Texte de la 27ème conférence de l'Université de tous les savoirs réalisée le 27 janvier 2000 par Jean Weissenbach

Le séquençage du génome humain : comment et pourquoi.

Introduction

Les cellules des êtres vivants contiennent un programme d'instructions (le génome) leur permettant de se maintenir en vie ou de se multiplier. Ces instructions (les gènes) sont codées sous une forme chimique le long de molécules géantes, les molécules d'ADN qui constituent les chromosomes. La connaissance de ces instructions est indispensable à la compréhension des phénomènes biologiques au niveau cellulaire et moléculaire. Mais elle est en outre le point de départ d'applications de plus en plus nombreuses dans les domaines de la médecine et des industries pharmaceutique, biotechnologique, agro-alimentaire et dans d'autres domaines en prise directe avec les processus biologiques (agriculture, environnement).

Le code des instructions (le code génétique) est constitué d'un alphabet chimique à 4 signes, les nucléotides (ou bases), qu'on symbolise par les lettres A, T, G et C. Une molécule d'ADN est constituée de l'enchaînement de millions de ces signes élémentaires tel un collier de perles à 4 couleurs. C'est cette forme d'enchaînement qui permet le stockage de l'information biologique, de même que la succession des octets magnétiques permet le stockage d'informations dans un ordinateur. En d'autres termes, l'ADN est la mémoire du vivant. Pour connaître les instructions que renferme une molécule d'ADN, il faut lire la succession des signes de l'enchaînement (des couleurs des perles le long du collier). C'est cette lecture qu'on appelle séquençage et qu'on sait pratiquer à petite échelle (quelques milliers de bases ou lettres) depuis les années 70, et à grande échelle (quelques millions de bases) depuis le milieu des années 90.

Depuis qu'il a appris à lire la séquence de l'ADN au cours des années 70, l'homme rêve de connaître son propre génome, même s'il n'est pas encore capable de connaître le sens de toutes les instructions contenues dans ce génome. Ce rêve est en passe de devenir une réalité : il y a quelques années a été lancé un gigantesque programme international destiné à séquencer le génome humain dans son intégralité soit 3 milliards de bases. Toute une série de retombées résultant de l'interprétation et de l'exploitation de ces données sont attendues pour les décennies à venir. Nous passerons en revue les plus importantes sur les plans scientifique, médical et des applications, sans oublier que les retombées scientifiques seront elles-mêmes à l'origine de la très grande majorité des nouvelles applications.

Les retombées scientifiques

Au niveau le plus simple se situe l'inventaire des instructions, puis la recherche d'une signification biologique à chacune des instructions qu'on désigne aujourd'hui sous le nom d'approches post-génomiques. Il est en outre quelque peu artificiel de séparer l'inventaire de l'interprétation, car l'inventaire lui-même peut déjà faire appel à un certain degré d'interprétation. Inventaire et interprétation doivent aussi s'appuyer sur les travaux portant sur les autres génomes, notamment les petits génomes, qui ont ouvert la voie et qui continuent à servir de systèmes pilotes au niveau de l'interprétation des données et dans les autres études post-génomiques. L'interprétation qui reste un défi considérable, même pour les petits génomes occupera sans doute une majorité de biologistes pendant plusieurs décennies.

L'inventaire des instructions

Un des premiers objectifs de l'interprétation consistera donc à procéder à un inventaire des gènes aussi complet que possible. Cet inventaire repose sur des comparaisons (notamment à des gènes déjà connus d'autres génomes) et des prédictions faites à l'aide d'autres programmes informatiques. Du fait de l'énormité du volume des données qui va être disponible, cette analyse informatique représente un défi sans précédent en biologie. L'analyse sera encore compliquée par le fait que la masse de données ne fera que s'accroître au fil des ans et qu'elle devra être constamment révisée pour prendre en compte les connaissances additionnelles. De plus, les résultats de traitements qu'effectuent les programmes existants doivent être examinés par des yeux experts, et ces experts font actuellement grandement défaut dans la plupart des pays, aussi bien dans le secteur académique que dans les entreprises privées. A partir de ces données analysées, on pourra dresser un inventaire de l'ensemble des gènes d'un individu.

L'interprétation des instructions

Dans la mesure où l'analyse informatique donne parfois des informations sur la fonction, l'interprétation commence dès ce stade. Si pour une fraction des instructions, la fonction peut être déduite des analyses informatiques, pour de nombreux autres, seules des démarches expérimentales additionnelles permettront de préciser la nature des instructions. En outre, même lorsque les fonctions des gènes peuvent être prédites par des programmes informatiques, il importe de valider ces prédictions par des expériences. Enfin, on s'aperçoit aussi que l'interprétation par comparaison passe souvent par l'acquisition de données de séquences additionnelles.

D'une manière générale, le fonctionnement d'une cellule et d'un organisme multicellulaire reste en grande partie méconnu. Même si la compréhension de chacune des instructions (gènes) d'un organisme ne donne pas une image complète du phénomène vital, elle représente un énorme pas en avant par rapport à l'état de connaissances actuelles. L'étude du rôle de chaque gène répertorié va donc devenir un des objectifs centraux de la biologie dans les décennies à venir. Une telle étude est déjà en cours pour les nombreux gènes de fonction inconnue des génomes dont la séquence complète est disponible.

