Génomique des populations, génomique comparative, métagénomique et évolution

Enjeux scientifiques :

L'étude de la diversité des génomes implique une large gamme de disciplines allant des travaux sur la microévolution par des approches de génétique des populations à des études de processus de macroévolution à l'échelle de l'arbre de vie. La diversité du monde vivant doit être prise en compte, au niveau des populations, au niveau des écosystèmes (approches de metagénomique) ou à l'échelle évolutive (le temps). L'ensemble de ces différentes approches peuvent être groupées sous le terme de "Génomique évolutive". Les questions posées dans ce domaine recherche sont très diverses. Un but général est de *comprendre* la diversité de la vie sur la planète dans une perspective évolutive, en nous informant sur des questions aussi diverses que l'origine et la spécificité de l'espèce humaine ou le fonctionnement d'écosystèmes comme les systèmes planctoniques des océans ou la flore intestinale humaine.

La génétique des populations est une discipline qui a émergé dans les années 20 faisant la synthèse entre la génétique Mendélienne et la théorie de l'évolution. Elle se situe à l'interface entre génétique et évolution. C'est une discipline qui a un très fort support théorique et mathématique associant le développement de modèles et leur validation statistique. La génétique des populations s'apparente par de nombreux aspects à l'épidémiologie. Elle se base initialement sur l'analyse de traits phénotypiques et sur la comparaison de leur distribution dans différentes populations. Le passage de la génétique des populations vers la génomique des populations implique l'analyse de polymorphismes génétiques distribués le long du génome. Le nombre de marqueurs testés est très variable. Pour des espèces modèles, et en particulier pour l'homme, le nombre de polymorphismes considérés peut être très élevé. Les technologies de typage par puce à ADN permettent ainsi l'analyse simultanée de plus d'un million de SNPs. Par contre pour la majorité des espèces soumises à ces études de population, bien moins étudiées par ailleurs, uniquement quelques dizaines de sites polymorphes sont utilisés. Les progrès des techniques de séquençage bouleversent cette situation en permettant d'identifier rapidement et à un coût raisonnable un grand nombre de SNPs informatifs pour une espèce même peu étudiée. Nous vivons aujourd'hui une nouvelle évolution avec le passage de méthodes de génotypage à des méthodes de séquençage de génomes complets et également une explosion des coûts d'acquisition des données génétiques soulevant le problème de leur financement.

La comparaison des génomes joue donc aujourd'hui un rôle clef dans les études de génétique des populations et ce rôle sera grandissant avec les évolutions techniques. C'est pour cette raison que le GDR1928 Génomique des populations a intégré cette thématique dans ses domaines de compétence et à modifier son nom pour devenir "Génomique des populations et génomique évolutive".

La génomique comparative s'est initialement développée sur les bactéries. Trois mois après la publication en 1995 du premier génome séquencé, celui de la bactérie pathogène opportuniste *Haemophilus influenzae,* la publication du génome minimum de *Mycoplasma genitalium* a été centrée sur la comparaison de ces deux génomes. La génomique comparative est en fait l'outil principal d'analyse et d'annotation des génomes. Ces comparaisons peuvent faire intervenir deux génomes ou N génomes. La complexité de la génomique comparative est donc sans limite. C'est un domaine intense de recherche en bio-informatique. Les comparaisons peuvent être faites à différents niveaux, chaque niveau donnant accès à des types différents d'information. 1) À l'intérieur d'une espèce, les données génomiques pour

plusieurs individus apportent des informations sur la quantité et la distribution de la variabilité génétique (polymorphisme) dans le génome. Ces analyses peuvent être considérées comme une approche de type génomique des populations et est parfaitement complémentaire aux autres approches utilisées dans ce domaine. Ces approches comparatives permettent également en associant phénotype et génotype d'identifier des gènes candidats associés à une fonction. Un exemple d'application de ce type d'approche est la recherche de gènes de virulence par la comparaison de microorganismes proches pathogènes et non pathogènes. 2) Des comparaisons entre organismes plus éloignés permettent de retracer l'histoire des espèces en identifiant les étapes majeures évolutive (gains ou perte de gènes, duplication, envahissement par des éléments génétiques mobiles) en abordant les mécanismes moléculaires sous-jacents. 3) Enfin, des comparaisons entre organismes très éloignés permettent d'accéder à des événements majeurs dans l'évolution de la vie comme la diversification des lignées principales (animaux, champignons, plantes...), les relations évolutives entre ces groupes et l'acquisition des organelles. Cependant, malgré le nombre de génomes séquencés, les études comparatives semblent souvent assez naïves et les résultats parfois décevants. La comparaison du génome humain avec celui du chimpanzé ne nous donne pas la clef de la spécificité de l'espèce humaine.

