

## DEUXIEME PARTIE

# II. Etude moléculaire des CNV dans le trouble bipolaire à début précoce

### A. Introduction

Le trouble bipolaire (TB), anciennement appelé maniaco-dépression, est un syndrome psychiatrique qui affecte environ 1% de la population générale. Ce trouble touche de manière égale hommes et femmes (Leboyer *et al.*, 2005).

Les patients atteints de TB alternent des épisodes dépressifs et des épisodes maniaques qui correspondent à des périodes d'exaltation, d'hyperactivité et d'euphorie. Ce trouble psychiatrique entraîne des difficultés socio-professionnelles (difficultés à trouver ou garder un travail par exemple) et socio-familiales (divorce, conflits familiaux). Le TB est également à l'origine d'une surmortalité par suicide (20% de décès chez les patients non traités) (Leboyer *et al.*, 2005).

La prévention de ce trouble passe par un diagnostic précoce et un traitement adapté pour éviter les rechutes. Néanmoins, le diagnostic n'est pas toujours aisé à poser et plusieurs années (huit en moyenne) s'écoulent avant la mise en place d'un traitement (Leboyer *et al.*, 2005).

Identifier les facteurs étio-pathogéniques de la maladie est indispensable pour améliorer la qualité et la précocité du diagnostic. De nombreuses investigations centrées sur la recherche de biomarqueurs, d'endophénotypes (caractères de vulnérabilité chez les apparentés non-atteints) et de gènes de vulnérabilité sont menées afin d'augmenter la précocité et l'efficacité de la prise en charge thérapeutique du patient et la prophylaxie des individus à risque.

Le TB est une maladie multifactorielle complexe dont les mécanismes patho-physiologiques restent encore inconnus, néanmoins, des études d'agrégation familiale, de jumeaux et d'adoption (Craddock et Ian, 2001) ont montré l'implication de nombreux facteurs génétiques et environnementaux.

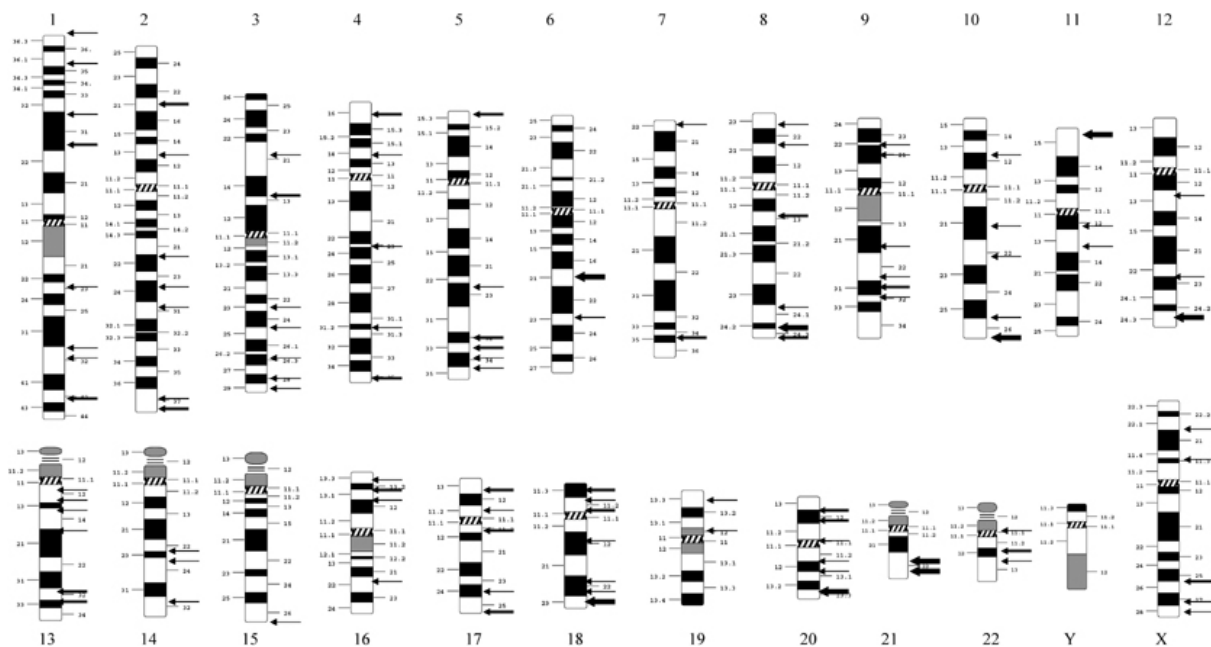
En effet, si la prévalence du TB dans la population générale se situe autour de 1%, celle-ci croît significativement dans les familles de sujets atteints, avec un risque 5 à 10 fois plus élevé chez les apparentés de premier degré (Craddock et Ian, 2001).

