Contributions à la Stratégie Nationale de Recherche
ITMO Génétique, Génomique et Bioinformatique

L'ITMO Génétique, Génomique et Bioinformatique (GGB) vise à coordonner les efforts de recherche autour de la Génétique et la Génomique de tous les organismes vivants depuis les virus, les microorganismes, les plantes jusqu'à l'homme. La recherche fondamentale correspondante s'attache à élucider les bases de l'organisation, de la stabilité, de l'évolution et de la régulation de l'expression du matériel génétique. Un des enjeux fondamentaux du domaine est de comprendre les liens complexes entre génotype et phénotype, ce qui requière d'intégrer la connaissance des mécanismes fondamentaux, la description de la variabilité génétique inter-espèce ou inter-individuelle au niveau populationnel avec l'analyse phénotypique des systèmes modèles et des êtres humains. Les approches à haut débit, dites omiques, et la modélisation et l'analyse bioinformatique des données induisent d'importantes avancées conceptuelles. Les perspectives cognitives et appliquées à la médecine et à l'agronomie de ce domaine sont immenses.

I. Les grands enjeux scientifiques pour la recherche en biologie et santé

Une des priorités de l'ITMO GGB est d'intégrer et d'anticiper les implications du développement du séquençage de nouvelle génération (*Next Generation Sequencing ou NGS*), révolution technologique sans précédent dont les impacts cognitifs médicaux et sociétaux sont considérables.

Les évolutions spectaculaires des débits et la diminution des coûts font qu'il est désormais possible d'avoir accès à l'intégralité du génome de la plupart des espèces, et en particulier de l'espèce humaine, d'en étudier leur variabilité ce qui constitue une étape historique dans la connaissance du vivant et d'embrasser la biologie du vivant dans sa globalité et sa complexité. Le séquençage du génome humain, basé sur le séquençage classique selon la méthode de Sanger développée en 1977, aura nécessité plus de 10 ans, 2.7 milliards de dollars et fut achevé en 2004. Les nouvelles technologies de séquençage de l'ADN qui ont émergé au début de ce siècle sont basées sur le séquençage en masse et en parallèle de courts fragments d'ADN isolés dans des microréacteurs ou sur support solide puis amplifiés de façon clonale. Ces technologies permettent de séquencer des génomes d'une taille importante, comme le génome humain, en l'espace de quelques semaines, voire de quelques jours pour des coûts de quelques dizaines de milliers d'euros et qui diminuent régulièrement pour tendre vers le millier d'euros. Ce saut technologique, compte tenu des masses de données générées, a nécessité d'emblée des infrastructures et expertises en bioinformatique indispensables au stockage des données, à la filtration et comparaison des séquences et à l'interprétation des variations génétiques. Au-delà des génomes, les technologies de NGS permettent de caractériser le transcriptome et l'épigénome des cellules et ainsi d'établir la base de la compréhension du lien entre génotype et phénotype et du déterminisme épigénétique de ce lien. Des grands programmes internationaux (essentiellement américains) de caractérisation systématique du fonctionnement des génomes humain, murin, ainsi que d'autres organismes multicellulaires modèles comme la mouche du vinaigre, le nématode, ou l'Arabette (Encode et ses déclinaisons) ont utilisé ces approches de NGS pour générer d'importantes bases de données qui forment un socle pour une description formelle de la relation entre génotype et phénotype et la compréhension du déterminisme génétique et du fonctionnement du génome. En parallèle, d'autres grands programmes internationaux visent à caractériser systématiquement la variabilité génétique dans les populations humaines mais aussi d'autres organismes modèles ou non (Projet 1000 génomes et ses déclinaisons). L'intégration des données génomiques fonctionnelles et de variabilité

génétique au niveau populationnel, aussi bien chez l'homme que chez les autres organismes, représente un enjeu majeur du développement des connaissances dans ce domaine.

Le NGS a également permis de spectaculaires progrès dans la caractérisation de la diversité microbienne dans des environnements variés. Il a ainsi été possible de mettre en évidence le rôle joué par le microbiome intestinal dans certains aspects de la physiopathologie humaine. Ce domaine est en plein développement et met en évidence l'importance de la compréhension de la symbiose entre organismes multicellulaires et bactéries. La diversité des populations de microorganismes est une dimension qui doit fort probablement être intégrée avec le génotype et l'épigénétique pour permettre une compréhension globale du déterminisme du phénotype.

Ces développements fondamentaux ouvrent des perspectives de retombées aussi bien en santé humaine, qu'en agronomie et pour les applications biotechnologiques permises par l'intégration d'une démarche de biologie systémique avec une perspective de biologie synthétique, toutes deux découlant des progrès de la synthèse entre approches omiques, bioinformatique et modélisation.