L'instruction biologique contenue dans un gène s'exprime sous forme d'une autre molécule, une protéine synthétisée par conversion du code d'ADN par la machinerie cellulaire de synthèse protéique. C'est la protéine qui effectue au niveau cellulaire l'instruction contenue dans le gène. Connaître la fonction d'un gène, c'est donc connaître la fonction de la protéine. Pour comprendre l'information, deux voies majeures sont empruntées, l'une consiste à étudier les protéines nouvellement identifiées en général dans des systèmes *in vitro*, et l'autre, à observer sur l'organisme (animal, plante, micro-organisme) les conditions naturelles d'expression d'un gène ainsi que l'effet des modifications de ce gène. En raison de l'avalanche de gènes nouvellement identifiés, on voit se créer une nouvelle série de goulots d'étranglements dans les démarches expérimentales au niveau des disciplines classiques de la biologie : biologie structurale, biochimie, biologie cellulaire, physiologie, biologie et génétique moléculaires, etc. Toutes ces disciplines qui *grosso modo* absorbaient les études sur les nouvelles protéines (et leurs gènes) au fur et à mesure de leur découverte sont déjà submergées par le déferlement de dizaines de milliers de nouveaux gènes issus des programmes de séquençage systématique des génomes. Cette situation va s'intensifier dans les années à venir.

A côté du défi informatique évoqué ci-dessus, il s'en profile donc un autre qui met en jeu des séries de dizaines de milliers d'expériences. On sait aussi qu'une expérimentation systématique destinée à une catégorie d'observations ne pourra s'égarer vers l'inattendu. Or, l'inattendu et l'imprévu sont des sources majeures de nouvelles connaissances, et les observations méticuleuses faites à l'échelle d'un gène unique ou de son produit, qui ne rentrent pas dans le moule de la grande échelle, ne peuvent être envisagées à l'échelle d'un génome. Il s'agit là clairement d'un appauvrissement du processus de recherche et d'une dérive quantitative, malheureusement inéluctables pour des raisons de coûts.

Image globale

Aujourd'hui on dispose aussi de méthodes nouvelles qui permettent de connaître simultanément le niveau d'activité de chaque gène dans un tissu donné. Ces méthodes reposent notamment sur l'utilisation des données de séquence et sur l'inventaire des gènes. Ces méthodes peuvent être appliquées à comparer les tissus d'un organisme les uns aux autres, ou comparer les différents états physiologiques d'un même tissu, ou observer l'effet d'une drogue sur l'expression de l'ensemble des gènes du tissu et ainsi de suite. De nouveaux programmes sont actuellement lancés dans de nombreux pays pour obtenir une image globale de l'expression des gènes de tissus cancéreux, une carte d'identité des tumeurs. Ces programmes permettront d'affiner les diagnostics des cancers, de distinguer entre elles des tumeurs aujourd'hui considérées comme identiques, et donc, de mieux adapter les traitements, peut-être d'identifier de nouvelles protéines cibles pour l'action d'agents anti-tumoraux.

Retombées médicales dans les pathologies rares

Un grand nombre de maladies humaines ont une composante génétique. L'influence de cette composante sur la maladie est variable. Pour de nombreuses maladies rares, une altération (mutation) dans un seul gène se manifestera en général par l'apparition d'une série de signes caractéristiques de la maladie, alors que pour la plupart des maladies communes telles que le diabète, l'hypertension, les maladies neuro-psychiatriques, etc., l'effet des variations des gènes est modulé par une influence exercée par le reste du génome et par le milieu environnant. C'est pourquoi on distingue d'une part les maladies purement génétiques rares, encore appelées mendéliennes ou monogéniques, dont l'apparition peut être prédite dès que l'on connaît le gène responsable et, d'autre part, les maladies communes, dont l'origine est multifactorielle et pour lesquelles la présence d'un facteur de prédisposition chez un individu n'entraîne pas nécessairement l'apparition de la maladie.

Un terrain de chasse déjà fréquenté

Les premiers gènes responsables de maladies génétiques *stricto sensu* ont commencé à être isolés vers le milieu des années 80, alors que nos connaissances sur le génome étaient très parcellaires. C'est à cette époque, suite à ces premiers succès encourageants, qu'on a réalisé qu'une connaissance de l'ensemble du génome faciliterait considérablement l'identification des gènes à l'origine des maladies génétiques. De ces considérations est issu le programme génome. Une première étape au début des années 90 a consisté à faire des cartes utiles pour repérer sur le génome les emplacements des gènes morbides se transmettant dans certaines familles. Les équipes françaises se sont particulièrement illustrées pendant cette première phase du programme génome. La carte génétique élaborée à Généthon permet pratiquement de localiser un gène morbide dans un intervalle représentant moins de 0,1 % du génome. Mais

même dans des intervalles aussi petits, la phase d'identification peut encore prendre un temps considérable et coûter des efforts énormes aux équipes engagées dans la chasse aux gènes.