La complexité des études en génomique des populations microbiennes et en particulier bactériennes, malgré la relative simplicité de leurs génomes vient de leur nombre et surtout de leur diversité. Le nombre d'espèces microbiennes estimé dans un gramme de sol est de l'ordre de 100.000. La notion d'espèce bactérienne est un sujet intense de discussion et l'analyse de la diversité montre que raisonner au niveau se l'espèce est bien trop réducteur. Il est nécessaire de prendre en compte les différents niveaux phylogénétiques ou "taxonomiques" et de raisonner à l'échelle des communautés. La caractérisation des communautés bactériennes était initialement basée sur la culture, pourtant la très grandes majorité des espèces se révèle non-cultivable. Suite aux travaux de Norman Pace, le développement d'outils moléculaires indépendants de la culture ont permis un inventaire de la diversité microbienne et de la comparaison de cette diversité. Ces divers travaux ont en particulier montré que la diversité du monde bactérien était bien plus importante que soupçonnée avec la découverte de nombreux phyla ayant aucun représentant cultivable. La métagénomique permet aujourd'hui d'aborder la génomique de ces communautés, en séquençant en masse l'ADN purifié. Le résultat de ces études est un ensemble complexe de séquences avec très peu d'organismes entièrement séquencés. La publication récente de la caractérisation métagénomique du microbiome digestif de 124 individus basée sur 574 Gbases de séquences a ainsi permis l'identification de 3.3 millions de gènes. L'étude de ces ensembles de séquences combinée aux séquences de génomes complets pose de nombreux problèmes informatiques, mais aussi méthodologiques et mathématiques. Aujourd'hui nous n'en sommes qu'au balbutiement de l'analyse de ces données, en particulier pour les comparaisons de communautés. Les études métagénomiques portent sur un grand nombre de communautés microbiennes commensales de l'homme et des animaux, symbiotique des plantes et d'un grand nombre d'environnements terrestres et marins. Ces études ne portent pas uniquement sur les bactéries, mais aussi sur les eucaryotes unicellulaires et sur les virus.

La génomique évolutive aborde des questions centrales en biologie sur l'évolution et la genèse des espèces. Elle est essentielle pour la compréhension de la dynamique des génomes et sur les interactions entre les espèces et entre les individus. Elle se situe au carrefour de nombreuses disciplines des sciences biologiques, mais également des sciences humaines et sociales (géographie, économie, ethnologie, histoire ...) et des sciences de l'environnement (podologie, écologie, climatologie). L'identification de gènes sous pression de sélection positive ou négative permet l'identification de polymorphismes correspondant à une adaptation particulière, à un trait spécifique. La génomique des populations et la génomique

comparative sont des approches puissantes pour la découverte de gènes impliqués dans un processus biologique (les études Evo-Devo illustrent la puissance de ces approches). Ces études présentent également un potentiel applicatif dans des domaines très variés. L'implication déjà ancienne d'instituts comme l'INRA, l'IFREMER le CIRAD ou l'IRD témoigne de l'importance de ces approches dans les domaines des productions alimentaires et de l'agronomie. Ces études sont aussi importantes dans le domaine environnemental et la gestion de la biodiversité animale, végétale et microbienne à l'échelle de la planète. La génomique des populations des microorganismes pathogènes et de leurs vecteurs permet d'aborder des questions multiples sur leur propagation, sur les phénomènes d'émergence sur l'adaptation à l'hôte et sur le passage de la barrière d'espèce ou l'impact du réchauffement climatique. Un sujet de prédilection de la génomique des populations est l'espèce humaine. De manière un peu prétentieuse, la génomique des populations et la génomique comparative de l'homme et des primates visent à comprendre ce qu'est l'homme. Un objectif de ces analyses est plus simplement la recherche de gènes associés à la prédisposition à certaines maladies.

Situation actuelle en France:

Génomique et génomique comparative

La France a eu un rôle reconnu sur le plan international en génomique au cours de ces 20 dernières années. Elle a eu tout d'abord un rôle moteur dans deux projets européens de séquençage, la levure *Saccaromyces cerevisiae* et la bactérie Gram positive *Bacillus subtilis*. La création du Genoscope l'a doté d'un outil performant en génomique. Les réalisations principales sont le chromosome 14 humain, en collaboration avec le Broad Institute, le poisson tetraodon et plus récemment la vigne, le protiste *Paramecium tetraurelia,* l'algue brune *Ectocarpus siliculosus* et la truffe. Ces différents projets ont apporté des données importantes sur l'évolution de ces espèces et du phylum auquel elles appartiennent. La politique de Genoscope, très ouverte vers le domaine de la génomique évolutive, a joué un rôle très important dans les développements de ce domaine en France. Le Genoscope a également contribué au travers de nombreuses collaborations à différents projets de génomique comparative en particulier en microbiologie. Le programme Genolovevure abordant la diversité génomique des levures ascomycète illustre le rôle fédérateur du Genoscope dans des projets de génomique comparative.