Ia. Génétique humaine

En ce qui concerne le génome humain, les premières applications du NGS se sont focalisées sur les parties codantes des gènes c'est-à-dire les exons qui représentent environ 34 Mb des 3 Gb du génome humain et sont actuellement analysés sous l'appellation exome. Si les théories de génétique des populations prédisaient que dans les séquences, de très nombreuses variations nucléotiques (*Single Nucleotide Variation* ou SNV) rares pourraient être mises en évidence, les données réelles n'ont pas contredit ces prédictions. L'exome de chaque individu présente un nombre considérable de SNV de l'ordre de 17 000, dont environ 500 ne sont pas présentes dans les bases de données et ont une fréquence allélique < à 0.1%, certaines variations étant privées et restreintes à l'exome d'un individu. Chaque exome contient environ 500 SNVs susceptibles d'avoir un impact biologique et, de façon remarquable, en moyenne apparait à chaque génération un *SNV de novo* avec un impact biologique potentiel. La majorité des SNVs susceptibles d'altérer la fonction des protéines sont probablement apparus récemment dans l'espèce humaine au cours des 10000 dernières années voire même des 2000 dernières années (100 générations) pendant lesquelles la population humaine a connu une croissance explosive, passant de quelques millions d'individus à plus de 7 milliards aujourd'hui. Cette variabilité génétique de l'ADN est encore considérablement sous estimée puisque les données disponibles ne concernent que les exomes, soit 1.2% du génome. Le reste du génome humain, incluant régions régulatrices et introniques des gènes et les régions inter-géniques, constitue un énorme réservoir de variations génétiques susceptibles d'être impliquées dans le déterminisme des maladies : il est en effet très probable, comme cela commence à être illustré par quelques maladies Mendéliennes ou multifactorielles, que des altérations génétiques situées en dehors des régions codantes altèrent l'expression génique et contribuent au déterminisme des maladies. La caractérisation de ces altérations nucléotidiques ou génomiques nécessitera des analyses génomiques complètes couplées à des analyses d'expression et l'interprétation biologique de ces variations sera encore plus complexe puisque la variabilité de ces régions n'a pas été caractérisée. Le développement des technologies omiques permet de passer progressivement d'une biologie réductrice restreinte à l'étude d'un gène ou d'une protéine à une biologie plus globale embrassant l'étude de voies et réseaux biologiques et d'appréhender la complexité des bases physiopathologiques des maladies de façon plus intégrée.

Ainsi la connaissance de la variabilité génétique des exomes puis du génome humain, des mécanismes de cette variabilité, de son évolutivité et de ses impacts biologiques et

médicaux représente le principal enjeu cognitif de la génétique et génomique humaine pour les 20 prochaines années.

Le NGS et la puissance de l'analyse bioinformatique ont provoqué depuis 2008 un bouleversement de l'étude du déterminisme génétique des maladies humaines et dans les possibilités diagnostiques des maladies humaines. Il est en effet possible d'identifier, à partir d'un nombre limité de patients, les bases génétiques des maladies en réalisant une analyse comparative d'exomes soit de patients présentant la même maladie mais appartenant à des familles différentes, soit de patients appartenant à la même famille, soit enfin d'un patient et de ses parents indemnes pour identifier une mutation de *novo*. Cette connaissance extensive des bases génétiques des maladies conduit au développement croissant de diagnostics génétiques permettant une prise en charge personnalisée précoce et codifiée des patients et permet d'adapter les traitements des maladies en fonction des caractéristiques génétiques des individus. Avant le NGS, la recherche sur le déterminisme génétique avait dichotomisé de façon réductrice les maladies en 2 classes extrêmes : les maladies Mendéliennes résultant d'altérations délétères d'un gène et les maladies multifactorielles résultant de la combinaison de facteurs environnementaux et de multiples variations génétiques présentes dans la population générale qui augmentaient le risque de maladie sans être ni nécessaires ni suffisantes. L'identification des bases génétiques des maladies Mendéliennes résultant d'altérations récurrentes de gènes majeurs fut, au cours des 20 dernières années, une des plus grandes avancées médicales permettant la compréhension de la physiopathologie de ces maladies (en particulier grâce à la possibilité de créer des modèles animaux pertinents pour ces affections), la mise au point de diagnostics moléculaires, le développement du diagnostic pré-symptomatique, le suivi médical et le traitement personnalisé des patients et de leur famille. En revanche, les succès et retombées des études du déterminisme génétique des maladies multifactorielles ont été beaucoup plus modestes. En effet, ces études reposent sur une stratégie d'association pangénomique (GWAS) consistant à comparer entre 2 groupes de sujets, malades (cas) et témoins, la fréquence de variations génétiques communes et à rechercher une ou des variations qui sont distribuées différemment entre ces deux groupes et à calculer pour les sujets présentant ces variations un risque relatif (*odds ratio*) de développer la maladie. Ces études, qui se sont multipliées, ont mobilisé des ressources considérables en termes de nombre d'individus analysés et d'analyse bioinformatique. Les odds ratio estimés par ces études sont, pour la plupart, faibles de l'ordre de 1.2 et les écarts-types sont importants. D'autre part, ces études ont fait le plus souvent abstraction de l'impact des facteurs environnementaux. La plupart des études cas-témoins, qui ont eu une contribution significative, sont celles qui ont intégré la connaissance biologique des maladies à l'analyse statistique. Le NGS a initié une relecture de la complexité du déterminisme génétique des maladies. La découverte de l'étendue du polymorphisme, du nombre considérable de SNVs privés et du taux de mutation par génération révélés par le NGS ont démontré que des maladies initialement considérées comme monogéniques, telles les maladies dominantes à pénétrance incomplète résultaient de la combinaison de quelques variations génétiques et avaient en fait un déterminisme oligogénique et, qu'à l'inverse, des maladies initialement considérées multifactorielles et survenant de façon sporadique au sein d'une famille résultaient d'une mutation délétère survenue *de novo*. De nombreuses maladies se caractérisent probablement par une très grande hétérogénéité génétique impliquant de nombreux variants individuellement rares mais de fort impact. Depuis l'irruption du NGS en génétique et son implémentation progressive dans le diagnostic des maladies génétiques, le défi n'est plus la détection des variations génétiques mais leur interprétation. L'interprétation des variants repose (1) sur la prédiction *in silico* de leur impact biologique grâce à des algorithmes de valeur prédictive qui donnent des résultats très hétérogènes, (2) des analyses biologiques réalisées *ex vivo* ou dans des modèles cellulaires ou animaux et, dans ce contexte

