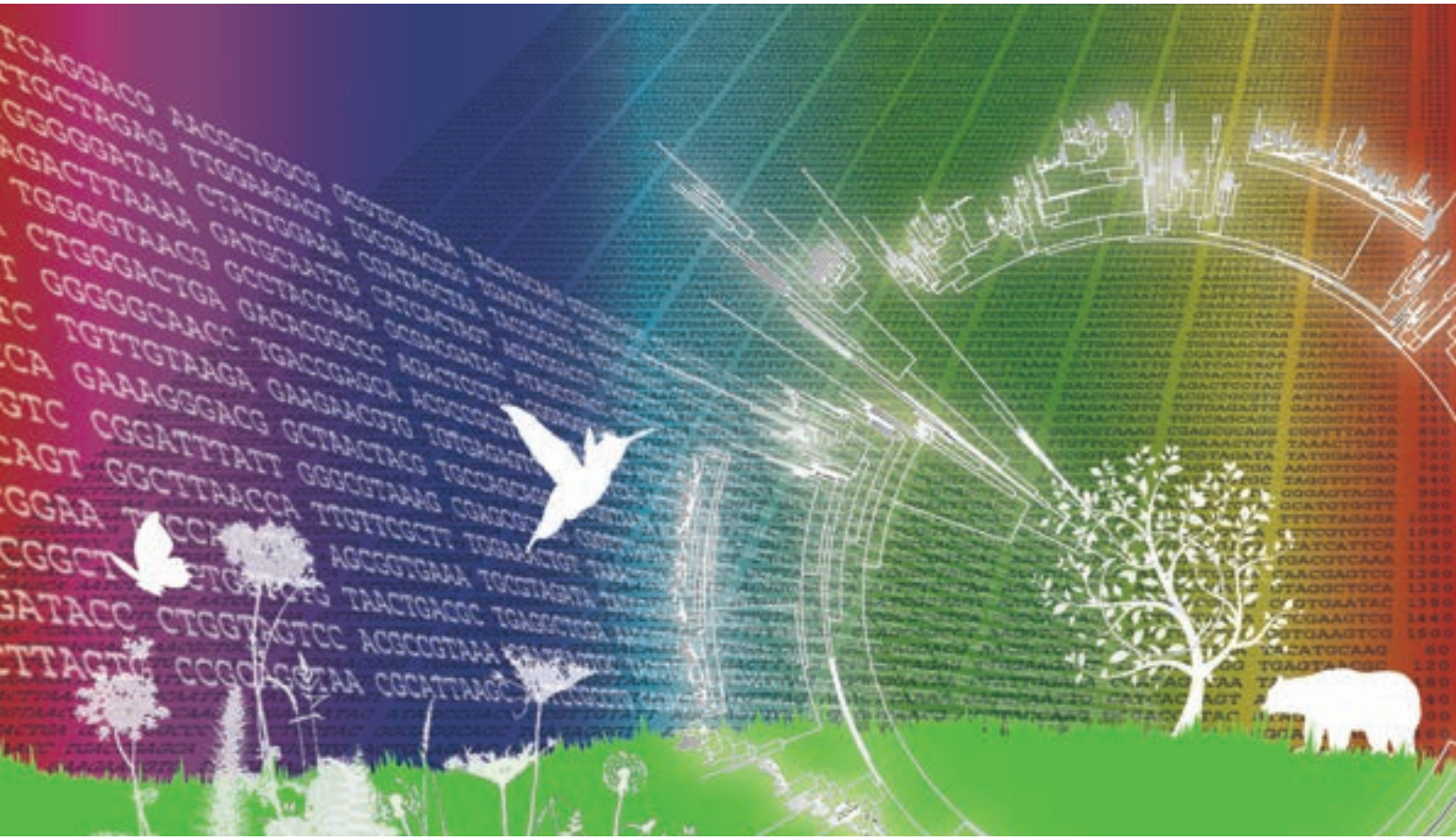
SNP (p 17) : Single Nucleotide Polymorphism. Variation d'un seul nucléotide dans les génomes ou séquences comparées.



