

Janvier 2011 – Guillaume Fernandez

Mise à jour : Octobre 2011

La métagénomique

Mise en perspective face aux enseignements historiques de la biologie évolutive

Moyens techniques en 2011

Génomique microbienne à grande échelle



Table des matières

1 . L'importance nouvelle des projets de métagénomique.....	3
2 . Contexte scientifique et objectifs de la métagénomique.....	3
2.1. Contexte et positionnement dans l'histoire de la biologie évolutive.....	3
Un projet dans la lignée d'une longue histoire : celle de la découverte d'une biodiversité étendue.....	3
Un renouveau de la perception des modes de transmission de l'information du monde vivant.....	4
2.2. Etat de l'art.....	6
Précédents dans les projets de métagénomique.....	7
2.3. Objectifs scientifiques prévisionnels des projets à venir.....	8
Vue générale.....	8
Analyse des gènes et des éléments mobiles.....	8
3 . Aspects techniques :.....	10
3.1. Séquenceurs et bioinformatique.....	10
1/ La reconstruction des séquences (volet technique).....	10
2/ La bio-analyse : l'exploitation des données.....	10
3.2. Description des tâches.....	11
Le prélèvement d'échantillons.....	11
Le séquençage (méthode et choix de l'équipement).....	11
L'assemblage (éventuel) et la bioanalyse.....	12
4 . Estimation financière d'un projet de trois ans.....	13
Postes budgétaires.....	13

1 . L'importance nouvelle des projets de métagénomique

Un projet de recherche fondamentale de métagénomique a pour but d'augmenter les connaissances sur la biodiversité microbienne. Dans la mesure du possible les chercheurs ont pour mission d'affiner la description des mécanismes non mendéliens de transmission de l'information génétique en s'appuyant sur les données de métagénomes microbiens prélevés dans l'environnement. En effet notre vision de l'évolution des génomes a radicalement changé grâce à l'analyse des génomes procaryotes. A coté de la transmission verticale des gènes, il a été démontré que les génomes microbiens sont massivement affectés par des transferts latéraux de gènes. Cependant, nous n'avons pas d'information pour la majorité des millions d'espèces procaryotes censées exister, principalement à cause de l'incapacité à les cultiver au laboratoire. Dans les années à venir, le nombre de projets visant donc à recueillir des données dans des régions océaniques jusque là peu échantillonnées devraient augmenter. En France, les organismes de recherche publics (le CNRS par exemple) auront sans doute vocation à s'associer en collaboration avec les acteurs de la mer possédant le matériel technique capable d'effectuer les prélèvements, notamment l'Institut Français de la Mer (Ifremer).

2 . Contexte scientifique et objectifs de la métagénomique

2.1. Contexte et positionnement dans l'histoire de la biologie évolutive

Un projet dans la lignée d'une longue histoire : celle de la découverte d'une biodiversité étendue

C'est en 1663 qu'a lieu la première observation d'une cellule bactérienne, par Antonie van Leeuwenhoek, l'inventeur du microscope. 200 ans plus tard, le botaniste Ferdinand Cohen et le mycologue Franz Unger développent le concept de culture. La culture de micro-organismes devient la technique standard pour l'étude des causes des maladies par des scientifiques tels que Robert Koch. Les bactéries cultivées sont une large source de découverte et aboutissent à une explosion des connaissances en physiologie et génétique microbienne de 1960 à la moitié des années 1980. La conscience de l'existence d'un monde microbiologique non cultivé émerge ensuite. En 1985 Staley et Konopka dénoncent la disproportion entre la taille estimée des populations cultivées sur plaques de dilution et les comptages au microscopique réalisés sur échantillons environnementaux (« the great plate count anomaly »). En 1988 Whang arrive à cultiver des bactéries issues des sources chaudes de Yellowstone en allongeant la période d'incubation des échantillons prélevés. Colwell prouve l'existence de souches vivantes de *Vibrio cholerae* dans l'environnement dont il traite dans trois publications en 1986, 1990 et 2000. En 1996, le concept de micro-organismes viables mais non cultivables est bien reconnu. Des découvertes spécifiques mettent en valeur l'intérêt du monde non-cultivé. En 1985 Marshall fournit un exemple frappant du pouvoir des micro-organismes non cultivés en démontrant qu'une bactérie non cultivée cause des cancers et des ulcers gastriques (Handeslman, 2004).

Un tournant survient dans la perception du monde non cultivé avec l'application des principes de Carl Woese publiés en 1987. Woese montre que les ARNr sont des chronomètres de l'évolution et

permettent une mesure de la biodiversité. En 1985 Pace a l'idée de cloner directement l'ADN d'échantillons environnementaux. Il réalise l'analyse de séquences d'ARNr 5S et 16S de l'environnement, ce qui est un défi technique. Les développements de la PCR permettent la conception d'amorces pour amplifier un gène entier à partir de 1990. L'année 1995 voit la construction d'une librairie métagénomique d'organismes enrichis sur de l'herbe.

