Le macrosatellite RNU2 : caractérisation, évolution et lien avec la prédisposition génétique au cancer du sein

Chloé Tessereau

N° d'ordre 75-2014 Année 2014

THESE DE L'UNIVERSITE DE LYON

Délivrée par

L'UNIVERSITE CLAUDE BERNARD LYON 1

ECOLE DOCTORALE BIOLOGIE MOLECULAIRE INTEGRATIVE ET CELLULAIRE

DIPLOME DE DOCTORAT

(arrêté du 7 août 2006)

soutenue publiquement le 16 Mai 2014

par

Madame Chloé Tessereau

TITRE :

Le Macrosatellite *RNU2* : Caractérisation, Evolution et Lien avec la Prédisposition Génétique au Cancer du Sein

Directeur de thèse : Docteur Sylvie Mazoyer

JURY :

Docteur Sophie Gad (Examinateur)

Docteur David Goldgar (Examinateur)

Docteur Frédérique Magdinier (Rapporteur)

Professeur Jean-Louis Mandel (Rapporteur)

Docteur Sylvie Mazoyer (Directeur de Thèse)

Professeur Damien Sanlaville (Examinateur)

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme le Professeur H. PARROT
Département GEP	Directeur : M. N. SIAUVE
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur A. GOLDMAN
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : Mme S. FLECK
Département Sciences de la Terre	Directeur : Mme la Professeure I. DANIEL
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. COLLIGNON
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Administrateur provisoire : M. N. LEBOISNE

Un roman, c'est un miroir qu'on promène le long d'un chemin.
Stendhal, Le Rouge & Le Noir.

A Mme Rozier, Mme Claire Vourc'h & Mr Olivier Bensaude qui m'ont initié au chemin de la Science,
Au Dr Jaboulay qui a abrité 2190 nuits et jours d'échanges passionnants,
A mon grand-père qui, s'endormant sur le canapé, m'a transmis sa patience,
A Catherine, qui a livré un combat déchirant.

Remerciements

Ma première pensée s'adresse à ma directrice de thèse, Sylvie Mazoyer. Les mots ne suffiraient pas pour te remercier pour tout ce que tu m'as apporté. Pendant 5 ans, ton encadrement a été *excellent*, terme que nous détestons toutes les deux tout autant mais qui convient ici parfaitement. Tu as fait naître en mot l'envie de *faire* une thèse, en me proposant un sujet qui m'a passionnée. Je te remercie pour le savoir que tu m'as transmis, pour tes qualités humaines qui m'ont fait grandir, et pour le temps précieux que tu m'as accordée. J'espère que nos routes scientifiques se recroiseront souvent. En attendant, très sincèrement, merci.

Je remercie Madame Frédérique Magdinier et Monsieur Jean-Louis Mandel pour avoir accepté d'évaluer ce manuscrit.

Je remercie Madame Sophie Gad, Monsieur Damien Sanlaville et Monsieur David Goldgar pour m'avoir fait l'honneur de participer à mon jury de soutenance.

Je remercie très sincèrement Monique Buisson, pour son aide au quotidien à la paillasse, pour son savoir qu'elle m'a transmis patiemment, pour les nombreuses relectures de ce manuscrit, mais surtout pour sa gentillesse et sa délicatesse. Travailler à tes côtés a été une chance dont je mesure l'importance.

J'adresse mes remerciements à Gaël Yvert et Richard Redon, qui m'ont fait l'honneur de participer à mes trois comités de suivi de thèse et qui ont également su se montrer disponibles en d'autres occasions. Vos conseils ont été plus que bénéfiques.

Pendant ma thèse, j'ai été entourée et j'ai eu la chance de rencontrer des personnes extraordinaires grâce auxquelles j'ai énormément appris, tant au niveau scientifique qu'humain. Quelques pages ne suffiraient pas à leur exprimer mon immense gratitude. Merci à Olga Sinilnikova et Mélanie Léoné pour leur gentillesse et leur aide. Merci à

Laure Barjhoux, Carole Verny-Pierre et Valérie Sornin, qui font souvent un travail dans l'ombre et dont la contribution a été inestimable pour cette thèse. Merci à toutes les personnes du 1er étage du Cheney D (Cyril, Zaza, …) qui ont toujours été disponibles pour m'expliquer le fonctionnement de chaque instrument ou me prêter des feeders.

Puisque les échanges scientifiques n'ont pas de frontière, j'ai eu la chance de réaliser ma thèse avec un financement CIFRE entre un laboratoire académique à Lyon, et la société Genomic Vision à Paris. Merci à Aaron Bensimon, Daniel Nerson et Erwan Martin pour avoir accepté de financer ma thèse et pour avoir nourri mon âme d'enfant en me permettant de *peigner* l'ADN.

Merci à Sébastien Barradeau pour son soutien et nos échanges scientifiques à chacune de mes venues à Genomic Vision, à Jennifer Abscheidt, Marjorie Pierret et Stéphanie Bouchilloux pour leur gentillesse, leur dynamisme, et leur aide technique irremplaçable, à Kévin Cheeseman pour nos discussions passionnées sur la réalité de la vie de thésard, à Aurélie Thomas pour la formation qu'elle m'a prodiguée à mes débuts, à Emilie Renard pour toutes les commandes qu'elle a passées, à Fadilha Abut pour l'organisation de chacun de mes déplacements. Merci également à tous ceux qui se sont impliqués dans ce projet et avec qui les échanges scientifiques ont fait germer de nouvelles idées, dont Jun Komatsu et Lucia Cinque.