Les nouvelles espèces modèles: un des points forts de la recherche en génomique évolutive en France est le développement de nouvelles espèces modèles permettant non seulement des analyses de génomique comparative et des populations, mais aussi le développement d'approches de génomique fonctionnelle. Ces modèles comptent à la fois des espèces développées pour des approches de evo-devo a l'intérieur de groupes bien caractérisés, comme la roussette ou la méduse *Clytia hemisphaerica*, et des modèles permettant l'exploration de nouveaux groupes d'eucaryotes, comme l'algue brune *Ectocarpus siliculosus* ou le diatomée *Phaeodactylum tricornutum* pour les stramenopiles. Le développement de ces systèmes modèles représente un investissement important, souvent sur des périodes de plusieurs années.

La paléogénomique: un cas particulier est constitué par la paléogénomique ou l'étude des ADN anciens qui consiste à analyser des séquences d'espèces disparues pour mieux comprendre leur relation avec les espèces actuelles. Un exemple emblématique est l'étude de l'homme de Neandertal. Une difficulté majeure est l'obtention de l'ADN d'une qualité compatible avec le séquençage et dépourvu de toute contamination par de l'ADN exogène (ou

tout au moins d'avoir les moyen d'identifier ces contaminations). Quelques équipes en France (Lyon, Paris, Saclay et Marseille) ont déjà obtenu des résultats d'une grande portée. La difficulté dans ces analyses est l'accès au matériel biologique.

La Bioinformatique: La bioinformatique est une composante essentielle de la génomique comparative à la fois pour le développement de bases de données, les développements méthodologiques (algorithmique et représentation des résultats) mais aussi pour la réalisation d'analyses fines. Les équipes de bioinformatique du Genoscope réalisent ces analyses à la fois pour les procaryotes et pour les eucaryotes. De nombreux autres groupes sont impliqués à Paris (INRA et Institut Pasteur), Marseille, Toulouse et Lyon dans des projets de génomique comparative et développent des outils et des bases de données. Il faut noter que certaines équipes de bioinformatique ne sont pas associées aux équipes produisant des données, mais réalisent des études innovantes sur les données publiques. Il existe probablement un déficit de coordination de ces différentes initiatives.

La métagénomique: le Genoscope a eu un rôle *assez* précurseur dans ce domaine avec le projet Cloaca Maxima d'étude métagénomique des fermenteurs de la station d'épuration d'Evry. Ce séquençage a en particulier permis l'assemblage du génome d'une bactérie anamox (Anaerobic ammonium oxidation) non cultivable. De nombreuses équipes en Frances étudient par des approches moléculaires les flores commensales de l'homme (INRA Jouy en Josas), des sols ou des sites pollués (CNRS Lyon, Université de Strasbourg, Université de Pau), la rhizosphère (INRA, Nancy) les flores de digesteur (INRA) et finalement les flores de fromage ou d'autres produits alimentaires fermentés. Il existe une expertise reconnue en écologie microbienne (Lyon). Certaines équipes ont entamé des projets métagénomique, par exemple à Lyon, mais surtout à l'INRA avec les projets d'étude du microbiome humain (du tube digestif). Pour ce dernier projet, alors que le leadership initial était français, il a été repris par des équipes du BGI-Shenzhen qui ont réalisé l'essentiel des séquences et des analyses. En termes de bioinformatique, il semble que des approches relativement classiques aient été utilisées.

Actuellement, la France a également pris un rôle moteur dans l'analyse de la microflore océanique avec le projet Tara Océans dont le Genoscope coordonne la composante génomique (http://oceans.taraexpeditions.org/). Une originalité de ce projet est de se focaliser sur le plancton eucaryote. Cette étude génomique sera associée à une étude de méta-transcriptomique.