Aujourd'hui nous savons qu'une part minime de la biodiversité microbienne nous est connue. Nous observons également que dans l'environnement la dynamique des populations procaryotes et virales. Nous constatons également expérimentalement que les populations bactériennes ont une capacité rapide d'adaptation. Il est possible en culture de faire apparaître des variants de bactéries et de mettre en place des conditions favorisant la sélection de ces variants dans la population en quelques générations seulement. Nous savons donc d'une part par les apports de la métagénomique qu'il existe une biodiversité insoupçonnée et d'autre part, par les expériences au laboratoire, que le domaine procaryote est dynamique. Ainsi, il ne faut pas s'attendre à trouver les mêmes communautés dans un lieu donné selon la saison d'observation, ni à retrouver exactement les mêmes équilibres dans le temps à long terme.

Un renouveau de la perception des modes de transmission de l'information du monde vivant

Les premières conceptions scientifiques des règles de transmission de l'information génétique sont issues d'observations rigoureuses de caractères morphologiques macroscopiques menées sur le terrain sur la faune et la flore environnante. Dans son ouvrage publié en 1748, *Telliamed*, Benoist de Maillet envisage une transformation des espèces par des modifications héréditaires. Il suppose que la vie a pris naissance dans la mer et qu'un retrait des eaux a ensuite contraint certains êtres vivants à s'accoutumer à la vie terrestre (Lecointre *et al.*, 2009). A la suite de son voyage sur le *Beagle* Charles Darwin propose en 1859 dans *L'origine des espèces*¹ une illustration des apparentements entre espèces sous la forme d'un arbre. Peu à peu, toujours par l'observation de caractères morphologiques, la connaissance des règles de transmission de l'information génétique dans le monde vivant va se préciser. A la fin du 19^{ème} siècle, l'hérédité des caractères acquis va être définitivement abandonnée, ce qui ouvre l'ère du néo-darwinisme (Romanes, 1894). Grégor Mendel communique les trois lois fondamentales de l'hérédité en 1865, qui seront mises en relation avec la structure chromosomique du noyau cellulaire par Sutton en 1902 et vérifiées par Thomas Morgan en 1922 sur la *Drosophile*. La théorie darwinienne est complétée tout le long du XX^{ème} siècle par le croisement des travaux des généticiens des populations (Dobzhansky), des botanistes (Huxley), des zoologues (Mayr), des paléontologues (Simpson). Ainsi, en 1950, on obtient le consensus qu'est la théorie synthétique de l'évolution. Tout ceci se construit parallèlement aux progrès de la biologie moléculaire : la molécule d'ADN publiée en 1953 par James Watson et Francis Crick, puis la caractérisation moléculaire du gène et de ses régulations (avec par exemple les travaux de Jacques Monod en 1961). S. J. Gould puis Moto Kimura qui présente le neutralisme en 1967 s'inscrivent dans la synthèse darwinienne. Ces travaux scientifiques amènent tous à la description d'un modèle vertical de transmission des gènes², c'est-à-dire de la génération parentale vers les générations ultérieures. Ce n'est ni plus ni moins ce qu'exploitent les hommes depuis des millénaires par la sélection des plantes et des meilleurs têtes de bétail.

Toutefois, d'autres éléments allaient être mis au jour en génétique moléculaire, parmi lesquels la découverte du facteur conjugatif de fertilité F (années 1950) ou celle des transposons par Barbara

1 *On the origin of Species by means of natural selection or the preservation of favoured races in the struggle for life*

2 Ici, le terme gène est à prendre dans son acceptation d'unité d'information.

MacClintock entre 1940 et 1950. Globalement, l'ensemble des travaux de la révolution moléculaire de la décennie 1990 (aux rangs desquels la prise de conscience de l'existence de régulations épigénétiques³) a conduit :

- d'une manière générale à bouleverser le dogme central de la biologie moléculaire (l'expression du gène aboutit à la production de la protéine correspondante par l'intermédiaire d'un messager ribonucléique) ;
- en ce qui concerne particulièrement la transmission de l'information génétique, à envisager que des transferts autres que verticaux existent.

Au cours de ces dernières années, grâce aux progrès du séquençage, notre vision de l'évolution des génomes a radicalement changé, majoritairement grâce à l'analyse des génomes microbiens. A côté de la transmission verticale des gènes, il a été démontré que les génomes microbiens sont massivement affectés par des transferts latéraux de gènes, principalement via des éléments génétiques mobiles comme les virus, les plasmides, les intégrons ou les transposons (Toussaint et al., 2005)⁴. Ces échanges ont souvent une valeur adaptative et peuvent avoir d'importantes conséquences en santé publique, les transferts de gènes de résistance aux antibiotiques entre bactéries pathogènes étant l'exemple le plus connu.

L'importance des éléments mobiles à promouvoir les transferts latéraux a été étudiée dans les génomes séquencés mais l'ampleur réelle de ces transferts dans les communautés naturelles et au sein de divers environnements reste inconnue. La plupart de ces communautés étant intéressantes aussi bien d'un point de vue écologique, biotechnologique ou clinique il semble très important de comprendre comment les différents membres de ces consortiums interagissent, évoluent ou innoveraient via les flux latéraux de gènes. D'un point de vue conceptuel, pour paraphraser W. F. Doolittle l'étude de ces flux permettra de dire dans quelle mesure la vie n'est pas correctement reproduite par un arbre⁵ (Doolittle, 1999).