De la chasse à la traque systématique

Alors que dans le passé, on recourait à tout un arsenal de techniques fastidieuses et délicates pour rechercher les gènes, aujourd'hui les techniques de séquençage sont devenues suffisamment puissantes pour que cette approche constitue déjà la manière la plus efficace et la plus sûre pour identifier ces gènes. En particulier, la connaissance de la séquence permet un choix systématique et raisonné de gènes candidats sur lesquels seront ensuite recherchées les mutations. Il est donc certain que le programme de séquençage complet du génome aura un impact majeur dans la recherche de ces gènes responsables de maladies monogéniques. On doit donc s'attendre à une importante accélération dans ce domaine, et la plupart des gènes de maladies mendéliennes devraient être identifiés dans les 3 à 5 ans à venir.

Un nouvel éclairage de la physiopathologie

Même si les maladies génétiques mendéliennes sont rares, et si l'impact en santé publique sera mineur, la découverte de ces gènes représente une étape essentielle dans la compréhension de la fonction des gènes du génome humain et des processus physiologiques dans leur ensemble. Les progrès ne seront sans doute pas très rapides (voir ci-dessous), mais ceci devrait, à terme, déboucher sur de nouvelles pistes pour la thérapeutique pharmacologique ou autre, à côté de la voie de la thérapie génique qui reste encore balbutiante. Il existe en particulier des formes mendéliennes de certaines maladies communes. On connaît ainsi certains types de diabètes, ou certaines formes de la maladie d'Alzheimer strictement génétiques. La découverte des gènes responsables dans ces formes mendéliennes peut amener à mettre en évidence des aspects essentiels de la physiopathologie de ces affections, et donc, éventuellement, permettre des avancées dans la thérapie des formes les plus fréquentes non strictement génétiques.

Intérêt diagnostic

Le développement d'outils de diagnostic moléculaire est une des premières conséquences de la découverte d'un gène responsable d'une maladie génétique. Comme ces pathologies sont rares, il n'est pas question d'utiliser ce type de diagnostic de manière systématique. Cependant, dans certaines populations (populations ayant vécu en isolement pour des raisons géographiques ou culturelles) la fréquence d'un gène particulier peut être très forte et atteindre jusqu'à plusieurs pourcents. Dans ces populations particulières, il pourra être procédé à un diagnostic systématique. Ce diagnostic peut être anténatal pour des pathologies graves et la famille pourra faire le choix d'une interruption de grossesse. Mais le diagnostic d'ADN permet aussi de confirmer ou d'infirmer un diagnostic clinique ou de prédire la survenue d'une maladie se déclarant tardivement au cours de la vie. Ainsi la découverte du gène de la maladie périodique (une maladie se manifestant notamment par de fortes douleurs abdominales) donne au praticien un moyen sûr de distinguer cette pathologie d'autres affections présentant des signes similaires. Le diagnostic précoce de cette même pathologie permet aussi la mise en route d'un traitement médicamenteux (colchicine) qui évite, s'il est entrepris assez tôt, toute une série de complications graves pouvant aboutir au décès du malade. Des exemples de ce type se multiplieront dans les années à venir.

Une mutation particulière dans le facteur de coagulation V (Facteur V Leiden) est fréquente dans la population européenne et Nord-américaine. Les porteurs présentent un risque accru de

thrombose. Mais ce risque accru est encore fortement augmenté chez les femmes à la fois porteuses de cette mutation et prenant des contraceptifs oraux. Il semble même que chez ces sujets le risque soit plus grand avec les contraceptifs de troisième génération. On voit par cet exemple (il en existe d'autres) une possibilité de ciblage médicamenteux orienté par la pharmacogénétique et qui s'étendra au fur et à mesure que progresseront nos connaissances dans le domaine de la susceptibilité génétique aux médicaments.

Retombées médicales dans les pathologies communes

Du mono- au multigénique

Ce qui a été dit ici pour les maladies monogéniques rares ne peut directement s'extrapoler aux maladies communes à étiologie complexe. En particulier en raison de ce comportement génétique plus ou moins fugace, mentionné plus haut, il est difficile de localiser sur le génome les gènes de prédisposition aux maladies communes par les approches et les outils utilisés pour les maladies monogéniques. Cependant, lorsque les facteurs de prédisposition commenceront à être identifiés, les mêmes applications diagnostiques et thérapeutiques pourront être envisagées.

D'autres armes pour une autre chasse

Alors que de nombreux gènes de maladies génétiques ont déjà été identifiés, seuls quelques très rares facteurs de prédisposition à des maladies communes ont été découverts. Le fait que cette recherche n'ait que faiblement progressé malgré les efforts considérables déjà engagés souligne les difficultés inhérentes à ces travaux. D'autres stratégies sont à mettre en œuvre ici. Elles n'ont pas encore apporté de réponses car les outils requis ne sont pas encore véritablement disponibles. Ces outils reposent sur l'application de quelques principes simples qui sont résumés ci-dessous.