les organismes simples facilement manipulables et de cycle de vie court compatibles avec une analyse rapide, tels le zebrafish, sont de plus en plus utilisées et (3) des arguments génétiques qu'il s'agisse de la co-ségrégation du variant avec la maladie au sein d'une famille ou de la mise en évidence d'un enrichissement statistiquement significatif de ce variant chez les malades par rapport aux témoins. Or actuellement la fréquence des SNVs rares dans les populations témoins est très mal caractérisée et n'est pas stratifiée par groupes de population, les seules données actuellement disponibles provenant soit du projet 1000 génomes qui couvre 14 populations dans le monde dont 5 populations européennes peu représentatives de l'étendue de la diversité génétique en Europe et en France en particulier, soit du consortium nord-américain ESP (*Exome Sequencing Project*) dichotomisant la population américaine en 2 groupes d'origine européenne et africaine. Ainsi, compte tenu du nombre de SNVs rares voire privés présents chez tout sujet et du nombre de SNVs à impact biologique potentiel, l'interprétation des variations génétiques, c'est-à-dire l'étude de leur contribution aux maladies, représente l'enjeu majeur de la génétique médicale puisque cette interprétation est indispensable pour comprendre les bases génétiques des maladies, leur physiopathologie, caractériser des traceurs génétiques pertinents, évaluer de façon fiable le risque associé à ces variations génétiques, développer des diagnostics génétiques fiables, conditions *sine qua none* au développement d'une réelle médecine personnalisée de qualité.

La question du traitement représente l'enjeu ultime de la recherche sur les maladies génétiques. Le transfert de gènes dans les cellules d'un patient atteint d'une maladie génétique est l'approche thérapeutique la plus rationnelle et la thérapie génique a démontré son efficacité chez l'homme dans des maladies génétiques telles les déficits immunitaires, l'adrénoleucodystrophie et dans des phases pré-cliniques développées dans des modèles animaux de maladies génétiques sensorielles telles les rétinites. Néanmoins la thérapie génique ne peut pas être l'approche universelle des maladies génétiques : cette approche peut être appliquée aux maladies génétiques héréditaires résultant de l'inactivation de gènes impactant des cellules qu'il est possible de cibler. Comme le montre le bilan de 20 ans de génétique moderne, le traitement des maladies génétiques n'est pas restreint à la thérapie génique : (1) le diagnostic génétique d'une maladie concourt indirectement au traitement puisqu'il évite un retard au diagnostic chez le sujet atteint. Il permet la mise en place de mesures et traitement médicaux et chirurgicaux appropriés prévenant les complications à l'origine de morbidité et mortalité, et ouvre la possibilité de réaliser un diagnostic pré-symptomatique au sein des familles facilitant la détection précoce voire la prévention des maladies. (2) Les thérapeutiques substitutives telles l'enzymothérapie développée par l'industrie et consistant en l'administration de la protéine défectueuse au patient, ont démontré leur efficacité pour freiner l'évolution de maladies métaboliques à l'origine de retards mentaux. (3) L'utilisation détournée de médicaments déjà mis sur le marché et prescrits dans d'autres indications, dès lors que leur mécanisme d'action est susceptible d'avoir un effet sur l'expression de la maladie, présente comme intérêt pour les patients d'être disponibles immédiatement.

Les avancées considérables de la génétique et de la génomique médicale doivent être accompagnées d'une réflexion éclairée pour que les applications se développent exclusivement pour le bénéfice des patients et de la société. Avant l'ère du NGS, le développement de tests génétiques commerciaux accessibles directement par internet sans consultation médicale préalable (*Direct-to-consumer genetic testing* services ou DTC) illustrait déjà un premier risque de dérive. La plupart de ces tests analysent les variants polymorphiques présents dans la population générale, identifiés comme facteurs de risque pour certaines maladies multifactorielles mais conférant des odds ratio faibles voire très faibles. Ces facteurs génétiques sont majoritairement situés dans les parties non codantes du génome, souvent dans des régions potentiellement régulatrices, et il reste à comprendre si et

comment ils sont impliqués dans la maladie. On appréhende encore très mal leur rôle dans les maladies multifactorielles et on ne pourra certainement pas le faire sans tenir compte des autres facteurs génétiques et environnementaux qui sont susceptibles d'interagir et de modifier leur effet, pouvant passer de délétères à protecteurs selon le contexte ou le phénotype étudié. Enfin, les risques quantifiés par les odds ratios et qui servent de base à ces calculs prédictifs s'entendent au niveau de la population et non pas au niveau individuel contrairement à ce que les services DTC peuvent laisser croire au consommateur qui, recevant les résultats de ces tests sans l'accompagnement médical adéquat, peut ne pas bien en appréhender toute les limites. L'accessibilité à l'intégralité du patrimoine génétique permise par le NGS soulève de nouvelles questions éthiques et expose à d'autres dérives, la première d'entre elles étant le réductionnisme génomique développé en particulier par des sociétés privées. Selon cette vision, le séquençage complet des exomes ou d'un génome d'un individu permettrait d'emblée, sans analyse de son phénotype, la détection de ses risques génétiques de développer des maladies et la mise en place de mesures préventives. Le nombre de variants rares d'impact biologique potentiel dans tout génome, la difficulté de leur interprétation qui exige la connaissance du phénotype et la méconnaissance des variants rares dans les différents groupes de population expliquent pourquoi l'analyse génomique sans analyse phénotypique de qualité n'a aucun sens. Cette perception erronée de la complexité du déterminisme génétique de la plupart des maladies apparait même dans certains documents officiels européens surestimant la puissance informative de la médecine génomique (*European Science Foundation, Personalised Medicine for the European Citizen, 2012*). L'accessibilité non contrôlée, par exemple pour des compagnies d'assurance, à l'intégralité de l'information génétique expose à des risques de discrimination génétique en particulier dans ce contexte de non compréhension de la signification des variations génétiques et de complexité du déterminisme génétique des maladies génétiques. Une autre dérive non médicale d'une accessibilité non contrôlée est l'utilisation des données du séquençage des exomes et génomes pour déterminer à des fins non scientifiques l'origine des individus. Les défis de la génomique humaine sur le plan sociétal sont donc la diffusion des connaissances auprès de la société pour éviter une utilisation inappropriée et une réglementation des nouvelles analyses génomiques.

