PREMIERE PARTIE

I. Etude bibliographique des CNV

A. Présentation générale des CNV

1. Que sont les CNV ?

L'utilisation fructueuse de marqueurs génétiques tels que les **SNP**[1] (*Single Nucleotide Polymorphism*) ou les **microsatellites**, pour l'étude de maladies complexes humaines a incité à rechercher de nouveaux types de polymorphismes dans le génome humain, ainsi que dans le génome d'espèces modèles tels que la Souris, le Macaque, la Drosophile ou le Chien (Wain *et al.*, 2009). Il y a quelques années, les techniques fines de biologie moléculaire ont permis de mettre en évidence la présence de nombreuses variations structurales des génomes, qui n'avaient pas été suspectées jusqu'alors : les CNV (Redon *et al.*, 2006).

Les CNV sont des délétions et duplications d'une taille variant d'un kilobase à plusieurs mégabases. Ils touchent toutes les régions du génome. Ils peuvent inclure plusieurs gènes, ne toucher qu'un gène ou partie de gène ou encore être dépourvus de gènes ou d'éléments de régulation. Pour la plupart des gènes de notre génome nous héritons d'une copie paternelle et d'une copie maternelle, si bien que nous possédons deux copies de chaque gène dans le noyau de nos cellules diploïdes. Néanmoins, à cause des CNV, le nombre de copies de certains gènes ou segment d'ADN peut varier dans le génome. Par exemple, chez l'Homme, un ensemble de gènes codant pour des ß-défensine (*ß-defensin*) présente fréquemment des duplications, allant de deux à sept copies par cellule diploïde (Hollox *et al.*, 2003). Il en est de même pour le gène de l'amylase salivaire *AMY1* qui varie de deux à 15 copies dans la population américaine d'origine européenne (Perry *et al.*, 2006). Un autre exemple est l'allèle rhésus négatif qui est généralement causé par une délétion complète du gène *RHD*. Le génome diploïde d'un individu peut ainsi contenir deux, une ou aucune copies du gène *RHD*. Les individus n'ayant aucune copie du gène *RHD* n'expriment pas l'antigène D et sont rhésus négatif (Avent *et al.*, 1997 ; Wain *et al.*, 2009).

[1] Les termes en caractères gras sont expliqués dans le glossaire présenté en annexe.

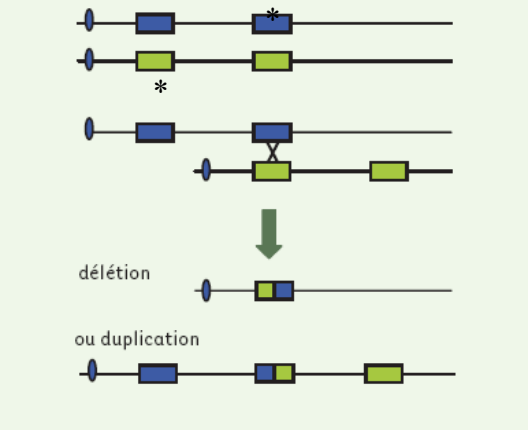
Les CNV sont présents dans la population générale humaine. Leur présence a aussi été mise en évidence chez d'autres mammifères ainsi que chez certains invertébrés tels que la Drosophile et le Poisson zèbre (Henrichsen *et al.*, 2009 A). Leur importante prévalence dans le génome (ils couvriraient 10% des autosomes de la Souris et 12% du génome humain) laisse envisager qu'ils pourraient avoir un impact fonctionnel important (Henrichsen *et al.*, 2009 B ; Redon *et al.* 2006 ; McCaroll *et al.*, 2008). C'est pourquoi ces variants ont suscité l'intérêt des chercheurs qui travaillent sur l'identification de facteurs génétiques impliqués dans le déterminisme des maladies complexes ou de caractères phénotypiques ; car, si certains CNV peuvent être anodins (du moins sans conséquence décelable sur le phénotype), l'implication de ce type de variant dans des maladies complexes est maintenant avéré. Chez l'Homme, les CNV ont été associés à diverses maladies complexes. Chez la Souris et le Chien aussi, les CNV ont été associé à des phénotypes spécifiques ou à une sensibilité accrue à certaines maladies (Lee *et al.*, 2001 ; Cahan *et al.*, 2009 ; Chen *et al.*, 2009 ; Nicholas *et al.*, 2009 ; Williams *et al.*, 2009).