Merci à toutes les personnes que j'ai citées précédemment et également Solène Guillon, Yannick Fourne, Djamila El Mhali, Samira Jbilou, Fanny Lemée, Sara Berthoumieux, Agnès Cibiel, Alexandra Nghe, pour leur bonne humeur qui accompagnait mes journées à la pépinière de Cochin puis à Bagneux. Un clin d'oeil tout particulier à mes compagnons de crop qui se reconnaitront…

Ces quatre années ont été parsemées de quelques départs. Je tiens donc à remercier Anne Vannier, qui m'a formée à la technique du peignage moléculaire et qui a été à l'origine du premier marquage de la version 1.0 du code-barres *RNU2*. Merci à Emmanuel Conseiller qui a cru à mon projet et me l'a montré au moins un jeudi par mois. Enfin, merci à Quynh Dao Joyez, Mimounia et Rachel Morra.

Un merci tout particulier à Laurent Duret, qui m'a accueillie au sein de son équipe pendant quelques mois et a su faire preuve de patience pour m'initier au monde merveilleux de la bio-informatique. Merci également à Yann Lesecque, Sylvain Mousset et Mathieu Groussin pour leur assistance technique et scientifique face à R, Perl, et cie … Merci également à Thomas Bigot pour ses conseils pointus sur ce manuscrit. Bien évidemment, merci à tous les membres du LBBE, toujours souriants et bienveillants lors de mon séjour prolongé ou de mes excursions passagères.

Je remercie également Damien Sanlaville et Caroline Schluth-Bolard, qui m'ont accueillie dans leurs locaux à Bron et m'ont formée à la technique de FISH. Merci également à toutes les techniciennes de l'équipe pour leur gentillesse et leur aide.

Merci à Marc Billaud qui m'a accueillie dans son laboratoire, l'UMR5201, pour la réalisation de mon Master 2. Merci à Alain Puisieux qui m'a accueillie au sein du Centre de Recherche en Cancérologie de Lyon pour mes trois dernières années de thèse.

Merci à Daniel Birnbaum et Anne Le Tessier pour notre rencontre à Marseille qui a ouvert de nouvelles perspectives à ce projet : l'exploration de la fragilité du locus. Merci à Michelle Debatisse pour ses conseils lors de notre rencontre à Paris. Merci également à Arnaud Coquelle pour l'intérêt immédiat qu'il a manifesté pour cette hypothèse, pour le temps passé lors de notre rencontre à Montpellier, et pour son implication depuis. J'espère sincèrement que cela conduira à de belles choses.

Merci à Olga Sinilnikova, Dominique Stoppat-Lyonnet et Nadine Andrieu pour m'avoir permis d'utiliser la cohorte GENESIS. Merci à Fabienne Lesueur et Nadine Andrieu pour leur réflexion statistique. Merci à Marie-Gabrielle Dondon, Juana Beauvallet & Séverine Eon-Marchais pour leur disponibilité et leur rapidité à générer les arbres, ainsi qu'à tous les collaborateurs de l'étude GENESIS à l'Institut Curie.

Merci à Yann Lesecque et Floriane Plard pour leur contribution majeure à l'analyse des données GENESIS et BCFR.

Merci à Bingjian Feng et David Goldgar pour le calcul de l'âge des mutations *BRCA1*.

Merci à Fabienne Le Calvez-Kelm pour son aide pour BCFR. Un merci particulier à Amélie Chabrier pour sa gentillesse et le temps qu'elle m'a accordé.

Merci à Zdenko Herceg pour sa collaboration sur l'analyse de la méthylation. Et un grand merci à Cyrille Cuenin pour son implication et les beaux résultats !

Merci à Marie-Eve Fondrevelle pour son implication dans la collecte des tumeurs, et merci aux techniciennes du service d'Anapath' du CLB pour ma formation à la macrodissection.

Merci à Emiliano Ricci, qui a été mon premier maitre de stage en Licence, et qui m'a donné goût à la recherche par sa patience et son enthousiasme. Merci à Théophile Ohlmann qui m'a accueillie dans son laboratoire pour ce stage.

Pour finir, je tiens à remercier très sincèrement Julien Varaldi, François Bonneton et Laurence Mouton pour m'avoir fait confiance en me confiant différents enseignements. Cette expérience m'a été précieuse et j'espère la renouveler !

*****

Ces cinq années passées au sein du Cheney D resteront associées à des rencontres incroyables et des fous-rires mémorables. Je remercie très sincèrement tous mes collègues qui m'ont soutenue et supportée au quotidien. Vous avez largement contribué à la réalisation de cette thèse, et surtout au plaisir immense que j'avais chaque matin à vous retrouver. En commençant une thèse, je ne me doutais pas que je repartirais de ce laboratoire avec de véritables amis, plus précieux que des résultats scientifiques.