Les études de jumeaux permettent de mettre en évidence une contribution génétique à l'étiologie de la maladie. Celles-ci ont montré un taux de concordance chez les jumeaux monozygotes (79%) supérieur à celui des jumeaux dizygotes (19%), permettant d'estimer une héritabilité entre 60 et 80% (Bertelsen *et al.*, 1977).

Les études d'adoption présentent l'intérêt de séparer les facteurs génétiques et environnementaux au sein des familles. Le risque pour un parent d'être atteint de TB, sachant que leur enfant est atteint, est supérieur pour les parents biologiques (18%) par rapport aux parents adoptifs (7%) (Mendlewicz *et al.*, 1977).

De nombreuses études de liaison et d'association ont tenté d'identifier les gènes impliqués dans le TB. Une revue de Seretti et Mandelli en 2008 dresse un bilan de l'ensemble des régions du génome associées au TB par des études de liaisons et d'associations (Figure 5).

**Figure 5. Régions du génome associées au TB de l'Homme, par l'ensemble des études d'association et de liaison menées jusqu'à décembre 2007.**



Chaque région est indiquée par une flèche. La largeur des flèches est proportionnelle au nombre de fois qu'une région a été associée au TB (d'après Seretti et Mandelli, 2008).

En dépit des nombreuses régions chromosomiques identifiées au cours de ces études (Seretti et Mandelli, 2008 ; Etain *et al.*, 2006), aucune région, ni aucun gène, n'a pu être impliqué avec certitude. Une des raisons qui pourrait expliquer ces résultats discordants et leur manque de reproductibilité, est que les patients de ces études étaient recrutés sur le seul critère « patient atteint de TB ». Or, le TB présente un spectre clinique large. On peut supposer que cette hétérogénéité clinique est sous-tendue par une hétérogénéité sous-jacente des mécanismes moléculaires et des gènes impliqués.

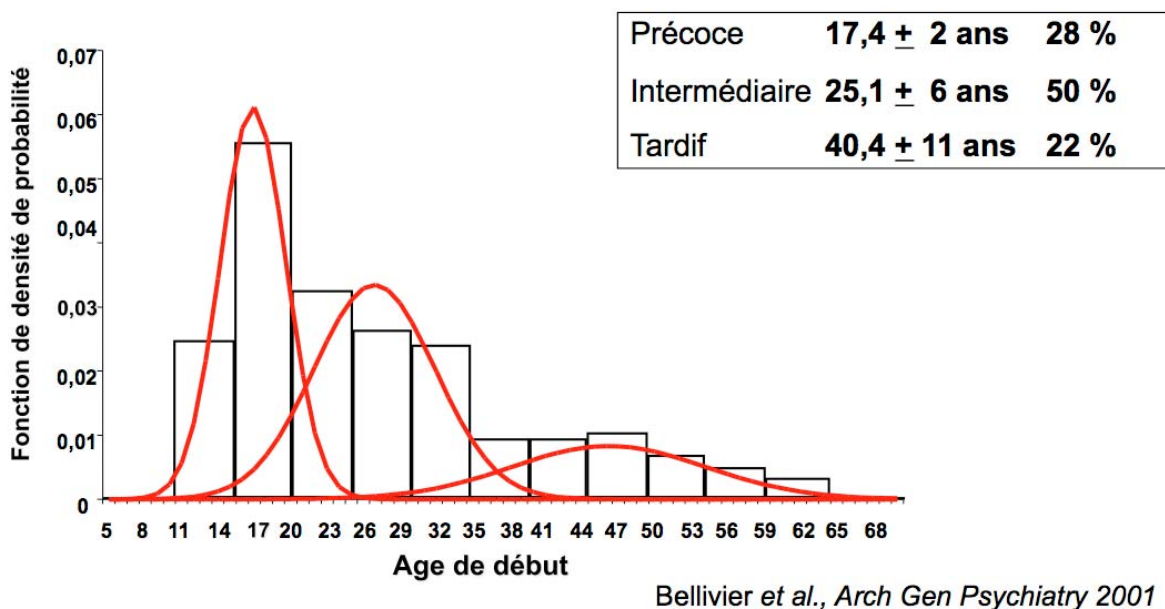
L'équipe 15 de l'Unité INSERM 955 « Psychiatrie et Génétique » de l'hôpital H.Mondor (Créteil) tente depuis plusieurs années de constituer des cohortes en collectant, chez les patients, un grand nombre de variables cliniques afin de définir des sous-groupes cliniques homogènes. Ils recherchent, dans ces cohortes, des symptômes candidats (Leboyer *et al.*, 1998), qui correspondent à des caractères phénotypiques transmis selon un mode génétique simple, présentant une agrégation familiale, et sous-tendus par un nombre limité de gènes.