2. Comment les CNV sont-ils formés ?

Les mécanismes de formation des CNV sont encore peu connus, mais de nombreuses études ont permis d'observer que les CNV étaient significativement plus présents dans les régions répétées du génome telles que les **duplications segmentaires**, les ***Long Interspersed Nuclear Element*** (LINE), ou encore les ***Long Terminal Repeat*** (LTR) (Iafrate *et al.*, 2004 ; Sharp *et al.*, 2005 ; Perry *et al.*, 2006 ; Redon *et al.*, 2006 ; Lee *et al.*, 2008 ; She *et al.*, 2008 ; Nicholas *et al.*, 2009). A partir de cette observation les chercheurs ont émis l'hypothèse que les CNV seraient principalement formés par recombinaison homologue non allélique, phénomène par lequel une recombinaison se produit entre deux régions dont les séquences nucléotidiques sont très similaires mais non alléliques (Gu *et al.*, 2008 ; Henrichsen *et al.*, 2009 A). C'est, par exemple, ce qui se produit dans les régions présentant des duplications segmentaires. Les duplications segmentaires sont des blocs d'ADN d'au moins un kilobase de longueur, dupliqués dans le génome et qui ont un degré remarquable d'identité de séquence (>90 %). Ces blocs peuvent être intrachromosomiques (situés sur le même chromosome), ou interchromosomiques (situés sur des chromosomes différents) (Turleau et Veckemans, 2005). Certaines régions du génome, plus particulièrement les régions péricentromériques et subtélomériques des chromosomes, sont enrichies en ce type de séquences (Bailey *et al.*, 2002). La forte homologie de séquence des duplications segmentaires en font un

substrat moléculaire de recombinaison homologue non-allélique. Cette recombinaison anormale entre des segments répétés spécifiques d'une région ou d'un chromosome entraîne la perte ou la duplication du segment génomique compris entre les deux duplicons (Turleau et Veckemans, 2005 ; Wain *et al.*, 2009) (Figure 1).

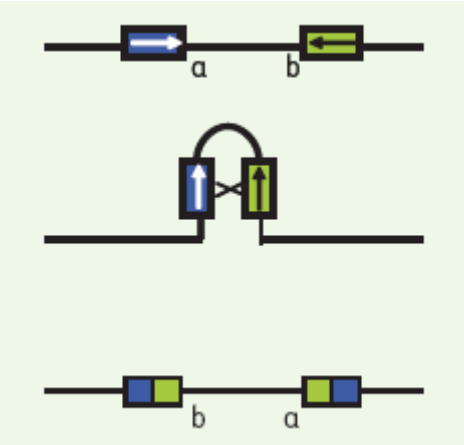
Figure 1. Recombinaison entre deux séquences homologues non alléliques de même orientation, entraînant une délétion ou une duplication (Turleau et Veckemans, 2005).



Les deux séquences homologues non alléliques sont représentées par les rectangles signalés par des étoiles. Ces deux séquences sont reconnues comme homologues par la machinerie cellulaire. Il se produit une recombinaison entre les deux brins d'ADN, produisant soit une délétion, soit une duplication.

Différents facteurs incluant la taille des répétitions, l'intervalle qui les séparent, leur degré d'homologie et leur orientation relative influencent la probabilité de mésappariement et le type de remaniement généré. D'une façon générale, plus les segments sont grands et plus leur degré d'homologie est élevé, plus la probabilité de survenue d'un échange anormal est grande (Turleau et Veckemans, 2005). Une recombinaison homologue non allélique entre deux duplications segmentaires intrachromosomiques de même orientation va entraîner une délétion ou une duplication du segment intermédiaire (Figure 1). Si les séquences sont en orientation opposée, ce même phénomène peut entraîner une inversion du segment intermédiaire (Turleau et Veckemans, 2005) (Figure 2).