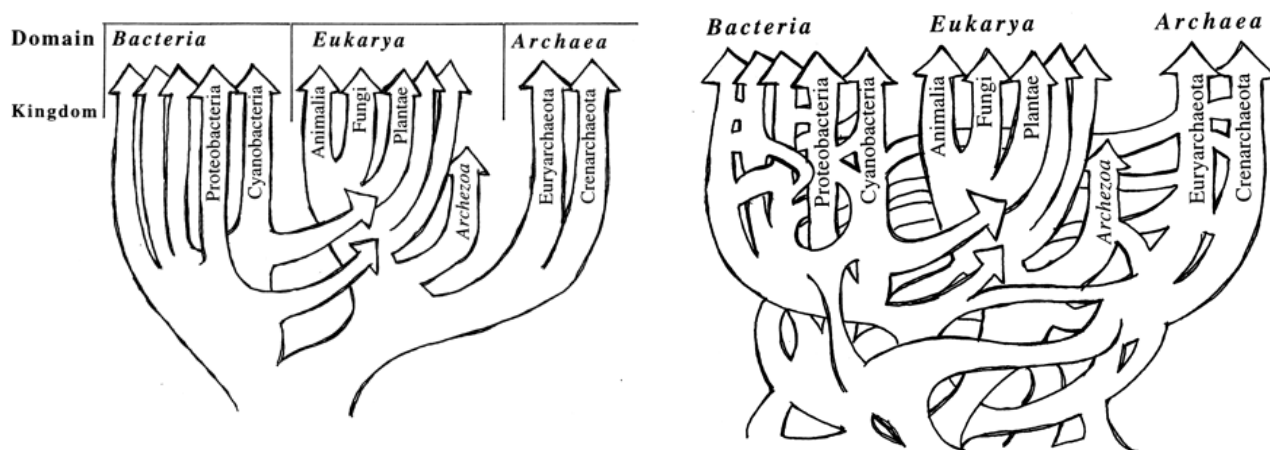


Figure 1 : les éléments génétiques mobiles renouvellent notre représentation du monde vivant. Illustration de W.F. Doolittle (1999). A droite on incorpore la transmission horizontale de portions de génomes.

3 A l'ère de la génomique, le séquençage a permis d'explorer des génomes reconstitués, mettant en évidence des régions génomiques ne cadrant pas avec le dogme central – les régions intergéniques ont d'ailleurs longtemps été considérées comme de l'ADN *poubelle* avant de mettre en relation leur séquence, leur localisation chromosomique et leur rôle dans la régulation de l'expression du génome *via* des mécanismes épigénétiques.

4 Ceux-ci seront succinctement présentés plus avant.

5 « *The history of life cannot properly be represented as a tree* », W.F. Doolittle.

Avec l'arrivée du séquençage des métagénomes, nous disposons de séquences d'ADN environnementales pour décrire les processus globaux d'évolution des procaryotes au sein de leurs communautés naturelles ; et notamment le rôle joué par les éléments génétiques mobiles. Ainsi la métagénomique permet-elle d'apporter une vision plus large, plus complète, d'accroître la profondeur d'échantillonnage. A terme, on peut souhaiter ou espérer que la vision procurée par une diversité de génome permettra d'avoir une vision plus globale des mécanismes de transmission de l'information génétique, de basculer de la connaissance qualitative des éléments à transmission horizontale vers une connaissance quantitative⁶, de pondérer l'équilibre entre verticalisme et horizontalisme dans chacun des regroupements obtenus jusqu'à présent par la méthode phylogénétique. En un mot il s'agit de faire évoluer notre perception du monde vivant, de dépasser les concepts établis, héritage de l'histoire du nécessaire réductionnisme scientifique.

2.2. Etat de l'art

Le métagénome est défini comme étant l'ensemble des génomes des micro-organismes présents dans une communauté microbienne donnée. L'exploration du métagénome représente une nouvelle stratégie conduisant à faire l'inventaire complet des gènes présents dans un environnement naturel. Cependant, ce serait donner une définition simpliste à la métagénomique que de la réduire à faire cet inventaire des gènes. La métagénomique est une science qui vise à analyser les métagénomes en exploitant différentes démarches qui font l'objet de deux dichotomies se recoupant. La première dichotomie intervient dans la nature des environnements étudiés : soit ouverts comme en plein océan (avec une faible densité de séquences pouvant appartenir à une même espèce) soit restreints comme dans un intestin (avec une haute densité de séquences appartenant potentiellement à une même espèce). Les problématiques informatiques (assemblage des séquences) sont différentes dans les deux cas. La deuxième dichotomie réside dans la méthode et le but de l'analyse. Premièrement, la métagénomique permet l'analyse des séquences pour l'étude de la biodiversité (cf. les expérimentations de Pace en 1985) ou la reconstruction de génomes. Deuxièmement, elle permet par une approche fonctionnelle de découvrir les fonctions présentes dans un environnement (sans chercher forcément l'attribution d'une fonction à l'organisme qui la porte).