Je remercie en premier lieu Rémy, mon grand frère au laboratoire. Toujours présent, toujours à l'écoute, tu m'as aidée à affronter les petits problèmes techniques mais également les grands moments de doute.

Merci à tous les *jeunes* et moins jeunes du 5ème étage, vous avez été des supers camarades de déjeuner et même plus pour certains. Vous allez me manquer.

Une petite dédicace à Rami, dont le rire tonitruant a retentit si souvent, à Sophie, pour son sourire et ses mots d'encouragements, à Romain, que je n'aime pas trop mais cite quand même, à Elise, et sa passion dévorante pour le chocolat, à Anne, dont le sens artistique a illuminé le laboratoire pour de si nombreuses occasions, à Doriane, sa volonté et sa délicatesse, à Estelle, pour sa douceur, à Chang et Philippe, pour l'accueil dans la Team *Men1* au self.

Merci aux anciens membres du 5ème et 6ème avec qui j'ai commencé cette aventure : Amandine, à qui j'ai piqué le bureau mais avec qui j'ai pris grand plaisir à travailler, Malek (& Koliane), et son humour légendaire, Nicolas, mon partenaire précieux de Master.

Merci à tous les membres du 6ème étage, dont Baptiste, Justine, Logine, Rana, Mélanie, Mélanie, Marc … pour les carnavals (et le reste !).

Merci à tous les membres du Sheep Department pour nos discussions hors sciences et soirées endiablées : Thibault, Gabriel, Etienne, Stéphane, Eleonora, Anne-Laure, Julien, Mathieu, Laurent, Clément, Benjamin.

Une pensée à mes deux stagiaires que j'ai eu le plaisir d'encadrer pendant 4 mois chacune, Nastasia & Marine. Travailler avec vous a été très formateur, bonne chance dans vos parcours respectifs !

Merci à tous les membres du Département Flux d'Information dans la cellule cancéreuse, Laura Corbo, Ruth Rimokh, Germain Gillet, Ivan Mikaélian, Jean-Jacques Diaz, Didier Auboeuf, pour les bons souvenirs des deux journées scientifiques.

Merci aux secrétaires du CRCL qui m'ont beaucoup aidée : Sophie Paulet, Basma Zamit & Dominique Pianetti.

*****

Merci à tous mes amis que j'ai rencontrés pendant ma thèse et qui m'ont tous, à leur manière, beaucoup aidés.

Riton & Lucie pour les belotes, Charols et cie.

Marinette pour nos supers brunchs.

Fanny, Anne-So, Denis, Julien, Charlotte, Alexis, Ambroise, Téréza, Olivier, Fred, Max, Fanette, Max, pour nos virées en vélo dans la Loire, les Nouvel-Ans, les joncs à Carqueiranne, …

Alexis pour nos footings irréguliers et la découverte de Gerland.

Clémence et Nathalie pour les trainings, et surtout pour votre bonne humeur.

Clément pour sa gentillesse.

Merci à tous les parisiens qui m'ont prêté un bout de canapé et/ou concocté un bon repas pendant mes séjours plus ou moins prolongés : Adrien & Emeline, Louis-Marie, Anne, Marie & Cyril (& maintenant la belle Léonie), Fanny & Julien, Dup-Dup, Anne-So & Denis.

Merci à Jeanne et Marie pour leur accueil chaleureux, et tout ce que nous avons partagé.

*****

Ces remerciements ne seraient pas complets si je ne mentionnais pas les courageux qui ont partagé, pour des durées variables, leur quotidien avec moi dans 72 m^2. Merci à tous les colocataires du 72 Rue Jaboulay : Yann, Louis-Marie, Florent, Harrison, Pierre, Antoine, Simon, Marine, Damien. Et merci aussi à ceux qui sont venus nous rendre visite pour quelques heures ou quelques jours ! Sonnez à Bobay ☺

Je remercie également très sincèrement mes camarades de l'Ecole Normale Supérieure de Lyon, Yann, Mathieu, Florent, Blaise, Pierre, Florie, Domitille, Luc, Alexandre, Nelly, Germain. Egalement Adrien & Emeline, à qui je souhaite beaucoup de bonheur.

Un merci particulier à Domitille, qui m'a réveillée tous les matins pendant notre année de Licence, et qui m'a ainsi permis d'aller en cours sans trop de difficulté !

*****

Merci aux *types* pour nos virées dans les Cévennes, à la Roche, à Niolon, à Noirmout', pour le lever de soleil dans l'amphi de l'ENS, pour tout le bonheur que vous m'avez apporté en résumé …

A Flo, mon coco, mon partenaire de rando, mon *précieux* de Nouvelle-Zélande, et pour sa supervision sur le taillage de carottes et d'oignons.

A Grouss, pour sa générosité, ses conseils avisés et nos retrouvailles prochaines outre-Atlantique.

A Blaise, pour son éducation politique, son accueil toulousain, son soutien quasi-quotidien sur skype.

A Lemerre, pour ses vidéos loufoques de lol-cats et ses mix percutants.

Cette thèse a commencé sur les toits de Jodhpur, et nous aura mené devant les temples de Bagan, face au Khazneh de Pétra, sur la plage de Phi Phi, au sommet du Pain de Sucre, … Merci à Pierre.