Tous parents tous différents

Les génomes des individus d'une même espèce sont globalement identiques. Il semble même qu'au sein de l'espèce humaine il y ait moins de variations que chez la plupart des autres mammifères. Ceci résulte sans doute d'une expansion assez récente d'*Homo sapiens sapiens* à partir d'un petit groupe d'individus fondateurs. Malgré cette forte homogénéité au sein de notre espèce, il existe de petites différences entre individus, même au sein d'une famille. Quand on compare les séquences de deux génomes non apparentés, on rencontre une variation environ toutes les 1.000 bases. Ce sont ces petites différences qui font que nous ne nous ressemblons pas tous comme des frères jumeaux, qui ont eux des génomes strictement identiques. Alors que la grande majorité de ces différences n'ont aucun effet sur le fonctionnement de notre patrimoine génétique, quelques-unes peuvent prédisposer à l'apparition de pathologies communes dans notre espèce. Un des objectifs majeurs des années à venir va consister à rechercher, au niveau des génomes, les différences génétiques qui peuvent être à l'origine de ces prédispositions.

Ces différences génétiques désignées sous le terme de SNP (single nucleotide polymorphism) peuvent se présenter à tous les niveaux génomiques, à l'extérieur ou à l'intérieur de gènes, dans la partie codante ou non codante. Si on imagine facilement que des SNP situés dans les parties codantes peuvent altérer une instruction génique, il faut aussi considérer que celles qui se produisent en dehors des régions codantes peuvent avoir un effet sur l'instruction. Une

variation de séquence à l'intérieur d'un gène ou dans son voisinage peut provoquer une augmentation ou une diminution de l'expression de ce gène. Des variations dans des introns (partie non codantes) peuvent provoquer des altérations dans le processus de maturation des instructions (qui a pour rôle d'éliminer la partie non codante). On est ainsi amené à considérer toute une série de possibilités de variations de séquences susceptibles d'avoir un effet biologique, mais, comme indiqué ci-dessus, la grande majorité de ces SNP sont sans effet fonctionnel.

Archéologie génomique

La démarche pour retrouver les facteurs génétiques de prédisposition repose sur le fait que les quelques rares variations de séquence qui sont à l'origine des prédispositions sont apparues par hasard dans un segment génomique particulier, lui même défini par un ensemble de SNP caractéristiques. Prenons l'exemple d'une variation prédisposant à l'asthme apparue chez un individu il y a 20.000 ans. Le phénomène de recombinaison génétique, qui fait que nous ne transmettons pas en bloc le génome d'un de nos parents mais un génome composite constitué de segments provenant soit d'un des parents soit de l'autre, va réduire, au fil des générations, la taille du segment d'origine contenant le variant de prédisposition. Après de nombreuses générations, on ne retrouvera que les SNP qui sont dans le voisinage immédiat du variant de prédisposition. En analysant une population d'individus asthmatiques, on pourra ainsi retrouver des individus qui descendent de celui chez qui s'est produit cette variation prédisposant à l'asthme. En analysant l'ensemble des SNP transmis par l'individu fondateur on pourra même retrouver lequel correspond à la variation de prédisposition. En effet, les SNP voisins, sans effet biologique pourront se retrouver dans la population générale, alors que la variation prédisposant à l'asthme sera trouvée de manière très préférentielle chez des asthmatiques. Comme d'autres variations pouvant survenir dans d'autres gènes peuvent avoir le même effet, toute la population d'asthmatiques étudiée ne sera pas descendante de cet ancêtre fondateur. Ceci va donc compliquer la recherche.

Comme le segment original dans lequel s'est produit la modification de prédisposition sera de taille très réduite après de nombreuses générations, il sera nécessaire de cribler le génome pour un très grand nombre de SNP. Pour pouvoir réaliser une telle analyse, il est indispensable de disposer, à côté de la séquence de référence du génome humain, d'une très grande collection de SNP représentative des différentes ethnies constituant la population de la planète. Devant l'ampleur de cette tâche, dix des plus grands groupes pharmaceutiques mondiaux ont décidé d'unir leurs efforts pour la constitution d'une vaste collection de SNP. Les données de cette collection resteront accessibles publiquement et pourront être utilisées par des groupes publics ou privés pour la recherche de facteurs génétiques de prédisposition.

La recherche des variations génomiques pouvant être la cause de prédispositions génétiques va donc s'intensifier dans les années à venir. Les stratégies ont été élaborées. Elles requièrent de très nombreuses analyses de SNP sur de très nombreux individus, aussi bien des populations témoins, que des cohortes de malades. Des techniques pour réaliser de très grands nombres d'analyses en parallèle sont encore en cours de développement ainsi que l'établissement de collections d'ADN provenant de cohortes de malades et d'individus témoins. Ces conditions (collections de SNP, d'ADN, techniques d'analyses massives, méthodes d'analyse statistiques) ne sont pas encore remplies. Mais d'important progrès ont été réalisés et l'établissement de la séquence complète du génome servira d'assise à une carte optimisée de SNP. On peut penser que d'ici deux à trois ans l'ensemble des outils sera disponible.