Un début précoce de la maladie, des conduites suicidaires, la présence d'éléments psychotiques sont par exemple des symptômes candidats. Ainsi, la sélection de patients, suivant des symptômes candidats, permettrait d'identifier des familles ou des cohortes plus homogènes, donc plus adaptées aux études de liaison et d'association, afin d'identifier les gènes responsables des ces phénotypes.

Le TB peut se déclarer entre l'âge de 12 et 65 ans. Un début précoce de la maladie est associé à des symptômes plus sévères (rechutes plus fréquentes, présence d'éléments psychotiques au cours des accès, épisodes maniaques et comportements suicidaires dans 50% des cas) et à un risque familial accru (les apparentés de premier degré ont 30 fois plus de risque d'être atteints qu'un individu non-apparenté) (Leboyer *et al.*, 2005). En effet, il a été montré que des germains atteints de TB appartiennent plus souvent au même sous-groupe d'âge de début que ne le voudrait le hasard (Bellivier *et al.*, 2001). Ceci suggère que ces sous-groupes, définis suivant l'âge de début, seraient sous-tendus par des facteurs de vulnérabilité génétique spécifiques. Quatre études indépendantes (Bellivier *et al.*, 2001 ; Bellivier *et al.*, 2003 ; Lin *et al.*, 2006 ; Manchia *et al.*, 2008) ont montré que l'on peut définir trois sous-groupes de patients en fonction de l'âge de début de la maladie (Figure 6) : un groupe à début précoce pour lequel le premier épisode apparaît en moyenne à 17 ans et qui représente 28% de la population ; un groupe intermédiaire, dont l'âge de début est en moyenne 25 ans et qui correspond à 50% des patients ; et un groupe à début plus tardif correspondant aux 22% restant et dont l'âge moyen d'apparition du premier épisode est 40 ans. L'étude de ces trois

distributions permet également de définir deux seuils, à 21 ans pour distinguer les âges de début précoce et intermédiaire et à 37 ans pour distinguer les intermédiaires des tardifs.

**Figure 6. Distribution de la population de patients bipolaires en fonction de l'âge d'apparition des premiers symptômes.**



L'histogramme montre la distribution observée de l'âge de début de la maladie des patients bipolaires. Les trois courbes de Gauss représentent les trois distributions théoriques de ces sous-populations basées sur l'âge d'apparition des premiers symptômes de la maladie.

Walsh *et al.* ont montré, en 2008, que les patients schizophrènes présentaient un excès de CNV (15%) par rapport à des individus contrôles (5%). Deux autres études ont montré également que les CNV rares *de novo* et ceux d'une taille supérieure à 100 kb étaient plus fréquents chez les patients schizophrènes que chez les contrôles (Xu *et al.*, 2008 ; Stone *et al.*, 2008) et que certains d'entre eux interrompaient des gènes candidats pour la schizophrénie (Vrijenhoek *et al.*, 2008).

Etant donné que le TB et la schizophrénie présentent des similitudes cliniques et épidémiologiques, il paraît crucial de rechercher l'implication des CNV dans le TB. De plus, Zhang *et al.* ont mis en évidence, en 2008, que des délétions rares (non répertoriées dans la population générale) de plus de 100 kb étaient surreprésentées dans une cohorte de patients bipolaires (16,2%) par rapport à un groupe de témoins (12,3%). Ce constat était encore plus tranché pour un groupe de patient bipolaire avec un âge de début précoce pour les épisodes maniaques (18,9%), par rapport aux contrôles (12,3%).

Ces données, suggèrent que les CNV pourraient avoir une part de responsabilité dans la sensibilité au TB.