Déjà 6 ans que nous partageons nos céréales en écoutant France Inter, que nous covoiturons pour retourner dans nos *chères* montagnes, et que grâce à toi je découvre en exclusivité les nouveaux titres pop. Merci à Yann pour sa joie de vivre, le soutien technique et humain qu'il m'a manifesté pendant cette période intense de rédaction et pour son *Ave Maria* à chaque Fête des Lumières.

*****

Merci à tous mes ami(e)s fidèles depuis le collège et le lycée maintenant, voire même avant pour certains. La distance, les années, et les changements de cap ne nous ont pas séparés, et j'en suis heureuse.

Anaïs & Anne-Sophie, vous m'avez toujours épaulée. Merci pour votre présence inégalable depuis 12 ans maintenant, vos délicates attentions et vos messages plein d'humour. Je suis fière de vous avoir à mes côtés pour ce jour particulier.

Merci à Ben, Ludo, Débo, Rémi, Anaïs, Anne-So & Jo (& la crevette Elsa), qui m'accompagnent maintenant depuis le LGM et avec qui je prends toujours autant de plaisir à partager des moments privilégiés : les tarots, les raquettes, les soirées déguisées…

Merci à Charlotte pour nos retrouvailles régulières !

Merci à Vincent pour sa mauvaise foi notoire, sa mauvaise humeur permanente et pour la visite du Mont Saint Michel.

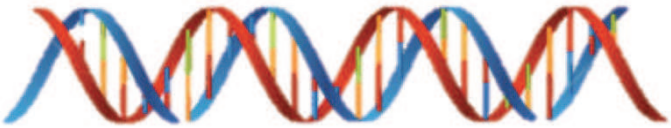
Merci à Chloé, ma première amie en arrivant à Grenoble, et que j'ai retrouvée avec plaisir sur Lyon.

Merci à mon cher ami d'enfance, Guilhem (& Maya). Heureusement que nous n'avons pas fait notre thèse dans le même domaine, nous aurions encore été en compétition !

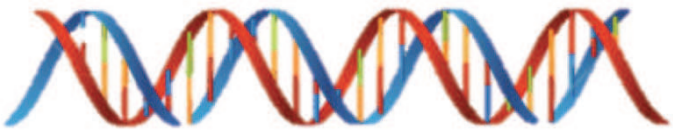
Je tiens à remercier très sincèrement toute ma famille pour son aide précieuse : ma grand-mère Jeannie, ma grand-mère Rénelde, Jean-Louis, Chantal, Rose-May, Renée, tata Mado, Alain & Cécile…

Pour finir, un énorme merci à mes parents et ma sœur Charline qui m'ont toujours soutenue durant mon cursus universitaire parsemé de questions existentielles et de crises identitaires, sans pour autant en comprendre l'origine et le bien-fondé.

SOMMAIRE



Liste des Abreviations

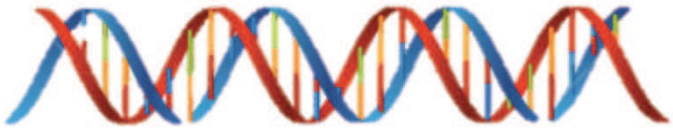


aa	acide aminé
aCGH	Array Comparative Genome Hybridization
AD	Activation Domain
ADN	Acide DésoxyriboNucléique
ADNc	ADN complémentaire
ADNr	ADN ribosomal
araC	Arabinoside Cytosine
ARN	Acide RiboNucléique
ARNr	ARN ribosomal
ARNsn	Small Nuclear RNA (petit ARN nucléaire)
ARNt	ARN de transfert
ATM	Ataxia-Telangiectasia Mutated
BAC	Bacterial Artificial Chromosome
BIC	Breast Cancer Information Core
BRCA1	Breast Cancer Gene 1
BRCA2	Breast Cancer Gene 2
BRCT	*BRCA1* C-terminal
BRIP	*BRCA1*-interacting protein C-terminal helicase 1
CDB	Cassure Double Brin
CEU	Utah residents with ancestry from northern and western Europe from CEPH
CHB	Han Chinese in Beijing, China
cM	centiMorgan
CNP	Copy Number Polymorphism
CNV	Copy Number Variation
CNVR	Copy Number Variant Region
CREB	cAMP-Responsive Element Binding
CSB	Cockaine Syndrome B
dbVar	Database of genomic structural variation
DECIPHER	Database of Chromosomal Imbalances using Ensembl Resources
DGV	Database of Genomic Variants
DHCN	DOC-estimated Haploid Copy Number
DS	Duplication Segmentaire
DSB	Double Strand Break
DSE	Distal Sequence Element
ERE	Estrogen Response Element
FISH	Fluorescent *in situ* Hybridization
FoSTeS	Fork Stalling and Template Switching
FSHD	FascioScapulohumeral Muscular Dystrophy
GWAS	Genome-Wide Association Study
HapMap	Haplotype Mapping
HAT	histone acétyl-transférase
HBOC	Hereditary Breast and Ovarian Cancer
HGP	Human Genome Project
HNPP	Hereditary Neuropathy with Liability to Pressure Palsy
HR	Homologous Recombination
Indel	Insertion/délétion