Une médecine sur mesure

A partir de là, il deviendra possible de réaliser les analyses génétiques en vue de retrouver des gènes de susceptibilité. Ces recherches seront longues et les résultats s'échelonneront sur de nombreuses années. Les premiers résultats peuvent néanmoins être obtenus dans les années à venir et serviront d'encouragement à poursuivre dans cette voie. Mais une fois ces gènes identifiés, il importera encore de transformer ces découvertes en progrès tangibles pour la santé humaine. Comme dans le cas des maladies monogéniques, deux applications sont envisageables, l'une dans le domaine du diagnostic et l'autre de la thérapie. On a souvent tendance à penser que la découverte d'un gène va inéluctablement conduire à une parfaite compréhension du mécanisme pathologique et donc aux moyens de maîtriser l'apparition et l'évolution de la maladie. Or, la réalité est bien différente. Le gène de prédisposition le mieux connu à ce jour est celui qui code pour l'Apoliprotéine E dont le variant E4 prédispose de manière très significative à la maladie d'Alzheimer et, 8 ans après la découverte de l'implication de ce gène, nous ne savons toujours pas comment le variant E4 peut induire l'apparition de la maladie. Mais si les progrès thérapeutiques seront peut être les plus tardifs, le diagnostic sera, comme souvent, le premier à bénéficier des nouvelles avancées dans l'identification des gènes.

La démarche appliquée à la recherche de facteurs génétiques de prédisposition peut aussi s'étendre à la recherche de susceptibilités à des traitements médicamenteux. Celles-ci ont en effet une composante génétique parfois majeure (cf. ci-dessus). Il deviendra ainsi possible d'administrer le médicament le mieux adapté pour une pathologie définie à un patient donné. Ceci permettra en outre de détecter des individus à risques pour certains traitements médicamenteux, tout en évitant d'en bannir l'utilisation pour l'ensemble de la population.

Des changements fondamentaux dans la pratique de la médecine, qui toucheront aussi bien la prévention, le diagnostic et la thérapie de maladies communes, sont à attendre dans les 10 à 20 premières années du prochain millénaire. Mais ceci a aussi des conséquences au niveau de nos sociétés. Les progrès du diagnostic en particulier déboucheront sur le criblage systématique de la population pour rechercher des individus présentant des risques potentiels pour leur santé. Cette information sera notamment critique pour les médecins et pour une politique éclairée de dépenses de santé. Il paraît presque inéluctable que le génome de chaque individu ne soit analysé pour un ensemble de SNPs d'importance critique. Comme cette information peut aussi être utilisée au détriment des individus, il sera nécessaire de garantir une confidentialité

Comment établit-on la séquence du génome humain

La stratégie du programme public de séquençage du génome humain repose sur le séquençage de grands fragments préalablement ordonnés (c'est-à-dire dont l'enchaînement original tel qu'il existe sur les chromosomes a été reconstitué). Mais la lecture des séquences se fait essentiellement sur de petits fragments d'ADN (1000 à 3000 bases). La première étape du séquençage consiste donc à fragmenter les grandes molécules d'ADN en morceaux plus petits qui seront ensuite séquencés. La fragmentation va introduire des cassures au hasard. Comme on va fragmenter un grand nombre de copies de la grande molécule de départ, les séquences des petits fragments qu'on aura lues pourront être chevauchantes. On pourra à partir de ces portions chevauchantes reconstituer la séquence du grand fragment. En fait on n'arrive pas à la reconstituer dans son intégralité à partir de ce séquençage "aléatoire". Même en faisant la

lecture d'un grand nombre de petits fragments c'est-à-dire en faisant des lectures aléatoires fortement redondantes 6, voire 10 fois, il est impossible de reconstituer complètement les molécules : il reste de petits trous. Pour combler ces derniers, un important travail de finition est incontournable. Celui-ci se fait de manière ciblée à un coût presque aussi élevé que le séquençage aléatoire qui donne 95% de la séquence.

Après des débuts à allure modérée vers le milieu des années 90, le programme public international est à présent dans une phase de très forte accélération. Afin que les utilisateurs (scientifiques et industriels) puissent disposer au plus vite de ces données, le programme public a décidé de réaliser des produits intermédiaires, sous forme d'une séquence incomplète de chacun des grands fragments, qui constituera une première ébauche disponible au cours du printemps 2000. Une deuxième version avant le travail de finition sera prête vers le début de 2001 alors que la version finale reste prévue pour fin de 2003.

Comme les applications de la séquence du génome humain sont potentiellement nombreuses et source de profit, elles suscitent l'intérêt grandissant des investisseurs privés. C'est pourquoi en parallèle au programme public, un programme de séquençage privé réalisé par une entreprise du nom de Celera s'est également mis en place. Comme les applications passent avant tout par une interprétation de la séquence du génome, cette entreprise envisage aussi de vendre, sous forme d'accès à une base de données, une interprétation de cette séquence. L'interprétation repose souvent sur l'utilisation de données de séquence supplémentaires. Il est peut donc être crucial pour Celera, afin d'attirer la clientèle, de constituer pour un temps limité des ensembles de données qui n'existent pas dans le domaine public. C'est pourquoi elle s'est dotée d'une capacité de séquençage bien supérieure à celle des plus grands centres du domaine public.