Les objectifs de cette étude expérimentale était de rechercher les CNV, dans une cohorte de 236 patients français atteints de TB à début précoce (TBDP), en analysant les résultats d'un génotypage à haut débit de 550 000 SNP pour chaque individu ; puis de valider, par deux méthodes indépendantes, la présence de certains de ces CNV, sélectionnés selon leur implication possible dans le TBDP. Enfin, nous avons cherché à caractériser les points de cassure de certains de ces variants afin de déterminer avec précision leur localisation.

## B. Matériel et méthodes

### 1. Constitution de la collection d'ADN

#### a. Inclusion des patients bipolaires

Cette étude a été réalisée sur trois cohortes humaines dont une cohorte de 236 patients français (patients ayant au moins trois grands-parents français), comprenant 163 patients bipolaires à début précoce (âge de début  $\leq 21$ ans) et 69 cas familiaux (patients d'âge de début  $> 21$  ans ayant au moins un apparenté au premier degré souffrant de trouble bipolaire). La constitution de cette cohorte est le résultat d'une dizaine d'années de collaboration entre plusieurs centres cliniques.

Les patients ont été inclus à Paris et Bordeaux, par des psychiatres expérimentés, et remplissaient les critères définis par le DSM-IV (American Psychiatric Association, *Diagnostic and Statistical Manual*, fourth edition) pour le trouble bipolaire 1 ou 2. Leur diagnostic a été évalué à l'aide de la *Diagnostic Interview for Genetics Studies version 3.0* (Nurnberger *et al.*, 1994). Tous étaient normothymiques au moment de l'inclusion. C'est-à-dire que les patients avaient un score inférieur à 5 pour le *Montgomery-Asberg Depression Rating Scale*, questionnaire utilisé pour mesurer la sévérité d'un épisode dépressif. Dix questions étaient posées aux sujets. Chaque réponse était notée sur 6, (0 représentant un état normal et 6 un état de dépression très sévère), (Montgomery et Asberg, 1979). Etaient considérés comme normothymiques les patients qui avaient également un score inférieur à 5 pour le *Mania rating Scale*, questionnaire utilisé pour évaluer la sévérité d'un épisode maniaque. Onze questions ont été posées aux patients. Chaque réponse était notée sur 4, (0 représentant un état normal et 4 un état maniaque sévère), (Bech *et al.*, 1979). L'âge d'apparition des premiers symptômes de la maladie a été évalué de manière rétrospective.

La seconde cohorte était constituée de 235 patients allemands, comprenant 172 patients bipolaires à début précoce et 63 cas familiaux (patients d'âge de début  $> 21$  ans ayant au moins un apparenté au premier degré souffrant de trouble bipolaire). Les critères d'inclusion des patients allemands dans l'étude étaient les mêmes que pour les patients français.

Un groupe contrôle était composé de 48 individus exempts de toute maladie psychiatrique et n'ayant pas d'apparenté du premier degré atteint de maladie psychiatrique.

## b. Extraction d'ADN

L'ADN des patients a été extrait à partir de sang total prélevé sur EDTA (Ethyl Diamine Tétracide Acétique). Des lignées de lymphocytes B immortalisées par le virus d'Epstein Barr ont été générées pour chaque patient, afin de disposer d'une grande quantité d'ADN et ceci sur plusieurs années.

Les ADN ont été isolés à l'aide du kit Nucleon BACC 3 (GE Healthcare) et ont été stockés dans l'ADNthèque de l'Hôpital H. Mondor (Créteil). L'ADN d'autres membres de la famille (parents, frères et sœurs, enfants) était disponible dans certains cas et a également été archivé.

## c. Dosage et estimation de la qualité de l'ADN

L'ADN a été dosé à l'aide du kit Quant-iT dsDNA BR Assay Kit, (Invitrogen Carlsbad, CA, USA) selon les instructions du fabricant.

La qualité de l'ADN a été évaluée par migration de 250ng d'ADN sur gel d'agarose à 0,8% pendant 2h à 50V.