Figure 2. Recombinaison homologue non allélique entre deux duplications segmentaires intrachromosomiques d'orientation opposée, entraînant une inversion du segment intermédiaire (Turleau et Veckemans, 2005).



Les deux duplications segmentaires sont représentées par les rectangles incluant une flèche. L'orientation du segment d'ADN est repérée par les lettres a et b. Les deux séquences homologues sont reconnues par la machinerie cellulaire et il se produit une recombinaison entre les deux brins d'ADN non allèles, produisant une inversion du segment d'ADN a-b.

Cependant, les CNV peuvent apparaître même en l'absence de régions répétées. Une perte ou un ajout de nucléotides peut survenir lors d'une réparation imprécise de l'ADN par exemple (Wain *et al.*, 2009).

Les CNV représentent un dynamisme du génome. L'évolution des CNV est constante. Chaque génération d'individus (homme ou animal) voit apparaître et disparaître des CNV (Wain *et al.*, 2009). Ce dynamisme dépend des régions du génome. Les régions répétées, par exemple, sont favorables aux remaniements, des CNV *de novo* y apparaissent fréquemment, contrairement à d'autres régions plus stables où l'apparition d'un variant est un événement rare. Dans les régions stables du génome, les CNV seraient généralement **mono-alléliques**, fixés dans la population et transmis de manière mendélienne (Wain *et al.*, 2009). Alors que les régions fréquemment remaniées présenteraient des CNV multi-alléliques résultant de mutations récurrentes.

Une synténie a été observée pour 22% des CNV entre le Chimpanzé et l'Homme et pour 25% des CNV entre le Macaque rhésus et l'Homme (Perry *et al.*, 2006; Lee *et al.*, 2008). Les CNV fréquents chez le macaque étaient souvent retrouvés chez l'humain (Lee *et al.*, 2008), ce qui suggérait que certains CNV étaient apparus chez un ancêtre commun et ont été conservés dans les deux espèces, mais aussi que certaines régions du génome (comme les régions répétées) étaient favorables à la formation des CNV, quelle que soit l'espèce (Henrichsen *et al.*, 2009 A).

Des CNV *de novo* peuvent se former dans les cellules, à des stades précoces du développement, mais aussi tout au long de la vie d'un individu. Si les cellules de la lignée germinale sont touchées par le CNV, celui-ci sera transmis. Il se pourrait que de nombreux individus soient mosaïques pour certains CNV, ce qui compliquerait l'analyse des CNV et l'étude de leurs effets (Wain *et al.*, 2009).