Ib Les mécanismes fondamentaux affectant le génome

Pour interpréter correctement les données de variation génétique, il est essentiel de comprendre les mécanismes fondamentaux contrôlant le maintien de l'intégrité des génomes, la création de la diversité génétique, l'organisation, l'évolution et l'expression des génomes en incluant les dimensions génétiques et épigénétiques affectant ces contrôles. Ces champs disciplinaires sont extrêmement dynamiques et compétitifs et jouent un rôle clef dans la compréhension des propriétés du vivant conférées par le génome et ses modalités d'expression. Les dernières années ont été marquées par de nombreuses avancées en épigénétique et épigénomique, comme la découverte de très nombreux ARNs non codants, la caractérisation fine de l'organisation chromosomale et chromatinienne à l'échelle du génome, et les rôles régulateurs variés que jouent ces différents acteurs pour contrôler de façon parfois dynamique, parfois métastable, l'activité du génome et la manière dont il détermine les propriétés du vivant. Ces développements nous amènent à reconsidérer sans cesse notre vision des relations génotype/phénotype. Par ailleurs ces découvertes permettent d'appréhender avec de nouveaux outils des phénomènes biologiques complexes (différenciation, développement, vieillissement, adaptation…) et de revisiter des notions aussi importantes que la pénétrance incomplète, la vigueur hybride, l'effet de l'environnement sur l'expression des gènes, etc. Les enjeux sociétaux sont également importants, les altérations et autres variations des mécanismes de régulation génique sous-tendant de nombreuses pathologies humaines, y

compris le cancer, et étant sans aucun doute responsable d'une part considérable de la composante héritable de la plupart des traits complexes. L'épigénétique est un champ d'investigation d'importance stratégique dans le domaine de la génétique, et la France figure parmi les leaders dans ce domaine.

Le NGS a un très fort impact sur l'étude des mécanismes fondamentaux car il permet des analyses diverses à l'échelle du génome étudiant la conformation et la topologie des chromosomes et des domaines chromatiniens, la distribution des marques épigénétiques chromatiniennes et l'expression du génome… La compréhension des mécanismes fondamentaux nécessite de plus d'intégrer ces données avec l'échelle cellulaire en y adjoignant l'imagerie cellulaire à haute résolution qui permet d'analyser la diversité des comportements à l'échelle de la cellule individuelle et pas uniquement d'avoir une vision moyennée d'une population de cellules. L'intégration de l'ensemble de ces données représente une nouvelle frontière de ce domaine scientifique. De plus, le NGS augmente aussi considérablement l'efficacité de nombreuses stratégies de criblage génétique, soit parce qu'il permet de cribler en masse les mutations sans passer par la sélection des clones, soit parce qu'il accélère la caractérisation moléculaires des mutations.

La compréhension de l'impact de la diversité génétique requière d'intégrer les différents niveaux où elle s'exerce, que ce soit les variations interindividuelles au sein d'une population ou d'une espèce, ou de la diversité interspécifiques au travers de l'arbre du vivant ou au niveau des écosystèmes. Il est essentiel d'intégrer aussi l'échelle temporelle et la dimension évolutive. La génomique évolutive aborde des questions centrales en biologie sur l'évolution et la genèse des espèces et nous informe 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. Elle permet la compréhension de la dynamique des génomes et des interactions entre les espèces et entre les individus. Elle permet aussi l'identification de gènes sous pression de sélection positive ou négative et donc des polymorphismes correspondant à une adaptation particulière, à un trait spécifique. Une perspective large est essentielle pour interpréter correctement la variabilité génétique à toutes les échelles d'étude.

Il existe en France un nombre important d'équipes dynamiques et performantes étudiant ces différents aspects fondamentaux, comme le montre la contribution importante de ce domaine aux financements ERC obtenus par les équipes françaises ainsi que la place des équipes françaises dans le réseau européen d'excellence en épigénomique. Il est essentiel de soutenir les efforts dans ce domaine pour pouvoir maintenir cet avantage compétitif, assurer la vitalité du domaine et disposer d'un tissu de recherche qui permettra les retombées en terme de santé et d'applications qu'apportera une meilleure compréhension des liens entre génotype et phénotype.