# 2. Analyse des CNV

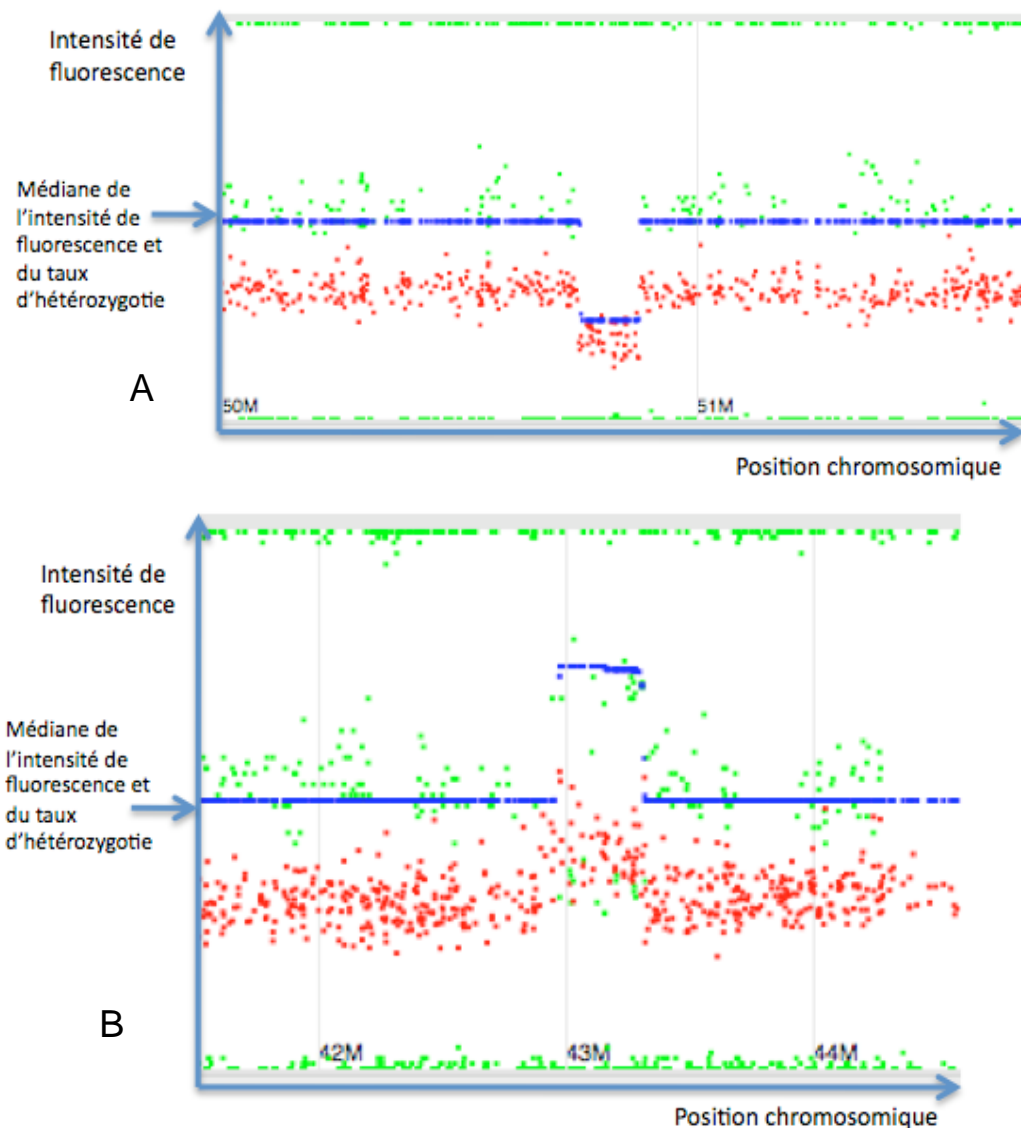
## a. Détection des CNV

Nous avons réalisé une analyse bioinformatique des puces à ADN HumanHap 550 Beadchips (Illumina Inc. San Diego, Etats-Unis) pour les patients allemands et français. Les individus contrôles ont été génotypés à l'aide de puces HumanHap 1 million Beadchips (Illumina Inc. San Diego, Etats-Unis).

Une collaboration avec le Centre National de Génotypage (CNG : [www.cng.fr](http://www.cng.fr)) nous a permis de génotyper 236 patients français et 235 patients allemands pour plus de 550 000 SNP, et 48 contrôles pour plus d'un million de SNP, répartis tout au long du génome et de l'ADN mitochondrial. Quatre patients français, 15 patients allemands et 2 individus contrôles ont dû être exclus de l'étude à cause d'un génotypage de mauvaise qualité.

Les données générées par ces puces (550 000 et un million de SNP) ont été analysées à l'aide du logiciel SnipPeep (Version 1.0, auteur Roberto Toro) (Figure 7) qui représente l'intensité de fluorescence et le taux d'hétérozygotie de chaque SNP.