3. Cartographier les CNV/Comment détecte-t-on les CNV ?

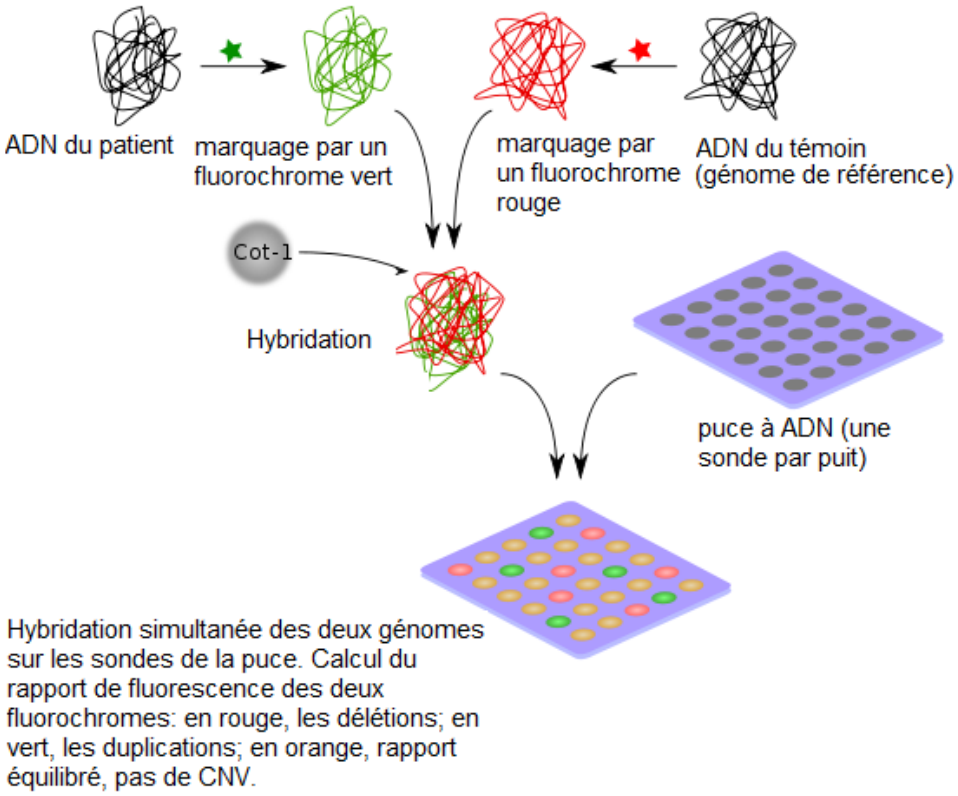
La détection des CNV s'effectue en deux temps : le passage de l'ADN des individus à tester sur des puces (CGH pour *Comparative Genomic Hybridization*, SNP pour *Single Nucleotide Polymorphism*, séquençage) puis l'analyse des données issues de ces puces à ADN, par des logiciels qui prédisent la localisation des CNV dans le génome.
La plupart des techniques employées aujourd'hui permettent de détecter les CNV sur l'ensemble du génome. Trois méthodes majeures sont utilisées pour détecter les CNV : l'hybridation génomique comparative (CGH), le génotypage de SNP et le séquençage.

a. Les méthodes de détection des CNV

L'hybridation génomique comparative (CGH)

La méthode de *CGH array* utilise un ensemble de sondes (molécules simple brin d'ADN) représentant l'ensemble du génome, qui sont ancrées sur une surface solide (*array*) ou puce. Cette méthode permet de comparer deux génomes : un témoin et un à tester. Le génome de référence est marqué avec une molécule fluorescente et le génome à tester avec une autre molécule fluorescente. Ces deux génomes sont hybridés simultanément sur les sondes présentes sur la puce. Pour chaque sonde, le rapport de fluorescence des deux fluorochromes est calculé, ce qui permet de détecter les délétions et duplications dans le génome à tester par rapport au génome de référence (Figure 3).

Figure 3. Principe de la puce d'hybridation génomique comparative



Cot-1 = ADN cot-1 : les ADN de référence et du patient sont hybridés en présence d'un ADN cot-1 humain, non marqué, riche en séquences répétées, afin de masquer ces séquences avant hybridation sur les sondes de la puce.

La technique d'hybridation génomique comparative a été la plus utilisée pour détecter les CNV jusqu'aujourd'hui (Wain *et al.*, 2009). L'ADN de référence utilisé dans ces études peut être celui d'un individu témoin auquel les génomes d'autres individus sont comparés. La technique de CGH est aussi couramment utilisée pour détecter les délétions et duplications des cellules tumorales d'un individu en comparaison aux cellules saines de ce même individu.

Suivant le nombre de sondes utilisées et leur répartition dans le génome, la résolution de cette technique varie considérablement. Les puces CGH de dernière génération sont capables de détecter des CNV de quelques dizaines de paires de bases seulement (Wain *et al.*, 2009).