Tableau 1 : les différentes sources d'échantillons et méthodologies de la métagénomique

		Approche séquentielle	Approche fonctionnelle
		analyse bioinformatique des séquences génomiques	identifier les clones qui expriment une protéine fonctionnelle dans l'hôte bactérien choisi pour le clonage
métagénomique à large spectre (environnements ouverts) <i>mer des Sargasses</i>	haute diversité de séquences, faible densité par organisme	<i>Approche séq. dans un env. ouvert :</i> Analyse de la biodiversité, statistiques sur motifs, génétique des populations.	<i>Approche fonctionnelle dans un env. ouvert :</i> Découvrir les métabolismes présents dans un environnement
métagénomique à spectre restreint (écosystèmes restreints tels que les symbioses) <i>Buchnera aphidicola</i>	haute densité de séquences par organisme, plus faible diversité	<i>Approche séq. dans un env. fermé :</i> Reconstruction des génomes	<i>Approche fonctionnelle dans un env. fermé :</i> Répartition des métabolismes entre hôte et symbiote

Apparue à la fin des années 1990, l'approche métagénomique a vu le jour grâce aux progrès spectaculaires de la biologie moléculaire, de la bioinformatique et des techniques de séquençage

6 Les études épidémiologiques réalisées pour étudier la propagation de virus dans une population constituent déjà un regard quantitatif.

ainsi que la baisse significative du coût du séquençage à haut débit. Le principe de la métagénomique étant d'analyser le pool génomique des communautés microbiennes sans avoir à isoler préalablement les micro-organismes, cette approche permet ainsi de mieux comprendre l'organisation des communautés microbiennes et le rôle écologique des micro-organismes qu'elles contiennent. Elle permet également de révéler les processus évolutifs des micro-organismes mettant parfois en évidence des événements de transferts horizontaux de gènes.

Précédents dans les projets de métagénomique

Les premières tentatives de reconstruction de génomes à partir de banques métagénomiques ont été initiées sur des communautés virales provenant d'océans et de fèces humains (Breitbart et al., 2002; 2003). D'autres études de métagénomique ont depuis permis de reconstruire une partie des génomes de micro-organismes incultivés. C'est le cas de l'étude menée sur le métagénome de la mer des Sargasses, située au large des Bermudes (Venter et al., 2004). Ce projet consistait à faire un inventaire le plus exhaustif possible des gènes et des génomes microbiens que renferme l'océan. Venter et son équipe ont filtré quelques centaines de litres d'eau avant d'extraire l'ADN génomique de la microflore retenue, puis ont construit plusieurs banques métagénomiques (avec des inserts d'ADN de petite taille, 2-6 kpb). Plus de deux millions de séquences d'ADN (soit environ 1.6 Mpb d'ADN) issues de ces banques ont été obtenues. Ces séquences ont été annotées, permettant d'identifier plus de 1.2 millions d'ORFs provenant d'environ 1800 espèces génomiques différentes. En avril 2004, 5% des séquences accessibles sur GenBank provenaient de la collection de la mer des Sargasses (Riesenfeld et al., 2004). Ce projet a également démontré l'utilité de l'approche métagénomique pour attribuer une fonction biologique aux lignées procaryotes incultivées. L'étendue de la diversité de certaines familles de gènes a également été révélée. C'est le cas des protéorhodopsines pour lesquelles 650 nouveaux photorécepteurs ont été identifiés.

Fort de ces premiers résultats, le Venter Institute a entrepris une campagne d'échantillonnage à l'échelle du globe avec pour but de donner un caractère quantitatif à la biodiversité des communautés microbiennes marines. Des échantillons ont ainsi été collectés durant l'expédition Global Ocean Sampling (GOS) sur le *Sorcerer II* entre août 2003 et mai 2004 (Venter *et al.*, 2007⁷). Durant cette expédition la majorité des échantillons ont été collectés à la surface de l'eau sur 41 sites différents d'échantillonnage séparés de 300 km. Ces distances ont permis de couvrir un panel important d'environnements de surface différents. Le trajet de cette expédition était : cote Est des Etats-Unis, Golfe du Mexique, les îles des Galapagos, l'océan Pacifique central et sud, l'Australie, l'Océan Indien, l'Afrique du Sud puis l'Océan Atlantique pour retourner aux Etats-Unis. 6,1 millions de protéines bactériennes et virales ont été annotées dans les séquences reconstruites à l'aide de la bioinformatique. Le but de ce projet était notamment de comparer les informations génomiques à grande échelle afin d'explorer la macro- et micro-diversité microbienne des océans dans un contexte biogéographique. Les séquences issues de la première phase de l'expédition GOS ont été mises à la disposition de la communauté scientifique au début de l'année 2007. En analysant 7,7 millions de séquences, Rusch *et al.* ont identifié 60 ribotypes très abondants (types ribosomiques qui être considérés comme l'équivalent des espèces). Cependant, la variabilité inter-espèce et la

7 Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Birmingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC, The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific, 2007

variabilité des organismes dans un même environnement suggère, qu'en dépit de leur forte similitude au niveau des gènes d'ARNr 16S, ils peuvent différer très largement au niveau physiologique et génomique. Il a également été montré que les données issues de GOS couvraient presque la totalité des familles de protéines de procaryotes (bactériennes et archéennes) connues et que plus de 1700 nouvelles familles avaient été identifiées (Yooseph et al., 2007).

2.3. Objectifs scientifiques prévisionnels des projets à venir

Vue générale

Actuellement, bien que les environnements étudiés représentent une diversité non négligeable, la majeure partie d'entre elles se sont concentrées sur des environnements océaniques ou estuariens de surface. Les futurs projets devront se proposer de partir à la découverte d'environnements novateurs avec des techniques d'extraction innovantes. D'une manière générale, la faisabilité de projets comportant une phase d'exploration marine est avérée. Par exemple, 540 campagnes d'explorations menées par 2700 chercheurs du programme Census of Marine Life (CoML) ont ainsi permis le recensement de 6000 nouvelles espèces. L'Ifremer est l'un des instituts à l'origine du projet.