**Figure 7. Représentation des caractéristiques de chaque SNP par le logiciel SnipPeep, pour chaque individu.**



En abscisse est représentée la position sur le chromosome et en ordonnées l'intensité de fluorescence (rouge) et le taux d'hétérozygotie (vert) pour chaque SNP génotypé. La ligne bleue représente la médiane d'intensité de fluorescence et du taux d'hétérozygotie des SNP, cette ligne prédit le nombre de copie(s). A. Représentation d'une délétion. B. Représentation d'une duplication.

L'estimation de la présence de délétion ou duplication est fondée sur l'intensité de fluorescence et le taux d'hétérozygotie des SNP. La diminution ou l'augmentation de l'intensité de



fluorescence de plusieurs SNP contigus nous permet de détecter les variants. La première vérification de la validation d'une délétion ou duplication est la diminution du taux d'hétérozygotie.

Nous avons ainsi pu repérer les délétions et duplications de chaque patient. Nous avons évalué la taille des variants en soustrayant la position chromosomique du dernier SNP inclus dans le variant, à la position chromosomique du premier SNP inclus dans le variant. Une recherche sur le site du *Database of Genomic Variants* (DGV <http://projects.tcag.ca>), base de données publique où sont répertoriés les variants connus et caractérisés dans la population générale humaine, a été effectuée, afin de vérifier si le variant était déjà répertorié dans la population générale.

Nous avons recherché sur le site de l'Université de Santa Cruz (*UCSC Genome Browser* ; <http://genome.ucsc.edu>), qui contient les données de séquence et l'annotation du génome humain, dans quelles régions se situaient les CNV : région contenant des gènes, des éléments régulateurs, des ESTs, et s'il y avait des gènes, quelle partie était touchée (exon ou intron). Le site *UCSC Genome Browser* nous a ainsi permis de voir quels gènes pourraient potentiellement être affectés et d'avoir facilement accès aux données concernant leur fonction et leur expression. Nous avons ensuite parcouru des sites tels que *Online Mendelian Inheritance in Man*, base de données des maladies génétiques humaines (OMIM, [www.ncbi.nlm.nih.gov/omim](http://www.ncbi.nlm.nih.gov/omim)), *Ensembl*, données de séquence et annotations de différents génomes ([www.ensembl.org](http://www.ensembl.org)), *GeneCards* ([www.genecards.org](http://www.genecards.org)), *Allen Brain Atlas* ([www.brain-map.org](http://www.brain-map.org)), *Pubmed* ([www.ncbi.nlm.nih.gov/pubmed](http://www.ncbi.nlm.nih.gov/pubmed)) afin d'affiner les données. Des informations complémentaires ont été ajoutées pour certains CNV, comme ceux présents dans une région du génome déjà associée au trouble bipolaire, ou à d'autres troubles psychiatriques.

## b. Sélection des CNV

Nous avons sélectionné les CNV candidats pour le trouble bipolaire selon les critères suivants :

- le CNV interrompait un ou plusieurs exons codants d'un gène exprimé dans le cerveau ;
- le CNV était susceptible de modifier l'expression génique ou le fonctionnement d'une protéine;
- le CNV était présent chez plusieurs patients de la cohorte mais n'était pas répertorié dans la population générale.

-l'ADN du proposant, et éventuellement d'autres membres de la famille, était disponible et de bonne qualité.

### c. Validation des CNV

Nous avons cherché à valider la présence des délétions et duplications par plusieurs méthodes quantitatives.

#### GeneScan

L'objectif de cette méthode était de quantifier la différence du nombre de copies d'ADN pour les régions suspectées contenir un CNV. Chaque région d'intérêt a été comparée à deux régions de référence pour lesquelles nous avons vérifié qu'aucun patient ni témoin ne présentait de CNV. Nous avons, ainsi, utilisé trois couples d'amorces pour chaque patient et chaque témoin.

Dans cette méthode nous avons utilisé pour chaque réaction 0,5  $\mu$ M d'amorce sens marquée avec un fluorochrome FAM et 0,5  $\mu$ M d'amorce antisens non marquée afin d'amplifier 100ng d'ADN, d'un patient ou d'un témoin, dans un mélange réactionnel de 20  $\mu$ l comprenant 0,2 mM de dNTP, du tampon 1X, 0,75mM de MgCl<sub>2</sub> et 0,25U de Taq DNA polymérase (Eurobio). Après une pré-dénaturation de 2 minutes à 94°C, 15, 20 ou 25 cycles d'amplification ont été réalisés selon les conditions suivantes : 30 secondes de dénaturation à 95°C, hybridation des amorces à 60°C pendant 30 secondes et 1 minute d'élongation à 72°C.