LISTE DES AUTEURS

Coordinateurs et contributeurs des chapitres

- Julie Aubert, UMR 518 Mathématiques et informatique appliquées (AgroParisTech/INRA, Paris)
- Eric Bapteste, UMR 7138 Systématique, adaptation, évolution (CNRS/Université Paris 6/IRD/Université Antilles et Guyane/MNHN, Paris)
- Philippe Bertin, UMR 7156 Génétique moléculaire, génomique et microbiologie (CNRS/Université Strasbourg)
- Didier Bogusz, UMR 232 Diversité adaptation et développement des plantes (IRD/Université Montpellier 2)
- Jérémie Bourdon, UMR 6241 Laboratoire d'informatique de Nantes Atlantique (CNRS/Université Nantes/Ecole des Mines Nantes/INRIA, Nantes)
- Catherine Boyen, UMR7 139 Végétaux marins et biomolécules (CNRS/Université Paris 6, Roscoff)
- Vincent Breton, UMR 6533 Laboratoire de physique corpusculaire (CNRS/Université Clermont-Ferrand 2)
- Didier Debroas, UMR 6023 Microorganismes : génome et environnement (CNRS/Universités Clermont-Ferrand 1 et 2)
- Régis Debruyne, UMS 2700 Outils et méthodes de la systématique intégrative (CNRS/MNHN/Université Paris 6)
- Frédéric Delsuc, UMR 5554 Institut des sciences de l'évolution de Montpellier (CNRS/IRD/Université Montpellier 2)
- Frantz Depaulis, UMR 7625 Ecologie et évolution (CNRS/Université Paris 6/ENS, Paris)
- David Enard, UMR 7625 Ecologie et évolution (CNRS/Université Paris 6/ENS, Paris)
- François Enault, UMR 6023 Microorganismes : génome et environnement (CNRS/Universités Clermont-Ferrand 1 et 2)
- Damien Eveillard, UMR 6241 Laboratoire d'informatique de Nantes Atlantique (CNRS/Université Nantes/Ecole des Mines Nantes/INRIA, Nantes)
- Denis Faure, UPR 2355 Institut des sciences du végétal (CNRS, Gif-sur-Yvette)
- Alain Franc, UMR 1202 Biodiversité, gènes et communautés (INRA/Université Bordeaux 1)
- Laurence Garczarek, UMR 7144 Adaptation et diversité en milieu marin (CNRS/Université Paris 6, Roscoff)
- Catherine Hänni, UMR 5600 Environnement, ville, société (CNRS/Universités Lyon 2 et 3/Université St-Etienne/INSA Lyon/ENS Lyon/ Ecole des Mines St-Etienne/Ecole nationale des travaux publics d'état, Lyon)
- Dominique Joly, UPR 9034 Evolution, génomes et spéciation (CNRS, Gif-sur-Yvette)
- Mathieu Joron, UMR 7205 Origine, structure et évolution de la biodiversité (CNRS/MNHN, Paris)
- Annegret Kohler, UMR 1136 Interactions arbres-microorganismes (INRA/Université Lorraine, Nancy)
- Sébastien Lavergne, UMR 5553 Laboratoire d'écologie alpine (CNRS/Université de Savoie/Université Grenoble 1)
- Line Le Gall, UMR 7138 Systématique, adaptation, évolution (CNRS/Université Paris 6/IRD/Université Antilles et Guyane/MNHN, Paris)
- Denis Le Paslier, UMR 8030 Génomique métabolique (CNRS/Université d'Evry-Val-d'Essonne/CEA Paris, Evry)
- Guillaume Lecointre, UMR 7138 Systématique, adaptation, évolution (CNRS/Université Paris 6/IRD/Université Antilles et Guyane/MNHN, Paris)
- Roland Marmeisse, UMR 5557 Ecologie microbienne (CNRS/Université Lyon 1/INRA, Lyon)
- Francis Martin, UMR 1136 Interactions arbres-microorganismes (INRA/Université Lorraine, Nancy)
- Tiphaine Martin (CNRS/University of Cambridge, UK)