JPT	Japanese in Tokyo, Japan
kb	kilobase
L1	Long interspersed element-1
LCR	Low Copy Repeat
LCV	Large-scale Copy number Variation
LSTR	Long Segment Tandem Repeat
LTR	Long Terminal Repeat
MAR	Marker chromosome
mb	Mégabase
MMBIR	Microhomology-Mediated Break-Induced Replication
NAC	Nombre allélique de copies
NAHR	Non-Allelic Homologous Recombination
NBR2	Near *BRCA1* gene 2
NES	Nuclear Export Signal
NGC	Nombre global de copies
NGS	Next-Generation Sequencing
NHC	Nombre haploidique de copies
NHEJ	Non-Homologous End Joining
NLS	Nuclear Localisation Signal
NMD	Non-Sense Mediated Decay
nt	Nucleotide
ORF	Open Reading Frame
pb	Paire de Bases
PCR	Polymerase Chain Reaction
PFGE	Pulse Field Gel Electrophoresis
PolII	Polymérase II
PRE	Progesterone Response Element
PRR	Positive Regulatory Region
PSE	Proximal Sequence Element
qPCR	Quantitative PCR
RFLP	Restriction Fragment Length Polymorphism
RIGS	Repeat-Induced Gene Silencing
RING	Really Interesting New Gene
RNPsn	Small Nuclear RiboNucleoprotein
RT	Rétrotranscription
SD	Segmental Duplication
SGA	Séquençage Global Aléatoire
SNAPc	snRNA-Activating Protein Complex
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
STR	Short Tandem Repeat
TF	Transcription Factor
TRDB	Tandem Repeat DataBase
TRF	Tandem Repeat Finder
UCHL1	Ubiquitin Carboxyl-terminal esterase L1
USP17	Ubiquitin-Specific Protease 17

UTR	Untranslated Region
VIH	Virus de l'Immunodéficience Humaine
VNTR	Variable Number of Tandem Repeats
YAC	Yeast Artifical Chromosome
YRI	Yoruba in Ibadan, Nigeria

TABLE DES ILLUSTRATIONS

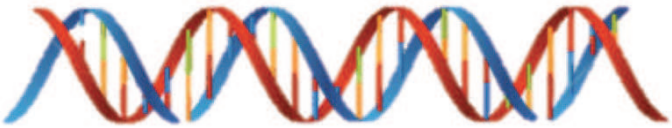


Figures

Tableaux

Chapitre 1

Revue Bibliographique



Mises à part quelques exceptions, chaque cellule d'un organisme contient son génome, à savoir l'ensemble de l'information génétique nécessaire à la vie. La découverte du support de cette information génétique, l'ADN, et plus récemment son déchiffrage ont eu des retombées scientifiques incalculables, et ont permis un prodigieux accroissement de la compréhension de l'organisme humain et des mécanismes à l'origine des maladies génétiques. Déchiffrer le génome, c'est-à-dire le séquencer, est devenu de plus en plus facile, rapide et peu cher. Aujourd'hui, plus de 98 % du génome humain est découvert. Puisque chaque individu possède son propre génome, un certain nombre de différences peuvent ainsi être observées entre les génomes de deux individus. Ces différences sont appelées des variants et sont classées selon le nombre de bases qu'elles affectent. Ces variations peuvent être bénignes, c'est-à-dire sans conséquences au niveau de l'organisme, ou au contraire influencer un trait particulier (par exemple la taille d'un individu ou la survenue d'une maladie). Les variants de grande taille, c'est-à-dire comportant un grand nombre de nucléotides successifs différents, suscitent un intérêt considérable depuis quelques années, puisqu'ils pourraient être à l'origine de la prédisposition de certaines maladies. Parmi eux, certaines séquences d'ADN sont présentes en plusieurs copies au sein d'un génome d'un individu, et le nombre de copies varie en fonction des individus. Ces répétitions, qui sont appelées des « Variations du Nombre de Copies » (CNVs), peuvent affecter l'expression des gènes qu'elles contiennent ou de son voisinage par des mécanismes plus ou moins complexes. Ainsi, plusieurs CNVs ont été reliés à l'apparition de maladies.

Néanmoins, l'origine génétique d'une grande partie des traits humains reste inexpliquée, ce qu'on appelle couramment l'héritabilité manquante. En effet, le déchiffrage de notre information génétique est incomplet, certaines régions particulières du génome étant encore méconnues. Parmi elles, les longues séquences répétées en tandem, également appelées macrosatellites, sont difficiles à séquencer et à étudier. Les techniques de bioinformatique habituellement utilisées peinent à assembler correctement le puzzle à ces endroits du génome, car il est difficile de distinguer chaque répétition. Ces

macrosatellites sont constitués d'un nombre variable de répétitions d'une séquence d'ADN, à la suite les unes des autres et ayant toutes la même orientation, et forment donc un sous-type particulier de CNVs. Certains macrosatellites ont ainsi des conséquences fonctionnelles majeures (par exemple l'inactivation du chromosome X), ou ont été reliés à l'apparition de maladies génétiques (telles que le syndrome FSHD). La compréhension de ces macrosatellites, et également des autres zones d'ombre du génome, suscite encore de nombreux espoirs, notamment en génétique médicale.