Après purification des produits obtenus en G50 sur Séphadex (Sigma Aldrich, Etats-Unis) selon les instructions du fabricant, les échantillons ont été déposés sur un séquenceur automatique ABIPRISM 3130 *xl* Genetic Analyser (Applied Biosystems).

#### Analyse des résultats de GeneScan

Nous avons analysé les résultats de GeneScan à l'aide du logiciel GeneScan Analysis version 3.7, (Applied Biosystems). Nous avons calculé les rapports d'intensité de fluorescence entre les produits de PCR de la région d'intérêt et celles des régions de référence pour chaque patient et chaque témoin.

#### PCR quantitative

La PCR quantitative permet de quantifier l'ADN synthétisé à chaque cycle grâce à un intercalant fluorescent spécifique de l'ADN double brin : le Sybr Green (émission de fluorescence à 520nm). La fluorescence est proportionnelle à la quantité d'ADN polymérisé. La PCR quantitative se base sur le fait qu'à chaque cycle la quantité d'ADN de la région d'intérêt est doublée. Pendant la phase exponentielle (quand aucun élément de la PCR n'est limitant) la quantité d'ADN à un cycle donné est proportionnelle à la quantité initiale de la matrice. Le nombre de cycles qu'il faut réaliser pour obtenir une certaine quantité d'ADN (et donc de fluorescence) est appelé cycle-seuil ou Ct (pour *Cycle Number Threshold*).

Nous avons pris un couple d'amorces dans chaque région d'intérêt (région supposée délétée ou dupliquée) et dans deux régions de référence et nous avons amplifié ces régions chez les patients ainsi que deux témoins afin de valider la présence de délétions ou duplications. Nous avons utilisé deux thermocycleurs et deux kits de PCR quantitative différents pour valider nos résultats.

La première méthode de PCR quantitative utilisée correspond à une amplification réalisée avec le kit MESA GREEN qPCR selon les instructions du fabricant (Eurogentec, Liège, Belgique), qui a été lue sur le thermocycleur « Mastercycler ep Realplex2S » (Eppendorf, Hambourg, Allemagne). Nous avons utilisé pour chaque réaction 0,3 µM de chaque amorce afin d'amplifier 20ng d'ADN dans un mélange réactionnel de 25 µl. Après une pré-dénaturation de 5 minutes à 95°C, 40 cycles d'amplification ont été réalisés selon les conditions suivantes : 15 secondes de dénaturation à 95°C, hybridation des amorces à 60°C pendant 20 secondes et 50 secondes d'élongation à 68°C.

La seconde méthode de PCR quantitative utilisée correspondait à une amplification suivant le même protocole que décrit précédemment, avec le kit LightCycler 480 High Resolution Melting Master (Roche, Mannheim, Allemagne), qui a été lue sur le thermocycleur « Light Cycler 480 » (Roche, Mannheim, Allemagne).

#### Analyse des résultats de PCR quantitative

Nous avons utilisé la méthode du  $\Delta\Delta Ct$ , qui permet de comparer des valeurs de Ct entre les gènes de référence et le gène cible. Cette comparaison est possible si l'efficacité des PCR est proche de 100% et ne diffère pas de plus de 10% entre elles.

On calcule d'abord le delta Ct entre le gène cible et le gène de référence,  $\Delta Ct = Ct_{\text{cible}} - Ct_{\text{gène de référence}}$ , pour chaque patient et un témoin. Ensuite, on fait la différence entre le  $\Delta Ct$  du patient et celui du témoin, ce calcul donne le  $\Delta\Delta Ct$  ( $\Delta\Delta Ct = (Ct_{\text{cible}} - Ct_{\text{référence}})_{\text{patient}} - (Ct_{\text{cible}} - Ct_{\text{référence}})_{\text{témoin}}$ ). La quantité d'ADN pour la région d'intérêt chez le patient, normalisée par rapport au témoin,

est égale à  $2^{-\Delta\Delta Ct}$ . Cette valeur permet de comparer les quantités d'ADN initiales chez le patient et le témoin. La délétion d'une copie par rapport au témoin (2 copies) correspond à un  $2^{-\Delta\Delta Ct} = 0,5$ . La duplication d'une copie (3 copies) par rapport au témoin (2 copies) correspond à un  $2^{-\Delta\Delta Ct} = 1,5$ .