Ic Génétique microbienne et métagénomique

L'analyse génétique des microorganismes a contribué très significativement à la compréhension des mécanismes fondamentaux affectant le génome, et l'analyse génomique des microorganismes individuellement a apporté des contributions majeures à la génomique évolutive et comparative (voir paragraphe précédent). Le NGS ouvre de nouvelles perspectives en offrant la possibilité d'accéder aux génomes d'organismes difficilement cultivables et contribue ainsi à la mise en évidence de nouveaux gènes et de nouveaux organismes. Il est maintenant possible de caractériser globalement des populations microbiennes complexes en effectuant des analyses métagénomiques. Ceci ouvre de multiples perspectives. D'une part, l'analyse des microbiomes intestinaux permet de mettre en évidence l'influence de ce microbiome sur le fonctionnement normal et pathologique des organismes qui les hébergent. L'apport des microbiomes à l'activité métabolique de ces organismes en

fait une composante essentielle à prendre en compte pour intégrer correctement la relation entre génotype et phénotype. Les conséquences cognitives et pratiques de la prise en compte des microbiomes sont très importantes. D'autre part, l'analyse métagénomique des microbiomes dans d'autres environnements, comme par exemple des environnements extrêmes, ouvre des perspectives à la fois cognitives et applicatives : par exemple, dépollution, nouvelles voies de synthèse de nouveaux composés avec un fort potentiel de retombées économiques. La France possède un tissu fort de recherche fondamentale en microbiologie et en génétique microbienne et a fait des contributions significatives en métagénomique. Il est important de soutenir activement ces domaines de recherche.

Id Les génomes des animaux et plantes d'intérêt agronomique

L'agriculture française apparait aujourd'hui plus que jamais comme un enjeu socio-économique de premier plan. L'objectif est toujours de garantir la production agricole en quantité et qualité suffisantes pour satisfaire les besoins alimentaires de la population, mais en prenant en considération les changements climatiques et les normes environnementales plus strictes (réduction des pesticides et des engrais). Par exemple, connaitre les bases génétiques des sensibilités aux pathologies des animaux et plantes domestiques ouvre la possibilité de minimiser l'effet néfaste des maladies sur les rendements mais aussi de mieux contrôler l'usage des pesticides et les transferts de pathogènes d'animaux domestiques à l'homme. A cet objectif de production alimentaire s'ajoutent les objectifs de production de biomasse non-alimentaire à des fins industrielles, ainsi que des objectifs environnementaux. Les applications en sont la chimie verte (remplacement des hydrocarbures fossiles) et la résiliance au changement climatique des écosystèmes nous environnant.

La génomique constitue un outil puissant pour relever ces défis, en fournissant les outils et les connaissances permettant une meilleure compréhension des bases moléculaires de la variation phénotypique, et accélérant l'exploitation de la diversité génétique. La compréhension du lien entre génotype et phénotype est tout aussi essentielle pour les organismes d'intérêt agronomique que pour les autres, et les progrès des connaissances obtenus quel que soit les organismes étudiés seront utiles pour l'étude des autres. Les retombées représentent des enjeux économiques et scientifiques colossaux.

Le décryptage de la séquence complète d'un génome marque le début d'une augmentation considérable des analyses génétiques et génomiques réalisables sur une espèce et sur ses espèces apparentées. Ainsi, l'analyse de la séquence génomique permet une identification accélérée des facteurs génétiques impliqués dans un trait d'intérêt. C'est aujourd'hui un point d'entrée incontournable à la dissection génétique d'un phénotype. La définition et la construction d'idéotypes, l'éco-physiologie, et la gestion de la biodiversité sont les cibles de ces efforts.

Caractérisant la structure des séquences génomiques, la génomique décrypte et facilite le travail aval visant à identifier des gènes d'intérêt agronomique, préserver la diversité génétique, améliorer les variétés ayant un potentiel agronomique. En effet, elle aborde tout autant le fonctionnement du génome, avec les mécanismes épigénétiques impliqués dans la régulation de la transcription, que les aspects structuraux, avec la dynamique des variations de séquences et leurs impacts sur l'adaptation et sur les traits d'intérêt agronomique.

De nombreux programmes de séquençage des différentes variétés d'une même espèce sont en cours. Ils cherchent à identifier les polymorphismes associés aux différences phénotypiques entre variétés, pointant ainsi la ou les régions portant les gènes responsables. Ces programmes doivent pouvoir s'intensifier pour concerner l'ensemble des espèces d'intérêt agronomique. Mais il ne faut pas oublier dans cet effort le séquençage d'espèces sauvages proches, permettant d'identifier des réservoirs de gènes d'intérêt, inexistant dans nos espèces cultivées, voire caractériser de nouvelles espèces à domestiquer pour notre agriculture

car mieux adaptées à certains environnements, portant de nouvelles caractéristiques alimentaires, ou ouvrant de nouveau débouchés industriels (biomasse ou chimie verte).

Ainsi la génétique et génomique agronomique bénéficieront des études menées au sein d'Aviesan, et pourront aussi contribuer au progrès des domaines qui sont bien représentés au sein d'Aviesan. Particulièrement, la génétique et la génomique des animaux et des plantes domestiqués offrent des modèles particulièrement intéressants pour suivre des évolutions rapides des génomes et des phénotypes au cours des 10000 dernières années depuis la révolution néolithique. C'est donc un secteur où la génomique évolutive, ainsi que la paléogénomique, peuvent contribuer très significativement à la compréhension des bases génétiques des phénotypes d'importance agronomique.