Cependant, tous les projets doivent adopter la même méthodologie qui consiste à cloner dans un vecteur approprié de l'ADN extrait directement d'une communauté microbienne environnementale, puis à explorer le métagénome ainsi produit. Un aspect particulièrement ambitieux de ces projets pourra être l'étude (d'un point de vue statistique) des séquences d'éléments génétiques mobiles qui pourront y être détectées.

Analyse des gènes et des éléments mobiles

Les méthodes bioinformatiques de recherche de gènes couramment employées en laboratoire reposent sur la prédiction de potentiel codant et la recherche des cadres ouverts de lecture. Entre génomes proches, les résultats d'une annotation automatique peuvent être confirmés avec des recherches de similarités réciproques (dites *bidirectionnelles*). Certains gènes caractéristiques (par exemple les marqueurs historiques de biodiversité que sont les gènes codants pour l'ARN 16S) pourront être spécifiquement recherchés par des procédures d'annotations fines.

Les éléments génétiques mobiles peuvent également être recherchés par la bioinformatique. Ces éléments comportent les virus, de l'ADN libre du milieu assimilé lors de la transformation⁸, des éléments conjuguatifs intégrés ou non dans le chromosome, des loci recombinants connus sous le nom de transposons « voyageant » dans des cassettes mobiles de gènes (figure 2). On peut distinguer les éléments qui relèvent d'une mobilité extracellulaire (dans le cadre de la conjugaison, de la transduction et de la transformation). D'autres éléments relèvent de la mobilité intracellulaire (les transposons) mais peuvent également faire l'objet d'une transmission à une autre cellule s'ils ont incorporés dans un plasmide conjuguatif.

⁸ La transformation avait été rapportée dès 1928 par Fred Griffith, sans pour autant comprendre quel était le facteur transformant (en fait, l'ADN, comme montré en 1944 par Avery).

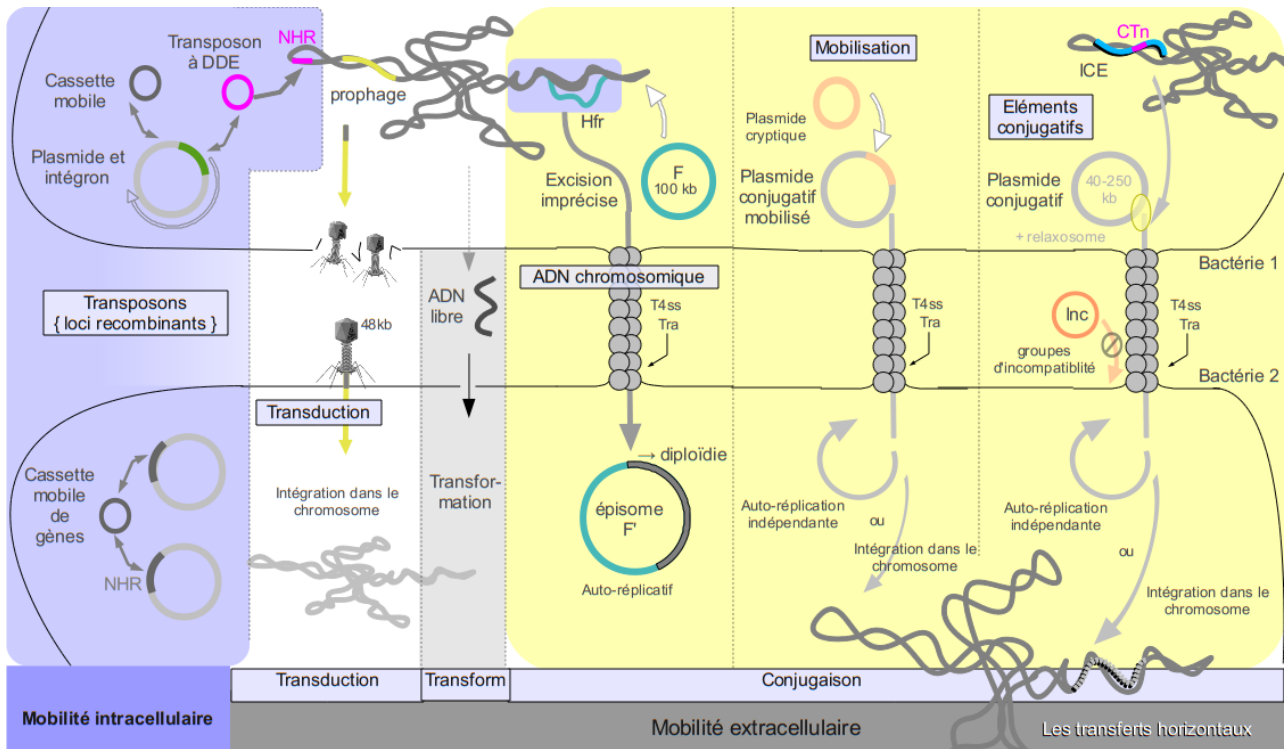


Figure 2 : les modes de transmission des éléments génétiques mobiles entre procaryotes

La recherche des gènes et de séquences particulières fournira des résultats descriptifs à court terme.