Cette technique possède néanmoins une limite importante : les CNV sont détectés par rapport à un génome de référence qui peut aussi présenter des CNV, ce qui engendre une source d'erreur. Par exemple, une duplication de l'ADN de référence apparaitra comme une délétion dans le génome à tester. Chez la Souris, la plupart des études d'hybridation génomique comparative se font en référence au génome de la lignée C57BL/6J. Or, il a été montré que 16 à 33 mégabases (Mb)

d'ADN n'existaient pas dans cette lignée par rapport à d'autres lignées (25Mb en moyenne) (Graubert *et al.*, 2007). Ainsi, les variants présents dans les séquences d'ADN qui n'existent pas chez C57BL/6J ne sont pas détectés, ce qui entraîne une sous-estimation du nombre de CNV dans les études recensant les CNV dans le génome de plusieurs lignées de Souris.

Génotypage de polymorphismes d'une seule base (SNP)

Le génotypage de SNP (utilisé pour les études d'association, par exemple) permet de détecter les CNV. Le principe de cette technique est d'extraire l'ADN d'un patient, de fragmenter cet ADN, de le marquer avec un fluorochrome, et de l'hybrider sur les sondes, spécifiques des SNP à génotyper, d'une puce. Après lavage, la fluorescence de chaque puits de la puce est analysée. Cette méthode permet de déterminer les allèles présents d'un SNP donné, mais aussi de déterminer le nombre de copies d'un SNP dans le génome. La fluorescence émise, pour chaque SNP, est rapportée à la fluorescence moyenne, qui correspond à deux copies d'un SNP dans le génome. Si la fluorescence détectée est inférieure, c'est que le SNP est en moins de deux copies dans le génome, c'est-à-dire que le segment d'ADN qui porte le SNP est délété. Si la fluorescence détectée est supérieure à la moyenne, c'est que le SNP est en plus de deux copies dans le génome, c'est-à-dire que le segment d'ADN qui porte le SNP est dupliqué. C'est l'analyse de l'ensemble des SNP, par des logiciels adaptés, qui permet de mettre en évidence la présence de délétion et de duplications, caractérisées par une diminution ou une augmentation de fluorescence de plusieurs SNP contigus.
Des études d'association se basant sur les CNV ont été entreprises en utilisant des puces à ADN Affymetrix (Santa Clara, Etats-Unis) et Illumina (San Diego, Etats-Unis) ainsi que des puces hybrides, composées de sondes détectant les SNP et de sondes non-polymorphes, spécifiques des CNV, situées dans les régions remaniées. Ces études mesuraient le signal combiné des deux allèles à un SNP donné, l'intensité de ce signal était exprimée en logarithme du ratio entre l'intensité enregistrée et l'intensité attendue. L'intensité attendue du signal étant une intensité moyenne calculée pour un SNP présent en deux copies dans le génome, et non calculée par rapport à un échantillon de référence, comme c'est le cas dans la technique d'hybridation génomique comparative. Tout un panel de normalisations et un grand nombre d'algorithmes ont été utilisés pour analyser les données issues des puces (Wain *et al.*, 2009).

Les puces à ADN qui génotypent des SNP à haut débit (les puces actuelles permettent de détecter plus d'un million de SNP chez l'Homme, par exemple) permettent de détecter les CNV à une résolution de moins de 10 kb sur l'ensemble du génome. De telles puces se développent aussi chez les animaux, comme chez la Souris par exemple (Yang *et al.*, 2009).

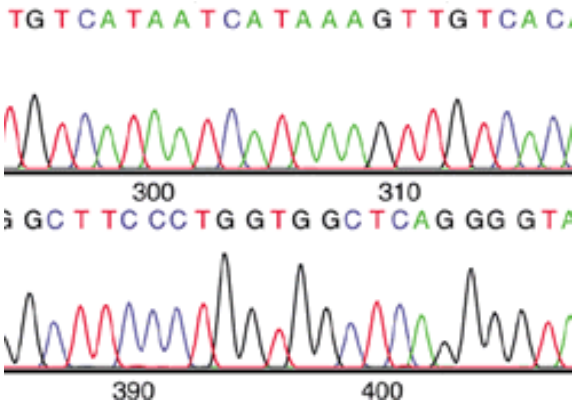
Cette méthode permet de cartographier les CNV sur un grand nombre d'individus en un temps réduit, et pour un coût raisonnable. Le génotypage de SNP à haut débit est moins coûteux que le séquençage (mais moins précis, car il ne permet pas de caractériser les points de cassure exacte des CNV).