La génomique des populations : l'analyse des équipes participant au GDR1928 donne un aperçu de la diversité des thématiques et des espèces en Génomique des populations. Les acteurs sont très divers : CNRS, INRA, IRD, IFREMER, Inserm, Institut Pasteur, MNHN et les thématiques également. Néanmoins de nombreuses équipes développant des projets en génomiques des populations ne sont pas encore associées à ce GDR. C'est le cas d'équipes travaillant sur des microorganismes pathogènes (parasites, bactéries et virus), mais aussi sur des vecteurs. On peut noter deux tendances symétriques, des équipes spécialistes dans l'étude des populations ont amorcé la transition vers la génomique comparative et la génomique des populations. D'autres équipes spécialistes en génomique ont réalisé l'importance d'étendre leurs études de génomique comparative au niveau des populations. Ces communautés ont des compétences complémentaires, mais également des "incompétences". Cette diversité de point de vue peut constituer une richesse si des structures permettent leur collaboration.

L'enseignement: la génomique comparative et la métagénomique sont des composantes importantes des cursus en bioinformatique, et ne justifient pas d'enseignement spécifique. La situation est différente pour la génétique des populations qui repose sur un ensemble de concepts complexes et d'outils mathématiques. Quelques formations existent :

- M1 au MNHN - EPHE (http://www.mnhn.fr/oseb/Genetique-des-populations).

- M2 rec. anthropologie, génétique des populations humaines (université Paul Sabatier, Toulouse)

- M2 Recherche Biologie des Organismes et des Populations (ENESEAD, Dijon) - amélioration.

Défis pour l'avenir

De manière schématique, on peut identifier trois phases dans la réalisation de projets en génomique comparative, métagénomique et en génomique des populations:

1. La collecte d'échantillons biologiques.

C'est l'étape clef de ces études. Les conclusions dépendront de la qualité de cette collection. Il est à souligner aussi qu'une collecte est faite en fonction d'une question particulière. La réutilisation de collections existantes représente une économie, mais aussi un risque : celui d'utiliser une collection mal adaptée à la question posée. La France a des atouts indéniables pour réaliser ces études à l'échelle de la planète. Le réseau des stations de biologie marine, les équipes de recherches dans les différents départements et territoires d'outre mer et les réseaux des instituts internationaux de recherche (IRD, CIRAD, Institut Pasteur) sont des acteurs privilégiés pour la collectes d'échantillons. L'intégration dans des réseaux internationaux est aussi essentielle.

Il n'est néanmoins pas envisageable de faire un inventaire de la diversité génomique de l'ensemble des êtres vivants de la planète. Le choix des espèces ciblés pour le séquençage et pour l'analyse des populations peut être basé sur différents types de questions soit avec un objectif appliqué à plus ou moins long terme: santé, environnement production agricole ou pour répondre à des question fondamentale dans des domaines très diverse sur l'évolution, la dynamique des génomes ou le développement (evo-devo). Ces choix doivent ternir compte de priorités scientifiques ou appliqués et de la compétition internationale. Il semble néanmoins nécessaire d'avoir les moyens de coordonner les études et éviter une trop grande dispersion. Cela peut être par des appels à projets spécifiques. Le GDR et les plates-formes régionales avec le Genoscope pourraient aussi contribuer à cette coordination.

La conservation et la mise à disposition de ces échantillons biologiques (cellules, ADN, graines) est un problème récurrent. Ils peuvent être conservés la durée de l'étude ou de manière plus pérenne sous forme de collections (au sein de Centres de Ressources Biologiques). Les collections sont un "gros problème" pour les institutions et pour les financeurs, par exemple le Muséum National d'Histoire Naturelle dont c'est une des missions peine à obtenir les budgets nécessaires à l'entretien et la valorisation de ses collections qui sont pourtant réputées être un outil scientifique majeur face aux enjeux actuels sur la biodiversité. La mise en place d'une collection est aisée avec un coût acceptable, par contre la pérenniser est bien plus ingrat et cher. Faire vivre une collection avec (pour) des projets de recherche est tout un art. Les financeurs ont du mal à en évaluer le bénéfice avec un nombre d'utilisateurs parfois limité. Certaines collections ont eu un impact majeur en génétique, c'est le cas du panel de familles du CEPH ou les collections du MNHN qui sont à la base du projet mondial BarCode. Nombre de collections sont des ressources de grande valeur pour des projets de séquençage comparatif à grande échelle.

2. l'acquisition des données: génotypage, séquençage complets.