Cependant, l'objectif ultime de ce type de projet est de fournir des éléments pour réaliser une synthèse à très long terme. En un mot, il faudrait concourir à amplifier la connaissance de la biodiversité microbienne (et phagique) et compléter la connaissance darwinienne de la transmission de l'information génétique. Un des résultats attendus de cette recherche est de montrer comment la diversification de la nature des échantillons étudiés par les scientifiques permet aussi la diversification de la compréhension des mécanismes du vivant. Historiquement, l'analyse de la reproduction des êtres vivants situés à proximité directe de l'expérimentateur (les plantes, les métazoaires) a fait naître une théorie comprenant certaines règles. La question reste ouverte de savoir dans quelle mesure celles-ci, si elles restent universelles, doivent être complétées. Les êtres vivants numériquement majoritaires sur Terre (les communautés procaryote et virale) n'exploitent-ils pas en plus de ces règles d'autres lois ? L'importance quantitative de la transmission verticale et horizontale est-elle différente dans les groupes historiquement étudiés et ceux que la métagénomique permet aujourd'hui de commencer à éclairer ? Le phénomène horizontal, s'il semble être marginal chez les métazoaires et être historiquement également conçu comme tel chez les bactéries, pourra-t-il être quantifié sur les organismes procaryotes incultivés et gagnera-t-il un statut à part entière ?

3 . Aspects techniques :

3.1. Séquenceurs et bioinformatique

1/ La reconstruction des séquences (volet technique)

Avant de les analyser, il faut reconstituer les séquences obtenues par le séquençage des échantillons d'ADN prélevés au cours des phases expéditionnaires d'échantillonnage. A ce titre, il faut rappeler que les différentes technologies de séquençage n'offrent pas les mêmes caractéristiques techniques. Le caractère ambitieux des projets de métagénomique repose donc sur un volet technique ardu. Il est donc préférable d'utiliser un programme d'assemblage (par exemple Velvet) sur des séquences produites par la dernière génération de matériel de séquençage à haut débit. En effet, l'utilisation de séquenceurs modernes (produisant de longues séquences) réduit le risque informatique pendant la phase d'assemblage en produisant des lectures de l'ordre du kilobase. Quoiqu'il en soit, les algorithmes de reconstruction de séquences ne sont pas figés : c'est un champs à par entier de recherche. Il faut être un expert du domaine pour suivre l'évolution (très rapide) des méthodes. Cette phase de reconstruction, avant toute analyse, demande beaucoup de temps. Surtout si l'on souhaite incorporer des contrôles pour avoir un retour sur la qualité de la technique employée dans un projet.

Un moyen de gagner du temps est de permettre le recouvrement temporel du volet expéditionnaire avec le volet technique de reconstruction bioinformatique. Pour ce faire, on peut imaginer que des capacités de communication satellitaire à large bande soient emportés sur des navires d'échantillonnage emportant directement à leur bord les séquenceurs. Les données génomiques seraient transférées au laboratoire dès leur lecture. Ainsi le volet de reconstruction à terre au laboratoire démarrerait dès la réception des premières lectures de séquences.

2/ La bio-analyse : l'exploitation des données

Ce volet vise à l'exploitation des séquences reconstruites. On y recherche les traces de génomes communes (annotation et recherche de gènes à l'aide de procédures automatiques ou encore recherche par similarité *via* des algorithmes probabilistes de courts motifs représentant les éléments génétiques mobiles...). Dans l'optique d'un projet visant à renforcer la connaissances des règles de transmission de l'information, on peut établir des modèles de génétique des populations basés sur le signal vertical résiduel, en utilisant notamment les gènes issus de procédures d'annotation fines. Ensuite, il suffit d'y superposer les éléments mobiles (à la manière de Doolittle).

Il est à noter que l'analyse biologique des résultats ne constituer qu'une analyse préliminaire dans les temps impartis à un projet (les financement de l'ANR ne durent ainsi que trois ans). En effet, le volume de données apporté fournit du travail à toute la communauté scientifique pendant de nombreuses années. Il est donc important de prévoir une infrastructure le permettant (serveurs de calculs, bases de données).

3.2. Description des tâches

Le prélèvement d'échantillons

Un simple voilier peut suffire dans la majorité des cas à prélever des échantillons en mer. Toutefois, pour des projets lourds, il existe également des solutions. Par exemple, un navire à usage partagé entre la Marine nationale et l'Ifremer, le *Pourquoi pas ?* est doté de plus de 950 m² de locaux scientifiques, de réseaux informatiques scientifiques et multimédia avec liaison Inmarsat et peut déployer des engins remorqués et téléopérés (site internet de l'Ifremer, novembre 2010). Son autonomie est de 64 jours à 11 noeuds.

Le séquençage (méthode et choix de l'équipement)

Traditionnellement les méthodes classiques offraient des lectures plus longues (800pb) que les lectures des appareils à séquençage à haut débit durant la synthèse (454, Solexa/Illumina, SOLiD) (voir le tableau 2 et Metzker, 2010). Dans le cadre d'un projet de métagénomique, la taille relativement faible des lectures (50 paires de bases) des premiers séquenceurs à haut débit posait un problème de reconstruction bioinformatique. En effet, la densité des séquences appartenant à un même organisme dans un mètre cube d'eau de mer ou de terre est très faible. La probabilité que les séquences n'aient pas un recouvrement suffisant avec d'autres lectures posait des problèmes pour reconstruire des contigs et scaffolds⁹.