Parmi les macrosatellites du génome humain décrits à ce jour, le macrosatellite RNU2 a été l'un des plus étudiés pendant les années 1990. Le nombre de répétitions a été caractérisé sur plus de 200 individus, et peut varier de 5 à plus de 63. Son unité répétée, longue de 6,1 kilobases, ne contient qu'une seule séquence codante de 200 paires de bases. Elle code pour un petit ARN nucléaire non codant, ce qui signifie qu'il ne sera pas traduit en protéine. Les petits ARN nucléaires (ARNsn) sont impliqués dans une étape clé de l'expression des gènes, que l'on appelle l'épissage et qui consiste à maturer les ARN messagers en excisant les introns (les régions non codantes). La majorité des ARN messagers de la cellule sont pris en charge par la machinerie d'épissage. Ce complexe est donc requis en grande quantité. Un des moyens pour la cellule d'obtenir une quantité suffisante d'ARNsn U2 est de contenir un nombre suffisamment élevé du gène U2, ce qu'on appelle une répétition de dosage. Bien que codant pour une molécule essentielle, le macrosatellite RNU2 ne fait pas partie du génome de référence, et n'a donc pas été étudié par les études à large échelle réalisées au cours de ces dix dernières années.

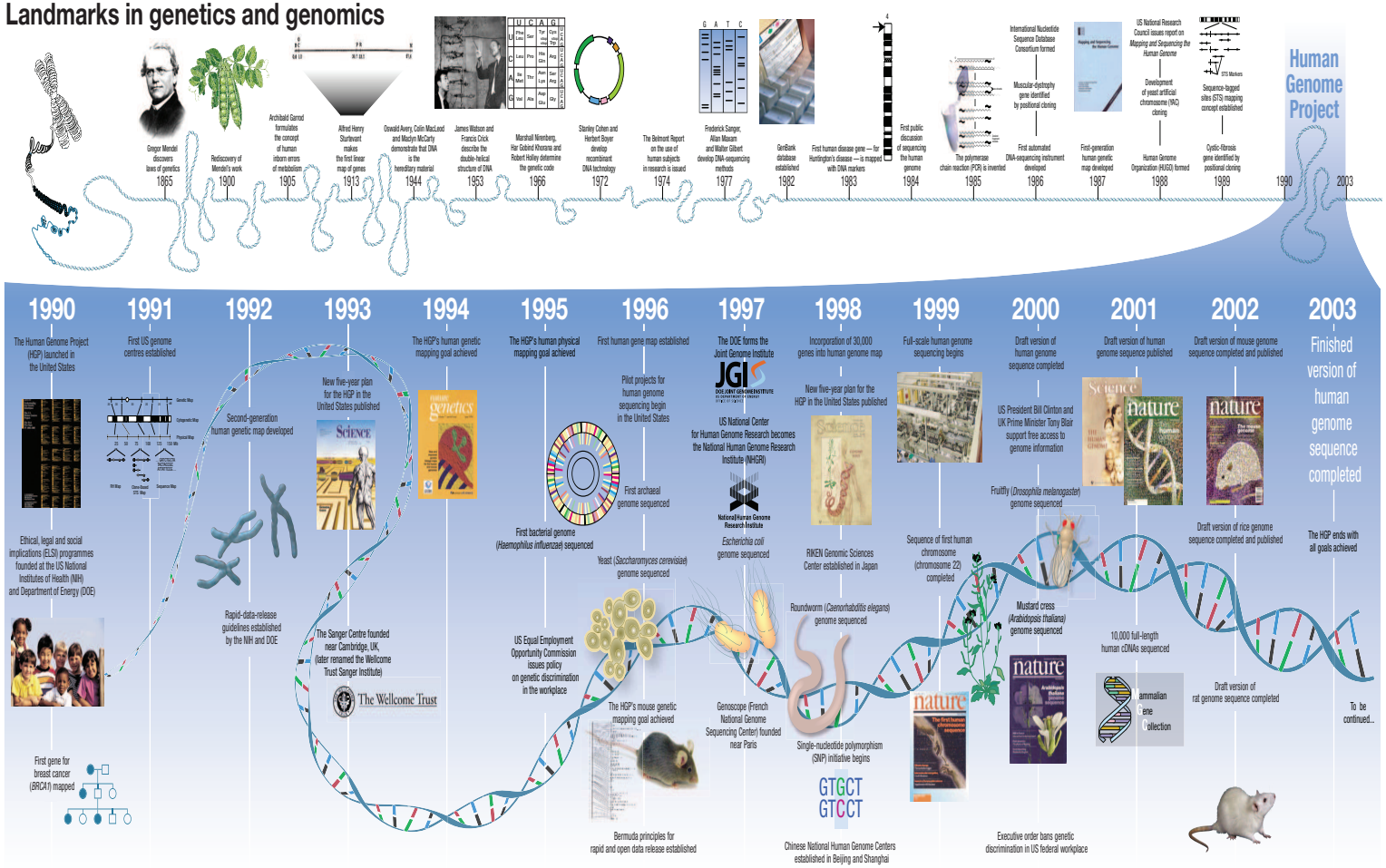
Cette absence est d'autant plus préjudiciable que ce locus pourrait être très utile en génétique humaine. Tout d'abord, le taux d'ARNsn U2 circulant dans le sang ou d'autres fluides corporelles serait un excellent biomarqueur pour certains cancers, tels que le cancer pancréatique ou le cancer colorectal. Ensuite, chez la souris, une dérégulation spatiale de l'expression d'une seule répétition a été impliquée dans une neurodégénérescence. Enfin, ce