#### d. Caractérisation des points de cassure

##### Amplification

L'objectif de cette méthode était d'amplifier les régions flanquant les points de cassure des délétions. Nous avons utilisé 0,4  $\mu$ M d'amorce sens choisie en amont du dernier SNP non inclus dans le variant et 0,4  $\mu$ M d'amorce antisens choisie en aval du premier SNP après la délétion, afin d'amplifier 250ng d'ADN dans un mélange réactionnel de 25  $\mu$ l comprenant 0,8 mM de dNTP, du tampon à 1X, 5 mM de MgCl<sub>2</sub> et 0,5U de Taq polymérase Takara LA spécifique des grands fragments (>1kb), fournie par le kit TaKaRa LA Taq (Takara Bio Inc., Shiga, Japon). Après une pré-dénaturation d'une minute à 94°C, 30 cycles d'amplification ont été réalisés selon les conditions suivantes : 30 secondes de dénaturation à 94°C, une hybridation des amorces à 58°C pendant 30 secondes et une élongation à 68°C, dont la durée dépendait de la taille du produit à amplifier, nous avons compté 1 minute par kilobase à synthétiser.

##### Clonage du produit de PCR

Nous avons réalisé ce clonage, selon les instructions du fabricant, à l'aide du kit TOPO-TA Cloning (Invitrogen Inc., Carlsbad, Allemagne). Nous avons ensuite amplifié les clones selon un protocole classique de PCR utilisant les amorces M13F et M13R fournies par le kit TOPO-TA Cloning, en vue de les séquencer.

##### Séquençage

Après amplification, les produits de PCR ont été purifiés, afin d'éliminer les dNTP et les amorces libres, par l'adjonction de 0,25U de SAP et 2,5U d'exonucléase I à 37°C pendant 40 minutes, puis à 80°C pendant 20 minutes pour inactiver les enzymes.

Nous avons effectué une réaction de séquence, qui est fondée sur la méthode de Sanger. Celle-ci a été réalisée suivant les instructions d'utilisation du kit Big Dye Terminator v3.1 (Applied Biosystems). Le produit de réaction de séquence a ensuite été purifié et déposé sur un séquenceur

16 capillaires ABI 3100 Genetic Analyser (Applied Biosystems). Nous avons visualisé les séquences grâce au logiciel Finch TV version 1.4.0 (Geospiza Inc.). Enfin, ces séquences ont été alignées avec le séquençage du génome humain disponible sur le site *UCSC Genome Browser*, afin de déterminer leur position dans le génome de l'Homme.

### 3. Analyse Statistique

Les analyses statistiques des résultats ont été effectuées par le test du Chi<sup>2</sup>, au seuil de signification de 3,84 à 1 degré de liberté. Une p-value de < 5% était considérée comme significative. Une correction de Yates a été appliquée pour les groupes qui présentaient moins de 5 patients.

### 4. Sites web

*Allen Brain Atlas* (<http://www.brain-map.org>), site qui recense l'ensemble des données concernant l'expression des protéines dans les différentes parties du cerveau chez la Souris.

*Basic Local Alignment Search Tool*, BLAST, (<http://blast.ncbi.nlm.nih.gov/Blast>), permet de rechercher des similarités entre des séquences protéiques ou nucléotidiques et de les comparer aux séquences de référence.

*Database of genomic variations (DGV)*, (<http://projects.tcag.ca>), base de données publique où sont répertoriés les variants connus et caractérisés dans la population générale humaine.

*Ensembl* (<http://www.ensembl.org>), données de séquence et annotations de différents génomes.

*Evolutionary Conserved Region* (<http://ecrbrowser.dcode.org>), permet d'étudier le taux de conservation au cours de l'évolution d'une séquence d'acide nucléique.

*GeneCards* (<http://www.genecards.org>), base de données réunissant les informations concernant la fonction, l'expression, et les aspects médicaux des gènes humains et de leur **orthologues** de Souris.

*National Center for Biotechnology Information (NCBI)*, (<http://blast.ncbi.nlm.nih.gov>), regroupe, entre autres, les sites Pubmed, BLAST, OMIM, OMIA.

*Online Mendelian Inheritance in Man (OMIM)*, (<http://www.ncbi.nlm.nih.gov/omim>), base de données des maladies génétiques humaines.