Technologie	Séquençage à haut débit durant synthèse						Sanger
	Roche 454				Solexa / Illumina	SOLiD	
Génération	GS20 (2006)	GS FLX Standard (2007)	GS FLX Titanium (2009)	GS FLX ERL (2011)	GA2X (2009)		
Lecture (pb)	100	250	500	1000	100	50	800
Taille séquencée	20 Mb / run	100 Mb/run	500 Mb/run	1 Gb /run	30 Gb / run	30 Gb / run	76 Kb
Temps par analyse	10 heures	10 heures	10 heures	10 heures	4 jours	7 jours	3 heures
Coût	20 k€ / Gb				0,3 € / Mb	1,5 € / Gb	0,5 € / lecture

Tableau 2 : caractéristiques techniques des technologies de séquençage. Sources : M. L. Metzker, *Sequencing technologies the next generation*, Nat Rev Genet, Jan. 2010 ; site internet de la société Roche ; cours de génomique, université Paris 7.

Pour garantir une bonne exploitation des échantillons récoltés il faut se baser sur des technologies à haut débit récentes et sur les dernières générations de machines. Celles-ci permettent aujourd'hui de

⁹ Malgré ces limitations les bioinformaticiens travaillant sur les échantillons de l'expédition Sorcerer II ont pu avec 4 124 495 contigs assemblés en 3 087 206 scaffolds annotés les 6,1 millions de protéines déjà citées (Venter_05). Ce qui faut retenir, c'est que l'on peut associer des types protéiques et des gènes à un environnement. Il reste plus difficile de reconstituer le génome complet de l'individu qui portait une carte de lecture annotée, encore plus de définir une espèce associée. De même, notre projet métagénomique fait le pari d'étudier l'enrichissement fonctionnel, qui donne une idée de la diversité spécifique, plus qu'il n'apportera des identifications d'espèces.

produire des tailles de lectures comparables à celles obtenues par la méthode Sanger. Par exemple, la dernière génération Roche 454® permet d'obtenir un ordre de grandeur de 1000 paires de bases par lecture (800 pb en moyenne selon les exemples fournis par la société). La dernière génération est dite « à lecture de tailles étendues » (Extended Read Lengths ou ERL).

Tableau 3 : caractéristiques du système GS FLX ERL®

Dimensions (largeur, profondeur, hauteur) du premier compartiment	74,3 cm x 69,8 cm x 36,1 cm (ou 82,5 cm avec l'écran)
Dimensions du compartiment inférieur	75 cm x 91 cm x 92,7 cm
Poids	242 kilogrammes

La figure suivante présente un exemple d'utilisation de la génération 2011 du système de séquençage GS FLX®. On observe la répartition des lectures obtenues au cours d'un séquençage selon leur taille. Les lectures atteignent 1000 paires de bases. Beaucoup de lectures sont comprises entre 750 et 900 paires de bases.

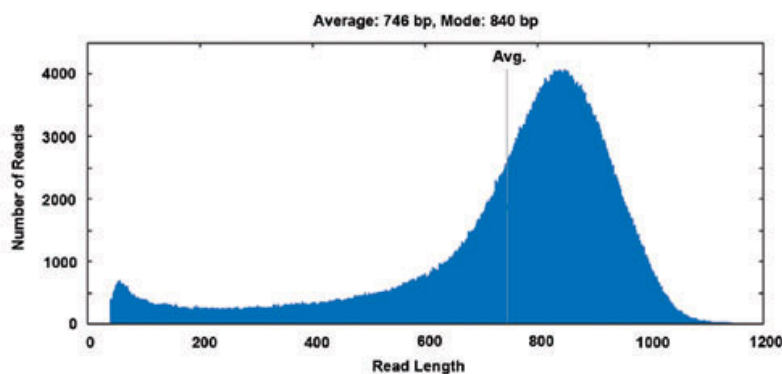


Figure 3 : répartition selon la taille des lectures obtenues au cours d'un séquençage avec le système GS FLX – ERL (Extended Read Length). La taille moyenne des lectures est de 746 paires de bases (Source : Roche®)

L'assemblage (éventuel) et la bioanalyse

Des outils d'assemblage existant aujourd'hui sur le marché, tel que Newble, peuvent être utilisés (surtout pour le cas de la métagénomique à spectre restreint, voir le tableau 1). Des programmes basés sur des recherches de similitudes de séquences, sur des recherche de motifs (exploitant les chaînes de Markov) peuvent être employés (notamment pour les projets à large spectre, voir le tableau 1). Pour les séquences d'insertion des transposons, des similarités avec les séquences actuellement déposées sur force de matrices poids-positions sur la base de données PFAM peuvent être utilisées. Enfin, pour trouver des phages et des intégrons, il existe les bases de données ACLAME et INTEGRALL.

4. Estimation financière d'un projet de trois ans

Ceci est une simulation se basant sur un projet fictif. On postule que ce projet contient une importante phase de navigation et de prélèvement dans des milieux novateurs (et donc très distants géographiquement) et difficiles techniquement parlent (exemple : prélèvements en profondeur à l'aide de robots de l'Ifremer). C'est donc un projet lourd. On imagine le recrutement de 7 personnels temporaires pour renforcer l'équipe de recherche en bioanalyse. On imagine que les équipages de l'Ifremer se chargent du travail de mise en œuvre des moyens de récolte.