De nouvelles puces à ADN dites hybrides incluent des SNP et des sondes non polymorphes spécifiques des CNV. Ces puces de dernière génération, qui existent chez l'Homme et chez la Souris, permettent une détection fine des variants (Wain *et al.*, 2009).

Le Séquençage

Le séquençage de l'ADN consiste à déterminer l'ordre d'enchaînement des nucléotides d'un fragment d'ADN donné. Actuellement, la plupart des séquençages d'ADN sont réalisés par la méthode de Sanger. Cette technique utilise la réaction de polymérisation de l'ADN à l'aide d'une ADN polymérase et de didésoxyribonucléotides (ddNTP) marqués par des fluorochromes différents. Les ddNTP, sont des terminateurs de chaîne : une fois incorporés dans le nouveau brin synthétisé, ils empêchent la poursuite de l'élongation. Cette terminaison se fait spécifiquement au niveau des nucléotides correspondant au didésoxyribonucléotide incorporé dans la réaction. Un séquenceur sépare les brins d'ADN synthétisés par électrophorèse et analyse la fluorescence émise par chaque ddNTP qui termine le fragment. Des logiciels permettent d'associer la longueur d'onde de la fluorescence émise à une base (A,T,G ou C) (Figure 4). On obtient ainsi la séquence de la portion d'ADN souhaitée, ou du génome entier (Staden, 1979).

Figure 4. Résultat du séquençage donné par un séquenceur automatique.



Chaque pic de fluorescence est associé à une base, suivant la longueur d'onde d'émission. Ici, la base C est représentée en bleu, G en noir, A en vert et T en rouge. L'ordre des bases constituant une séquence d'ADN est ainsi déterminé. Source : données personnelles.

Cette séquence peut ensuite être comparée à une séquence de référence pour mettre en évidence des modifications de séquence telles que les délétions ou les duplications.

Les approches de cartographie des CNV fondées sur du séquençage permettent de détecter et délimiter les CNV avec exactitude et mettent également en évidence les inversions et translocations. La méthode de séquençage permet de comparer la séquence d'un ADN à tester à la séquence d'un ADN de référence. Mais contrairement à la technique de *CGH* ou le génotypage de *SNP,* tous les variants peuvent être détectés même si le génome de référence présente des délétions ou duplications. Si le séquençage à haut débit est encore trop coûteux pour étudier de grandes cohortes, celui-ci a joué un rôle important dans la cartographie des CNV (Kidd *et al.*, 2009 ; Wain *et al.*, 2009).

L'intérêt majeur du séquençage est de pouvoir déterminer avec exactitude les points de cassure des CNV, ce qui permet de connaître quelle région ils touchent et d'avoir une idée plus précise de leur impact phénotypique.

b. Validation des CNV recensés

Une fois les CNV détectés par les algorithmes de prédiction suite à l'utilisation des techniques vues précédemment, il est nécessaire de valider ces variants par une méthode

indépendante. La PCR (*Polymerase Chain Reaction*) quantitative est la méthode la plus utilisée pour valider la présence des CNV. Le séquençage est également utilisé pour valider la présence d'un CNV, déterminer les points de cassure et pour étudier de manière détaillée les régions remaniées.

La précision de détection des CNV varie entre et au sein des différentes méthodes, car les puces à ADN utilisées n'ont pas toutes la même résolution et les algorithmes employés pour prédire les CNV n'ont pas tous la même puissance. C'est pourquoi la validation, par PCR quantitative ou par séquençage, des CNV prédits est nécessaire pour déterminer le taux de faux positifs de la méthode de détection des CNV utilisée.

La principale base de données recensant les CNV chez des individus de la population générale humaine est la *Database of Genomic Variants* (http://projects.tcag.ca/variation/).