Id Bioinformatique, Modélisation, Biologie systémique et Biologie Synthétique

Le génome, codé à l'aide de 4 bases, se prête remarquablement à l'analyse digitale. La génomique est donc à l'origine de l'apparition de la bioinformatique dont l'irruption au sein de la biologie est une mutation de grande ampleur. L'enjeu majeur de la recherche en bioinformatique est de comprendre comment assister le biologiste, l'agronome ou le médecin dans la modélisation et la compréhension au niveau moléculaire d'un comportement particulier sur lequel on dispose de données brutes massives et également de connaissances plus ou moins bien formalisées que l'on doit intégrer dans les méthodes d'analyse. L'opportunité de disposer d'un cadre rationnel pour expliquer les grands équilibres physiologiques et leurs déplacements au sein des cellules, jusqu'au niveau moléculaire, en élucidant les différents types d'interaction qui interviennent entre les éléments cellulaires et avec l'environnement ne deviendra une réalité qu'en multipliant les recherches méthodologiques sur ces systèmes extrêmement complexes, en relation avec les laboratoires de biologie. Les notions de robustesse et de régulation sont des enjeux de formalisation importants si l'on veut à plus long terme espérer effectuer un contrôle stable, voire une réparation des dysfonctionnements du vivant. L'intégration d'une recherche bioinformatique sèche de haut niveau avec les sciences expérimentales biologiques humides est un enjeu majeur pour la biologie en général et pour les domaines relevant de l'ITMO GGB en particulier. La connaissance de l'outil bioinformatique est une compétence qui est devenue en quelques années indispensable pour les biologistes et en particulier pour les généticiens et les génomiciens. Cette connaissance est nécessaire pour leur permettre d'interagir productivement avec des bioinformaticiens et pour identifier les analyses bioinformatiques les plus pertinentes par rapport à la question posée, ainsi que pour mettre en œuvre leurs expériences de la manière adaptée au traitement bioinformatique qui sera réalisable.

L'intégration des analyses omiques, issues du NGS mais aussi des autres approches globales, protéomique, métabolomique, interactomique, avec les autres données phénotypiques est essentielle pour permettre la compréhension des bases du déterminisme génétique du vivant de la situation normale comme de la situation pathologique. La modélisation et la bioinformatique, associées aux analyses omiques permettra de mettre en œuvre une approche de biologie systémique qui ouvre des perspectives immenses, aussi bien sur le plan cognitif que pour les applications qu'elle rendra possibles. Une approche systémique peut permettre d'identifier les causes profondes des pathologies dont le déterminisme est en partie génétique, et ainsi permettre de mieux caractériser la part génétique, la part épigénétique et la part environnementale de la pathologie, ce qui ouvre des perspectives d'approches thérapeutiques ciblées sur les dysfonctionnements et non sur la cause génétique per se, ce qui permet d'envisager l'utilisation de médicaments plus traditionnels que la thérapie génique pour traiter certaines pathologies génétiques.

Finalement, la biologie systémique rend possible une approche de biologie synthétique visant à reprogrammer génétiquement des organismes, particulièrement des microorganismes, pour rendre possible de multiples applications à visée thérapeutique ou autres.

II. Etat des lieux, forces et faiblesses de la France, positionnement au niveau Européen et International

En génétique, les indicateurs bibliométriques (Web of Science - Thomson Reuters) placent la France en 3ème position au niveau européen après le Royaume Uni et l'Allemagne. Dans le domaine «Genetics & Heredity » et sur la période 2003-2012, la très grande majorité des publications les plus citées dans le monde et à plus fort impact sont d'ordre technologique et/ou méthodologique et il n'existe pas d'affiliation française, ce qui illustre dans ce domaine le sous-développement actuel de la recherche technologique et méthodologique en France. La France, dans les années 1990, grâce à l'équipe de Jean Weissenbach, avait joué un rôle historique dans la cartographie génétique puis physique du génome humain. Le Génoscope a contribué à la caractérisation de nombreux nouveaux génomes. La France a également une bonne, voire très bonne, lisibilité dans la recherche en génétique médicale. Chaque année, le nombre d'articles signés dans la revue *Nature Genetics*, journal d'excellence d'impact facteur 36, est de l'ordre de 21 et les publications françaises en génétique représentent 20% des Top 10% internationales. L'impact des équipes travaillant sur les aspects fondamentaux de la génétique, épigénétique et génomique est aussi important et place la communauté de ces chercheurs parmi les meilleures mondiales. Au cours des 20 dernières années, les équipes françaises ont contribué de façon très significative à l'identification de gènes impliqués dans les maladies Mendéliennes et à la caractérisation de leur physiopathologie. La proximité des laboratoires de recherche et des services de génétique, l'expertise clinique assurant une qualité de l'évaluation phénotypique et l'organisation très structurée de la génétique médicale au niveau national avec de nombreux réseaux ont contribué, en France, à ces succès. L'intrication des laboratoires de recherche et des laboratoires diagnostiques a permis une rapide diffusion des connaissances dans la pratique médicale pour le bénéfice des patients. La France a également eu une contribution historique dans le domaine de la thérapie génique. En revanche, le cas de la bioinformatique est plus paradoxal car, selon la manière d'interroger les bases de données, on note une très forte disparité dans le rang international de la France. Si on se focalise sur les articles mentionnant les termes "algorithme" et "génétique" pour la période 2008-2012, La France est à la 11ème position mondiale et à la 7ème place pour l'impact des publications, 7ème place conférée par un seul logiciel blockbuster. Par contre, si on analyse l'ensemble des publications dans les revues de bioinformatique, la France occupe la 4ème place. Ces différences révèlent que si la communauté bioinformatique a un bon niveau en ce qui concerne le développement de méthodes et d'algorithmes et la production d'outils d'analyses des données biologiques, ces approches n'ont qu'un impact relativement faible auprès des généticiens, biologistes ou médecins. Deux facteurs expliquent cette faiblesse : les équipes de bioinformaticiens ne sont pas assez intégrées dans la communauté des chercheurs en sciences du vivant ; les outils informatiques développés ne sont pas suffisamment diffusés auprès des biologistes et médecins, vraisemblablement faute de la mise en place d'interfaces conviviales. Une amélioration notable de la place de la France pourrait donc être obtenue avec trois mesures. La première serait de soutenir la mise en place d'interfaces conviviales autour des meilleurs logiciels développés, ce qui devrait être assuré dans le cadre de l'infrastructure d'avenir de l'IFB (Institut Français de Bioinformatique). La seconde serait de développer sur le territoire national une politique active et coordonnée de formation à la bioinformatique destinée aux généticiens et aux biologistes. La troisième serait de stimuler le regroupement des communautés de bioinformaticiens, de généticiens et de biologistes au sein de grands