La stratégie retenue par cette entreprise est fondamentalement différente de celle du projet public.
Elle a consisté initialement à parier qu'on peut, non seulement reconstituer, un par un, l'enchaînement de grands fragments de génome humain, de l'ordre de la centaine de milliers ou du million, mais qu'on peut reconstituer des fragments aussi importants à partir du séquençage aléatoire de l'ensemble du génome. Cette stratégie, qui a été accueillie avec beaucoup de scepticisme, a été testée avec un succès mitigé sur la drosophile. Des fragments ont pu être reconstitués, mais le génome reste en 5000 morceaux. Ceci indique clairement que cette stratégie ne peut réussir sur le génome humain.
Mais avec l'accélération du projet public pour réaliser une première ébauche, Celera pourra simplement ajouter ses propres données à celles du projet public et disposer ainsi pour son compte d'un ensemble plus intéressant.

Il faut cependant garder à l'esprit que les objectifs de Celera et ceux du projet public sont fondamentalement différents. Dans le premier cas, il s'agit d'un objectif à court terme, visant à faire un produit incomplet (possédant un intérêt commercialement intéressant pour une période de temps limitée) et, dans le deuxième cas, de constituer un outil accessible à tous et d'une utilité bien plus générale.

Conclusion

Nous sommes aujourd'hui devant une nouvelle ère qui aura des répercussions sur la vie de chacun aussi importantes qu'il y a 10.000 ans, lorsque l'agriculture s'est progressivement répandue sur l'ensemble de la planète. Nous ne sommes sans doute pas plus qu'alors capables

de mesurer tous les risques de ces changements qui résulteront d'une plus grande maîtrise du vivant, mais au moins avons-nous l'intuition que ce qui se prépare n'est pas sans conséquences. Trop a déjà été dit sur les OGM, leurs dangers potentiels, beaucoup moins sur une évaluation rigoureuse des risques telle qu'elle pourrait être admise par l'ensemble des protagonistes.

Dans le domaine de la santé, nous sommes une fois de plus plongés dans le paradoxe du progrès de la connaissance qui nous fait gagner un nouvel espace de liberté et en même temps nous enseigne que notre destinée individuelle est un peu plus déterminée. Mais y a-t-il une réponse à ce dilemme et aux interrogations que suscitent les pratiques médicales issues de la génétique qui vont progressivement s'établir ? Nous serons en mesure de prédire beaucoup. Avons-nous le devoir de le faire ? Pouvons-nous refuser de savoir, dans des sociétés où la santé est à la fois l'affaire de tous et de chacun ? La seule évidence qui me semble s'imposer est celle de plus d'éducation. Le niveau de connaissances en biologie de nos concitoyens doit être suffisant pour que les choix qui seront à faire puissent être faits en connaissance de cause. Cela est moins facile qu'il n'y paraît, car la biologie n'est pas la science des certitudes absolues.

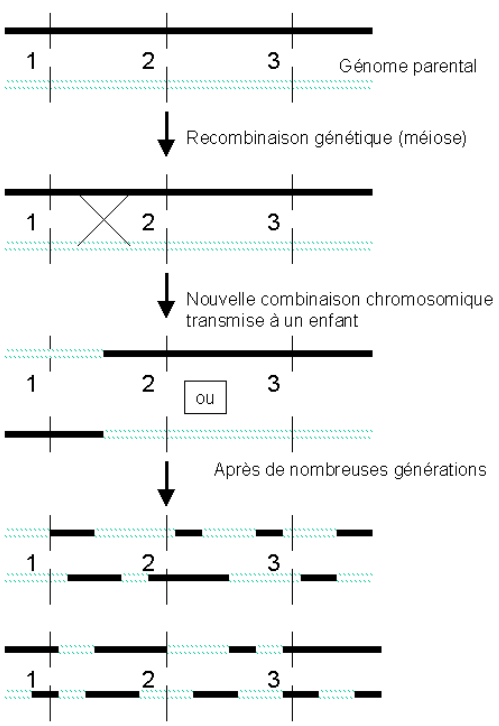
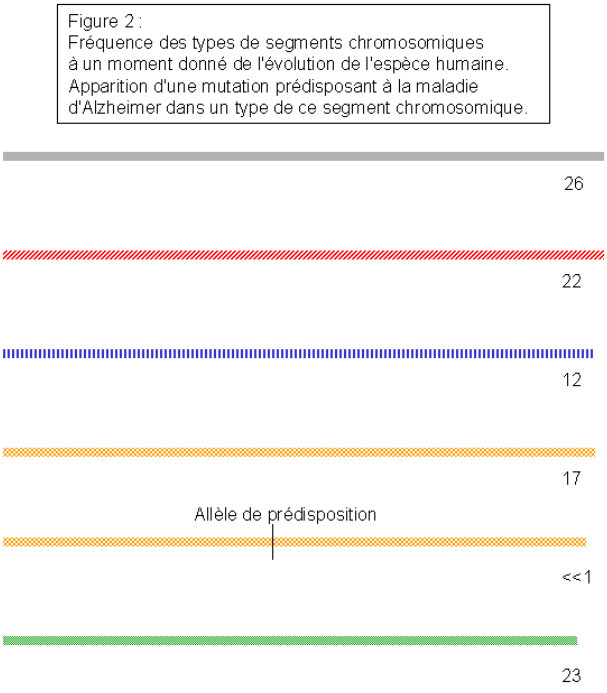