macrosatellite se situe à quelques kilobases en amont d'un gène majeur de prédisposition au cancer du sein, le gène BRCA1, et pourrait ainsi être impliqué dans la prédisposition génétique à ce cancer.

En France, le cancer du sein est le cancer le plus fréquent chez la femme, et également la première cause de mortalité. Plusieurs facteurs de risque ont été identifiés, parmi lesquels le genre, l'âge, l'histoire familiale, les facteurs hormonaux et la vie reproductive. Actuellement, 5 à 10 % des cancers du sein sont diagnostiqués chez des femmes ayant des antécédents familiaux, et sont donc vraisemblablement dus à une prédisposition génétique. Un test génétique est proposé aux membres de ces familles de cancer du sein, afin d'identifier la mutation (i.e. l'anomalie) à l'origine de cette prédisposition. Lorsque cette mutation est identifiée, les femmes porteuses de la mutation dans la famille bénéficient d'un suivi médical adapté, ce qui permet d'intervenir très précocement dans le développement du cancer et donc de mieux le soigner.

Actuellement, aucune 'explication' génétique n'est trouvée pour 80 % des familles analysées par le test génétique. Mon équipe recherche de nouveaux mécanismes d'inactivation des gènes connus pour être impliqués dans l'apparition de cancers du sein. Parmi ces gènes, les deux gènes majeurs sont BRCA1 et BRCA2. Ce sont deux gènes suppresseurs de tumeurs, ce qui signifie qu'ils freinent la prolifération cellulaire. Ils agissent comme des garde-fous empêchant la dérive d'une cellule vers la malignité et perdent donc fréquemment leur fonction dans les cancers. Les protéines BRCA1 *et* BRCA2 *sont impliquées dans une multitude de processus cellulaires fondamentaux, tels que la réparation des dommages à l'ADN.*

Pendant ma thèse, je me suis intéressée uniquement à BRCA1. En effet, ce gène est localisé à quelques centaines de kilobases, une distance relativement faible, du macrosatellite RNU2. Ainsi, nous avons émis l'hypothèse que ce macrosatellite pourrait être impliqué dans la prédisposition génétique au cancer du sein.



DESIGN BY DARRYL LEJA
PEAS COURTESY J. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; SCIENCE COVERS COURTESY AAAS

Figure 1: Chronologie des principales découvertes en génétique et génomique.
Extrait de Collins et al., 2003

1 Les Variations Génétiques de grande taille du Génome Humain

1.1 Introduction : les variations génétiques humaines

1.1.1 La Séquence du génome humain

La séquence du génome humain, contenue dans 23 chromosomes, est composée de 3 milliards de nucléotides (A, T, C ou G). Ces bases azotées constituent l'ADN (acide désoxyribonucléique), le support universel de l'information génétique. Avery et ses collaborateurs sont les premiers à émettre cette hypothèse en 1944 grâce à leurs travaux sur la transformation bactérienne (Avery et al., 1944). Cette observation est par la suite confirmée par Hershey et Chase, grâce à leurs études sur l'infection de bactéries par le bactériophage T2 et l'utilisation d'isotopes radioactifs permettant de tracer les molécules (Hershey and Chase, 1952). Grâce à des études de chromatographie sur papier, Chargaff en 1950 décrit la composition en bases azotées du génome de différentes espèces (Chargaff et al., 1950), ce qui a permis trois ans plus tard la découverte de la structure en double hélice de l'ADN par Watson et Crick (Watson and Crick, 1953) (Figure 1). Les premiers outils d'étude du génome humain, également regroupés sous le terme de génie génétique, voient le jour au cours des années 1960 et 1970 : découverte des enzymes de restriction (Smith and Wilcox, 1970), découverte des ligases, mise au point de la technique d'électrophorèse, … Les premières techniques de séquençage du génome humain ont vu le jour au cours des années 1970 : d'une part la méthode de Sanger (Sanger et al., 1977), basée sur une synthèse enzymatique sélective, et d'autre part la technique de Maxam et Gilbert (Maxam and Gilbert, 1980), basée sur une dégradation chimique sélective.