centres de recherches qui favoriseraient la transition vers la nouvelle biologie intégrant approches expérimentales et formelles.

De plus, la France a raté la révolution technologique du NGS initiée dès 2008 et accuse un retard conséquent par rapport à d'autres pays Européens, tels les Pays-Bas et l'Allemagne, qui ont rapidement mis en place des infrastructures nationales de séquençage haut débit bénéficiant de compétences bioinformatiques de haut niveau. Ce retard s'explique en partie par le sous développement de la recherche technologique en France, le cloisonnement de la bioinformatique et de la biologie, l'absence de stratégie nationale anticipée et réfléchie et, dans ce contexte, l'absence d'une infrastructure nationale de séquençage dévolue aux maladies humaines a fait cruellement défaut. Si le Centre National de Génotypage (CNG) et le Centre National de Séquençage (CNS) à Evry ont installé un parc conséquent de séquenceurs de nouvelle génération, ces installations se sont réalisées plus tardivement que dans d'autres pays et à un rythme moindre. La Chine par exemple a fait de façon précoce le choix d'investir massivement dans les technologies de NGS et s'est dotée avant 2010 d'un centre, le BGI, doté d'environ 200 séquenceurs NGS, encadré d'un millier de bioinformaticiens, c'est-à-dire à une échelle très supérieure au plus grand centre français qui est encore loin d'approcher cette échelle. De plus, les choix stratégiques, analyse de maladies multifactorielles pour le CNG, séquençage *de novo* pour la caractérisation de nouvelles espèces pour le CNS et leur éloignement géographique des centres de génétique ont conduit à ce que les équipes françaises sollicitent l'expertise en NGS d'autres structures, qu'il s'agisse de structures privées françaises ou de plate-formes de NGS localisées à l'Etranger. Depuis 2012, la réorientation stratégique du CNG et sa participation active aux programmes de séquençage d'exomes en collaboration avec d'autres plate-formes nationales, dans le cadre des appels d'offre de la Fondation pour les Maladies Rares (FMR) permettent à la France de réduire progressivement le retard accumulé en génétique humaine. Mais il convient certainement de poursuivre cet effort de développement d'un grand centre français de séquençage à haut débit et d'inciter à la mise en place de bases de données publiques pour répertorier les variations génétiques mises en évidence et faciliter l'interprétation des données issues du NGS.

III. Propositions, priorités organisationnelles, scientifiques, technologiques, médicales

1. La diffusion du séquençage de nouvelle génération, l'accessibilité aux plate-formes de NGS et les modalités de stockage des données du NGS est une priorité qui doit être organisée et coordonnée au niveau national en ayant à l'esprit que le défi essentiel sera l'analyse des données. En intégrant le continuum nécessaire en génétique entre recherche et diagnostic, 3 niveaux de NGS peuvent être considérées : NGS de niveau 1 (0,5-2 GB) pour le séquençage ciblé de régions d'intérêt à la fois pour une utilisation diagnostique en génétique médicale, et pour permettre des réponses souples et réactives aux équipes de recherche et favoriser la diffusion des compétences en analyse génomique; le niveau 2 (50-100 Gb) pour l'analyse globale d'un nombre limité d'exomes, de marqueurs génétiques, ou de transcriptomes et un réseau de plate-formes assurant des prestations de site et nationales de ce niveau s'est constitué en France ; le niveau 3 correspondant à du très haut débit capable d'assurer des analyses en masse d'exomes et de génomes. Ce niveau ne peut être développé que dans une ou 2 structures organisée(s) selon un mode industriel et dotée(s) d'expertise et de capacités bioinformatiques à la hauteur. Une première hypothèse serait que le CNG soit cette structure, compte-tenu de son parc technologique et des investissements qu'il représente, sous réserve qu'il soit investit d'une mission nationale, s'intègre dans les réseaux nationaux de génétique et bénéficie d'une infrastructure bioinformatique renforcée et de moyens de stockage adaptés.