Figure 1 : Recombinaisons génétiques



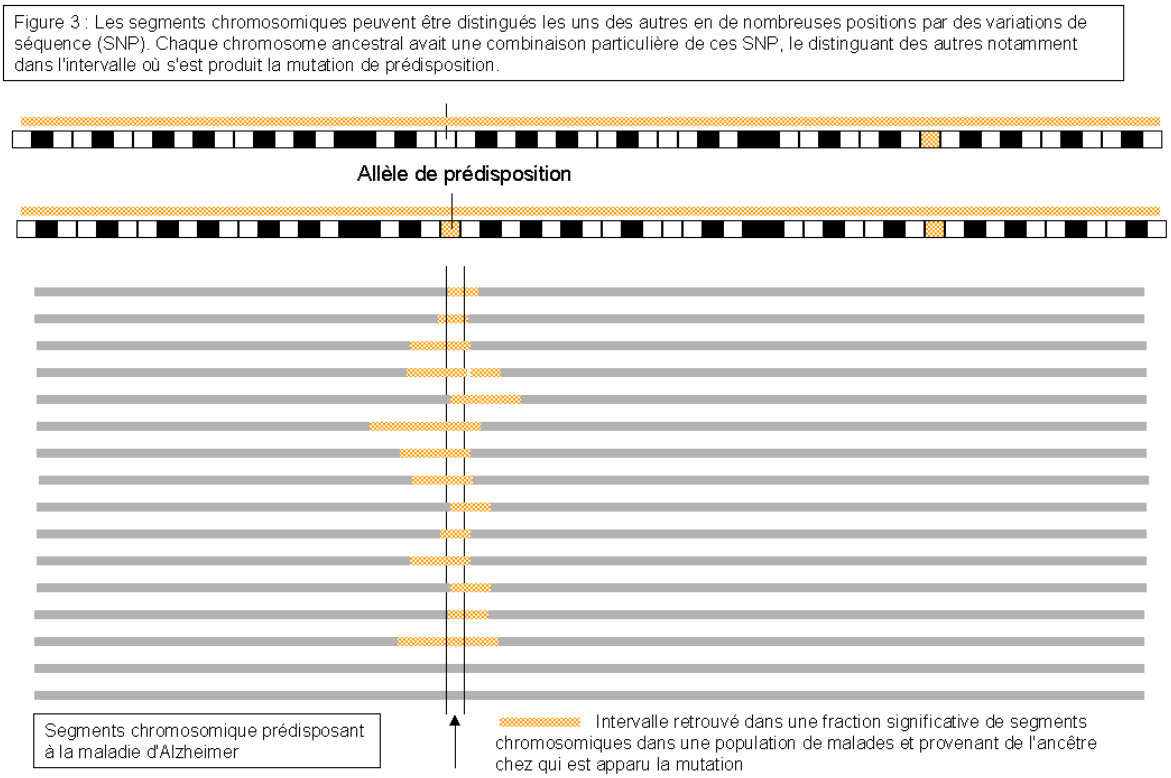


Figure 2 : Fréquence des types de segments chromosomiques à un moment donné de l'évolution de l'espèce humaine.
Apparition d'une mutation prédisposant à la maladie d'Alzheimer dans un type de ce segment chromosomique.