- Sylvain Merlot, UPR 2355 Institut des sciences du végétal (CNRS, Gif-sur-Yvette)
- Sébastien Monchy, UMR 8187 Laboratoire d'océanologie et de géosciences (CNRS/Université Lille 1/Université Littoral, Wimereux)
- Xavier Nesme, UMR 5557 Ecologie microbienne (CNRS/INRA/Université Lyon 1)
- Philippe Normand, UMR 5557 Ecologie microbienne (CNRS/INRA/Université Lyon 1)
- Morgane Ollivier-Ruz (ENS Lyon/Université Lyon 1)
- Eric Pante, UMR 7266 Littoral, environnement et sociétés (CNRS/Université la Rochelle)
- Frédéric Partensky, UMR 7144 Adaptation et diversité en milieu marin (CNRS/Université Paris 6, Roscoff)
- Eric Pelletier, UMR 8030 Génomique métabolique (CNRS/Université d'Evry-Val-d'Essonne/CEA Paris, Evry)
- Guy Perrière, UMR 5558 Biométrie et biologie évolutive (CNRS/INRIA/Hospices civils Lyon/Université Lyon 1)
- Pierre Peyret, UMR 6023 Laboratoire Microorganismes, génomes et environnement (CNRS/Université Clermont-Ferrand 1)
- Eric Peyretaillade, UMR 6023 Laboratoire Microorganismes, génomes et environnement (CNRS/Université Clermont-Ferrand 1)
- Frédéric Plewniak, UMR 7156 Génétique moléculaire, génomique et microbiologie (CNRS/Université Strasbourg)
- François Pompanon, UMR 5553 Laboratoire d'écologie alpine (CNRS/ Université de Savoie/Université Grenoble 1)
- Nicolas Puillandre, UMR 7138 Systématique, adaptation, évolution (CNRS/Université Paris 6/IRD/Université Antilles et Guyane/MNHN, Paris)
- Jean-Yves Rasplus, UMR Centre de biologie et de gestion des populations (INRA/IRD/CIRAD, Montpellier)
- Xavier Raynaud, UMR 7618 Biogéochimie et écologie des milieux continentaux (CNRS/Université Paris 6/ENS Paris/IRD/Université Paris Est Créteil Val-de-Marne, Paris)
- Sarah Samadi, UMR 7138 Systématique, adaptation, évolution (CNRS/Université Paris 6/IRD/Université Antilles et Guyane/MNHN, Paris)
- Jean-François Silvain, UR 072 Biodiversité et évolution des complexes plantes-insectes ravageurs-antagonistes (CNRS/IRD, Gif-sur-Yvette)
- Télesphore Sime-Ngando, UMR 6023 Microorganismes : génome et environnement (CNRS/Universités Clermont-Ferrand 1 et 2)
- Pascal Simonet, UMR 5005 Laboratoire Ampère (CNRS/Ecole centrale Lyon/INSA Lyon/Université Lyon 1)
- Carole Smadja, UMR 5554 Institut des sciences de l'évolution de Montpellier (CNRS/IRD/Université Montpellier 2)
- Pierre Taberlet, UMR 5553 Laboratoire d'écologie alpine (CNRS/ Université de Savoie/Université Grenoble 1)
- Aurélie Tasiemski, UMR 8198 Génétique et évolution des populations végétales (CNRS/Université Lille 1)
- Xavier Vekemans, UMR 8198 Génétique et évolution des populations végétales (CNRS/Université Lille 1)



Crédit photos/Illustrations : © CNRS Photothèque : Serge AUBERT - Jérôme FOURNIER - Marion VALEIX - CNRS © Cyril Fresillon - © Serge Ibanez - © Philippe Normand - © Nicolas Puillandre - Martine Chomard - Christophe Destombe - Dominique Joly - Yasmin Latour - Line Le Gall - Pierre Peyret - José Utge - Colomban de Vargas - Shipher Wu - © AFP / Martin BUREAU - © BIO Photography Group, Biodiversity Institute of Ontario - Fotolia - Wikicommons