Tableau 4 : composition hypothétique de l'équipe impliquée dans le projet

Personnel permanent CRS	2 DR – (Encadrement / analyse) 3 CR (analyse des séquences) 4 IE (assemblage / analyse des séquences)
Personnel CRS recruté sur contrat temporaire dont la prise en charge fait l'objet d'une demande d'aide dans le cadre du présent projet	4 post-docs (séquençage, assemblage, analyse) 2 IE (prélèvement / séquençage) 1 TR (informatique, création et maintien de la base de données)
Doctorants CRS (dotation financière extérieure)	2 (prélèvement/analyse)
Autres personnels (prestation de service)	2 techniciens Ifremer (robot/prélèvement)
Autres personnels (collaboration)	5 (scientifiques Ifremer) 20 (équipage Ifremer)

Postes budgétaires

On peut dresser la liste des postes de dépense. Les postes identifiés sont :

- (1) la masse salariale des chercheurs, ingénieurs et techniciens recrutées sur contrat temporaire chargés du prélèvement des échantillons, de l'assemblage des séquences génomiques et de l'analyse de celles-ci (4 post-doc et 2 IE CDD pour 144 k€/an, 1 TR à 21,6 k€ /an soit une demande d'aide de 507 k€ pour 3 ans) ;
- (2) la masse salariale de l'équipage de permanents Ifremer, assurée par l'Ifremer ;
- (3) une participation à la mise en œuvre des moyens de prélèvements marins (prestation de service assurée par l'Ifremer), y compris les frais incompressibles de logistique (dénommés frais d'administration générale) et y compris la mise en œuvre des moyens de communications satellitaires à large bande (5 jours de maintenance de Victor 6000 à 150 € + pièces de rechanges, prestation de service facturée à hauteur de 1 k€ au mois d'expérimentation à la mer, le tout estimé à 17,8 k€ auquel on ajoute les frais généraux soit une demande d'aide totale fixe de 21 k€) ;
- (4) l'achat d'équipements de séquençage de dernière technologie et de dernière génération, matériels qui seront réutilisables après l'opération dans le cadre de nouveaux projets (le

financement de deux systèmes GS FLX ERL à 500 k€ l'unité est assuré dans le cadre d'un financement de l'Union Européenne et ne fait donc pas l'objet d'une demande d'aide à l'ANR) ; ainsi que le soutien de ces équipements (l'aménagement du laboratoire à bord du navire de l'Ifremer pour recueillir les équipements de séquençage fait l'objet d'une demande d'aide à hauteur de 5,4 k€) ;

- (5) la mise en place des ressources informatiques d'assemblage des lectures issues des séquenceurs, de stockage des données de séquences assemblées et de leur diffusion à la communauté scientifique par une base de donnée (17 k€ pour l'ensemble du système informatique) ;
- (6) la location des emprises foncières permettant l'installation de l'équipe scientifique et de ses moyens informatiques (financement régulier du CRS).

On imagine que les dépenses seraient partagées avec l'Ifremer. La part d'aide au projet qui pourrait être apportée par l'EPA Agence Nationale de la Recherche serait consacrée au soutien de la masse salariale de la composante 1 des postes de dépenses ainsi qu'aux postes de dépense 4 (achat d'équipement dont la valeur unitaire est supérieure à 4000 € HT¹⁰) et 5.

Le reste des dépenses serait couvert par l'Ifremer, sauf pour les actes dits de prestation de service du titre 3 de dépense, financables par l'ANR¹¹. En effet, dans le cadre de ce projet l'Ifremer pourrait être déclaré prestataire de service pour deux opérations individuellement définies :

- le maintien en condition opérationnelle du système téléopéré sous-marin de prélèvement dans le cadre d'une part du titre 2 de dépenses (estimé à un jour de maintenance pour une semaine d'utilisation pour 10 € TTC l'heure de maintenance, soit 150 € la journée de main d'œuvre effectuée par deux techniciens payés selon la grille indiciaire de la convention collective de l'industrie de la mer, auxquels il faut ajouter au plus 1000 € de pièces de rechanges pour la durée de la campagne¹²) ;
- la mise à l'eau par des moyens de halage depuis la plate-forme maritime du système sous-marin (estimé à une trentaine d'heure de manipulation par mois d'expérimentation à la mer, soit ¼ de la masse salariale mensuelle de 2 techniciens Ifremer, ce qui fait environ 1 k€ au mois d'expérimentation à la mer) ;

Conformément au règlement de financement de l'ANR, il faut que le coût de ces prestations reste inférieur à 50 % du coût global entrant dans l'assiette de l'aide qui serait consacrée au projet.

Le paragraphe 3.1.1 du règlement relatif aux modalités d'attribution des aides de l'ANR rappelle que les dépenses de personnels prises en compte dans l'assiette de l'aide ne peuvent en aucun cas concerner des personnels permanents de ces établissements, à l'exception de leurs frais de déplacements engagés dans le cadre du projet. Ainsi, la masse salariale des personnels permanents restera assurée par leur organisme d'origine.

En conclusion, l'aide totale demandée à l'ANR pour ce projet hypothétique serait de 550,4 k€.

10 Cf. paragraphe 3.2 du règlement relatif aux modalités d'attribution des aides de l'ANR - Réf : ANR-REG-071220-01-01

11 Cf. paragraphe 3.1.2 du règlement relatif aux modalités d'attribution des aides de l'ANR - Réf : ANR-REG-071220-01-01

12 Dépense limitée par la conception de l'engin sous-marin.