L'évolution des capacités et des caractéristiques des séquenceurs permet à une équipe ayant accès à une telle machine de séquencer 1000 génomes bactériens, des échantillons métagénomiques de taille moyenne ou *quelques* génomes de mammifère. Les centres

régionaux de séquençage doivent donc avoir les moyens de réaliser ces projets de génomique comparative et de génomique des populations à une échelle "moyenne". Il semble nécessaire qu'ils acquièrent les méthodes (multiplexages, combinaison de différents type des banques: simple, paired ends et mate pairs) et soient associés à des équipes en bio-informatique. Par contre pour les projets réellement ambitieux se pose le problème d'être compétitif avec les grands centres de séquençage internationaux et en particulier aux Etats Unis et en chine. Ainsi le projet d'étude du microbiome du tube digestif humain a été réalisé en Chine par le BGI-Shenzhen avec une contribution mineure du Genoscope. Une augmentation significative de la capacité de séquençage du Genoscope semble nécessaire ainsi qu'une meilleur définition des rôles respectifs du Genoscope et des plates-formes régionales afin que le Genoscope puissent se focaliser sur des projets réellement ambitieux comme la projet Tara Océans. Par ailleurs la situation actuelle d'absence d'interactions entre le Genoscope et le CNG est fortement préjudiciable à l'émergence d'une véritable culture de génomique des populations.

Les données à l'échelle du génome sont aujourd'hui indispensables pour rester compétitif à un niveau international. La mise en place des équipements ne posera pas de problème, par contre les coûts en consommable sont extrêmement élevés et difficile à financer. La composante bioinformatique et son coût sont également sous évalués

3. <u>l'organisation des données sous forme de bases de données et leur analyse statistique informatique et bioinformatique.</u>

Les technologies haut débit et la réduction des coûts de génotypage et de séquençage permettent de typer avec précision un très grand nombre d'individus. Le projet "1000 génomes humains" de séquençage massif d'individus d'origine diverse en est un exemple. Ce nombre de 1000 génomes sera très probablement largement dépassé. Cette quantité d'information soulève la question de son stockage et son organisation sous forme de bases de données. Les progrès en informatique et dans la compression des données permettent de stocker ces données. Il est néanmoins envisageable de stocker ces données de manière plus élaborées en termes par exemple de différence. Le changement d'échelle implique aussi des innovations informatiques que ce soit pour l'analyse des données ou la représentation des résultats. La communauté française en bioinformatique intéressé par ces questions est significative, elle semble néanmoins manquer d'équipes réellement leader à l'échelle internationale.

Le changement d'échelle n'implique pas uniquement des développements informatiques, mais également méthodologiques, mathématiques et conceptuels. On peut s'interroger si les modèles classiques en génétique des populations ne doivent pas être revus. Si des démarches nouvelles ne doivent pas être initiés. Les résultats en génomique microbienne ont montré par exemple que raisonner à l'échelle de l'espèce était trop restrictif. L'importance des transferts horizontaux souligne l'importance de replacer l'espèce au sein de la biodiversité microbienne et au sein de la communauté. #qu'en est il pour les eucaryotes#

Il ne faut également ne pas oublier la question éthique soulevée par ces études sur l'homme et l'impact des résultats obtenus.

Propositions d'actions:

- Propositions d'actions spécifiques au domaine de la génomique évolutive:

Échantillonnage et collections d'échantillons. Avec l'augmentation de la capacité de séquençage, la collecte et le stockage d'échantillons (souvent sur la forme de collections dans les centres de stockage) deviendront de plus en plus une étape clé dans les approches de génomique évolutive. Il faut que cet aspect soit intégré systématiquement dans les financements de projet ou il faut crée des support spécifique pour ce type d'activité ou renforcer ceux existants dont la compétence est connue.

Appel d'offres spécifiquement dans le domaine de la génomique évolutive. Il y a un besoin de financement spécifique pour des projets en génomique évolutive, que ça soit des projets de génomique comparative, des projets de metagénomique ou des projets impliquant les nouveaux organismes modèles. Les financements devraient prendre en compte l'entièreté d'un projet (pre- et post-séquençage) et non seulement l'aspect séquençage.

Développement d'approches multidisciplinaires.

- Propositions d'actions non spécifiques au domaine de la génomique évolutive

Mise en place de plateformes performantes utilisant des nouvelles technologies de séquençage. L'évolution rapide des technologies de séquençage pose un vrai défi en ce qui concerne la position de la recherche en biologie en France par rapport à la communauté internationale. Il faut un programme spécifique visant à augmenter la capacité de séquençage et à intégrer l'évolution technologique dans ce domaine. En complément à ces plateformes, il faut mettre en place des moyens (humain et technique) en bioinformatique nécessaire pour l'analyse des données générées.

Formation. Il y a besoin de formation en génomique, en particulier en ce qui concerne la bioinformatique et l'application de principes de la génétique des populations à des données de type génomique.