2. L'intégration de bioinformaticiens formés au NGS dans les équipes de génétique est indispensable. Il y a toutefois un déficit significatif en personnes formées compte tenu des demandes très fortes. La formation à la bioinformatique appliquée au NGS des étudiants et les actions de formation des personnels en poste doivent être orchestrées au niveau national. Pour inverser graduellement le déficit d'enseignement en bioinformatique, il parait essentiel de donner une formation solide en bioinformatique à tous les étudiants en biologie, et particulièrement à ceux qui se spécialisent en génétique. Pour inciter à monter en puissance rapidement, il semble sage de faire dépendre la validation des cursus de Biologie offerts par chaque Université à un nombre minimal d'ECTS consacrées à la bioinformatique et à la biostatistique depuis la Licence jusqu'à la thèse. Il faut aussi augmenter de façon concomitante le nombre d'enseignants-chercheurs sur des profils "analyses génomiques, bioinformatique et biostatistique". Finalement, la formation continue en analyses génomiques est une demande très forte des chercheurs et enseignant-chercheurs qui relèvent de notre ITMO. Ce qui domine alors, c'est un apprentissage très ciblé : connaissance, utilisation et maîtrise des outils bioinformatiques. Les universités peuvent jouer un rôle centralisateur utile dans ce domaine face à la disparité de fonctionnement des instituts. L'effort de formation permanente devra également s'appuyer sur l'IFB et France Génomique. Pour remédier à l'urgence, l'ITMO GGB a mis en place une telle formation continue et s'efforcera de coordonner les actions de formation en bioinformatique organisées sur le territoire national, mais l'ITMO ne devrait pas avoir vocation à assurer de telles formations de façon pérenne.
3. Le développement des approches de biologie systémique et l'intégration des approches expérimentales et bioinformatiques sont des enjeux stratégiques majeurs. Il faut encourager le regroupement d'équipes de bioinformatique et de biologie expérimentale et/ou biomédicale au sein de grands centres de recherche. La caractérisation phénotypique des variants génétiques, aussi bien dans les populations humaines que dans les autres espèces est un enjeu essentiel. L'étude de la fréquence des variants rares du génome humain par groupe de populations est un socle indispensable pour interpréter les variants détectés chez des patients. Cela suppose de disposer de cohortes de témoins parfaitement caractérisées sur le plan phénotypique et le développement de nouvelles approches statistiques. Dans ce cadre, La France peut jouer un rôle stratégique. Les Centres d'Investigations Cliniques (CIC) implantés sur les CHU, répartis sur le territoire et le plus souvent affiliés à l'Inserm permettent de recruter des témoins correctement phénotypés. Dans la même perspective, le phénotypage haut-débit des modèles animaux ou des plantes doit être soutenu. Les projets investissement d'avenir comme Phenome, Phenomin et Tefor amorcent le mouvement, mais l'effort doit être poursuivi si l'on veut élargir le nombre d'espèces pouvant bénéficier de ces équipements. La génétique des populations est une discipline dans laquelle la France a un positionnement international et le développement de nouvelles approches en biostatistiques permettant l'étude de variants rares doit être stimulé. La génétique quantitative, encore appelée génétique statistique, délaissée un temps, doit être développée pour permettre d'exploiter correctement les phénotypages et génotypages haut-débits.
4. Le développement de plusieurs modèles animaux de cycle de vie court et dans lesquels la manipulation génétique est aisée sera crucial pour tester rapidement l'impact des variations génétiques. La diversité des modèles est essentielle pour comprendre la variabilité des réponses phénotypiques à des mêmes mutations selon le contexte génomique. Les collaborations entre équipes de génétiques avec des experts de ces modèles doivent être incitées.

5. La caractérisation des bases épigénétiques des maladies humaines, tel le cancer et les maladies liés à l'âge, ainsi que les bases épigénétiques des traits agronomiques représentent un enjeu, considérable compte-tenu du vieillissement de la population et du changement climatique. En particulier l'étude des variations des régions génomiques cibles des modifications épigénétiques telles la méthylation est une voie de recherche de grand intérêt. Il est donc important de soutenir aussi bien des recherches fondamentales de pointe dans ces domaines qu'intégrer la qualité et l'exigence requises pour valider ces recherches fondamentales dans les études réalisées sur les populations humaines et les espèces modèles.
6. Le potentiel qu'offrent les approches de génomique évolutive et de génomique des populations à la compréhension du déterminisme génétique doit être exploité. Par exemple, la génomique et la transcriptomique comparatives inter-espèces, entre primates, animaux modèles, domestiques ou d'élevages, ou plantes, seront probablement de grand intérêt pour identifier des gènes impliqués spécifiquement dans des traits génétiques et étudier la complexité et l'évolutivité génétique (gain ou perte de segments génomiques, duplications, envahissement par des éléments génétiques mobiles).
7. La biologie intégrée est particulièrement importante pour appréhender les phénotypes dans leur globalité. De nombreuses retombées des approches de biologie intégrative se profilent. Ainsi pour les maladies génétiques, une approche globale dans leur dimension physiopathologique et thérapeutique, permettra d'explorer des approches thérapeutiques utilisant des cibles connues et des médicaments déjà sur le marché. Les traits agronomiques modélisés dans leur dimension écophysiologique devraient permettre d'identifier les idéotypes de demain, guidant alors l'amélioration variétale. Cela nécessite de stimuler, faciliter et renforcer les coopérations pluri-disciplinaires entre biologistes, bioinformaticiens, mathématiciens et physiciens et la formation aux mathématiques et à la physique des étudiants en biologie.
8. Compte-tenu des enjeux sociétaux de la génomique, il est impératif de définir un cadre réglementaire encadrant l'information des patients, le partage des données, les modalités de réalisation des analyses génomiques, l'accessibilité des données du NGS, leur stockage sur le long terme, leur l'utilisation et le transfert des données vers des pays étrangers.