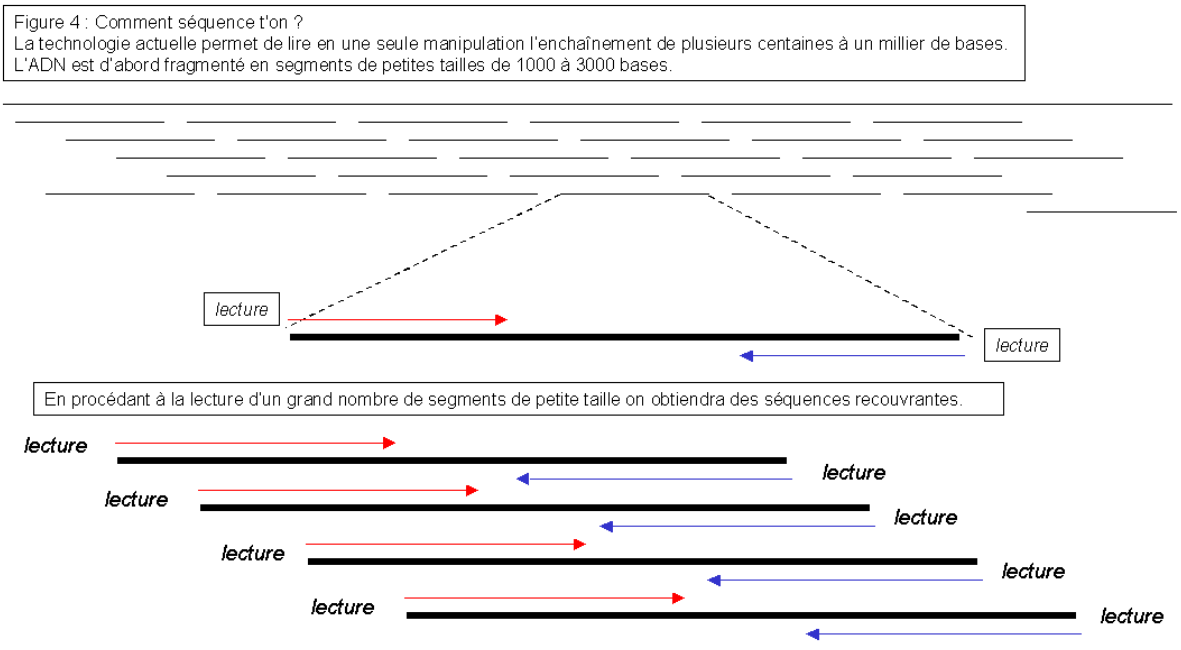


Figure 3 : Segments chromosomiques prédisposant à la maladie d'Alzheimer.

Les segments chromosomiques peuvent être distingués les uns des autres en de nombreuses positions par des variations de séquence (SNP). Chaque chromosome ancestral avait une combinaison particulière de ces SNP, le distinguant des autres notamment dans l'intervalle où s'est produit la mutation de prédisposition.
Un intervalle est retrouvé dans une fraction significative de segments chromosomiques dans une population de malades. Il provient de l'ancêtre chez qui est apparu la mutation. Chez les individus non prédisposés; l'intervalle contient d'autres combinaisons des variations de séquences (SNP).

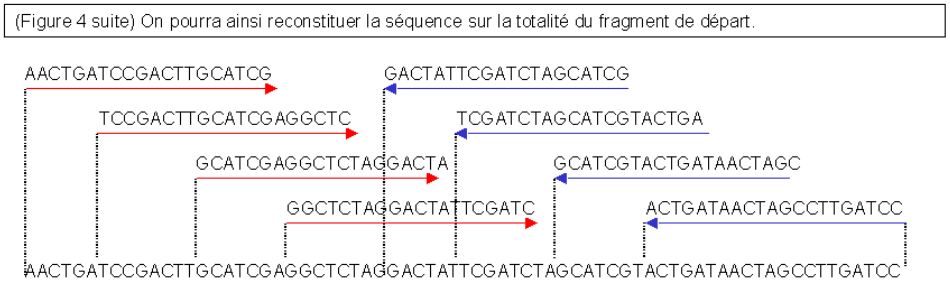


Figure 4 : Comment séquence-t-on ?
La technologie actuelle permet de lire en une seule manipulation l'enchaînement de plusieurs centaines à un millier de bases. L'ADN est d'abord fragmenté en segments de petites tailles de 1 000 à 3 000 bases. En procédant à la lecture d'un grand nombre de segments de petite taille on obtiendra des séquences recouvrantes. On pourra ainsi reconstituer la séquence sur la totalité du fragment de départ.

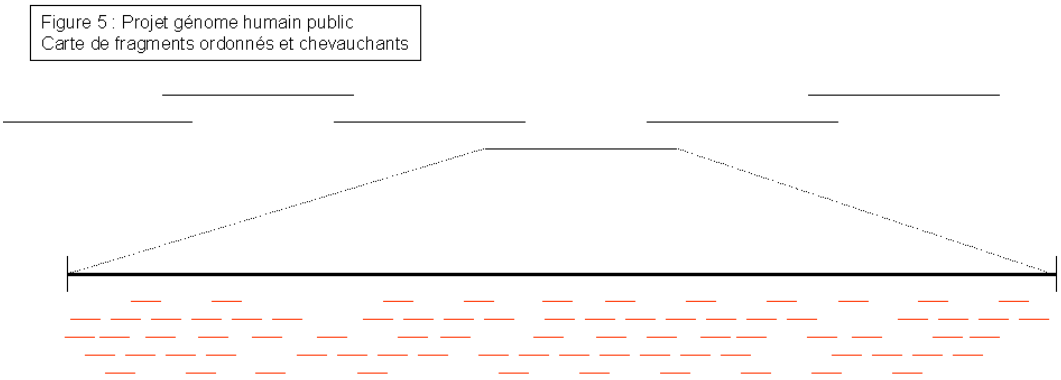


Figure 5 : Projet génome humain.
La première ébauche de la séquence publique pourra être combinée avec les données de la compagnie Celera.