Avant son séquençage complet, seules quelques régions dispersées dans le génome avaient été explorées, lors d'études longues et fastidieuses. Au cours des années 1990, de nombreux gènes impliqués dans l'apparition de maladies génétiques humaines ont

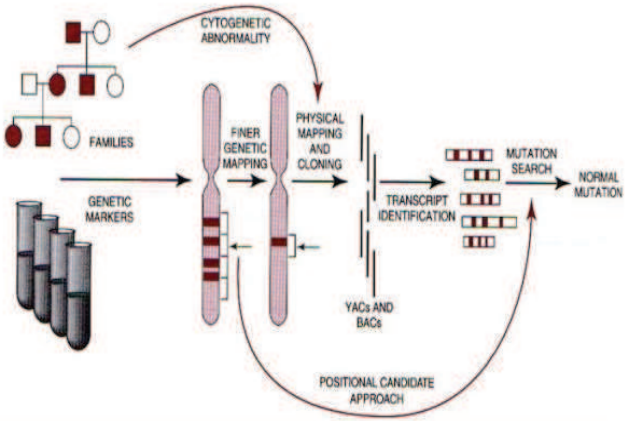


Figure 2 : Stratégie du Clonage Positionnel.
Cette stratégie permet d'identifier un gène d'intérêt à partir de la seule connaissance de sa localisation chromosomique, et est utilisée lorsqu'il n'existe aucun gène candidat évident. La première étape, dite de localisation primaire, consiste en une analyse de liaison classique (ségrégation dans les familles), suivie d'une cartographie fine de la région d'intérêt permettant de réduire l'intervalle critique contenant le gène. Pour cela, un certain nombre de marqueurs répartis à intervalles réguliers le long des chromosomes sont génotypés afin d'identifier ceux dont les allèles ségrègent spécifiquement avec le phénotype. Grâce à l'identification de clones BACs ou YACs chevauchant et couvrant la région d'intérêt, il est alors possible de mettre en évidence les séquences codantes contenues dans cette région. Parmi ces gènes est recherché celui présentant des mutations qui ségrègent spécifiquement chez les individus porteurs du caractère étudié.
D'après Molecular Lab.

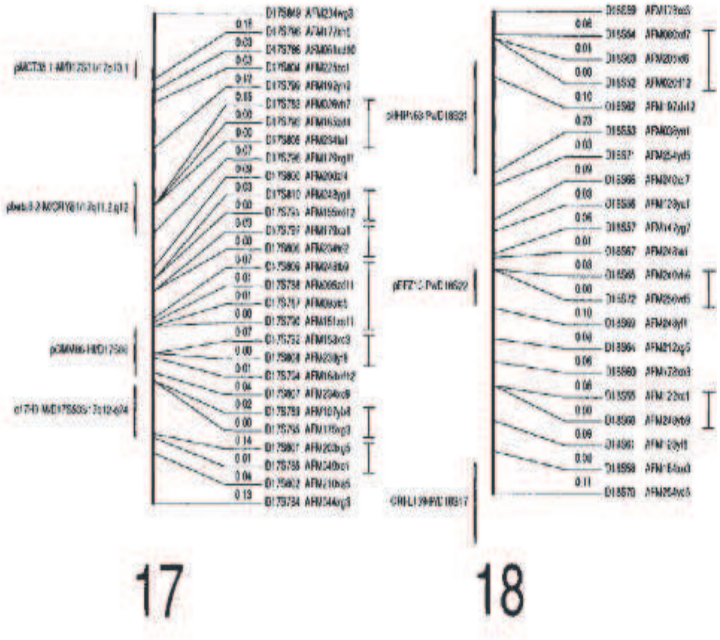


Figure 3 : Carte génétique du génome humain: exemples des chromosomes 17 et 18.
Extrait de l'article Weissenbach et al., 1992.

été identifiés par la stratégie du *clonage positionnel* (Figure 2). Cette stratégie repose sur l'utilisation de marqueurs polymorphes, tels que les polymorphismes de sites de restriction et les microsatellites. Ces derniers ont été découverts au cours des années 1970 et 1980 (Jeffreys et al., 1985; Kan and Dozy, 1978; Wyman and White, 1980). L'utilisation de ces marqueurs permettaient d'établir des cartes génétiques, c'est-à-dire de déterminer la position d'un gène d'intérêt sur un chromosome en fonction du taux de recombinaison génétique (les distances étant alors exprimées en centimorgan, ou cM). Cette première étape est alors suivie d'une cartographie plus fine de la région afin de réduire progressivement l'intervalle critique contenant le gène responsable du caractère étudié. Cette stratégie était extrêmement lourde à mettre en place.

En 1992 puis 1994, Weissenbach et ses collaborateurs finalisent la première cartographie génétique du génome humain au sein du laboratoire du Généthon, regroupant plus de 812 marqueurs (Gyapay et al., 1994; Weissenbach et al., 1992) (Figure 3). Une version finale regroupant plus de 5000 marqueurs fut publiée en 1996 (Dib et al., 1996). Ces cartes génétiques ont eu un impact considérable en permettant la localisation de nombreux gènes responsables de pathologies monogéniques, comme en atteste le nombre incroyable de citations dont elles ont fait l'objet (1992 : 1810 citations, 1994 : 1937 citations, 1996 : 2836 citations). Sans ces avancées majeures, le séquençage du génome humain n'aurait pu être réalisé aussi vite, et aussi tôt. La première carte physique couvrant la moitié du génome humain fut publiée par Daniel Cohen en 1992 (Bellanné-Chantelot et al., 1992). Cette carte établit la position physique précise d'un grand nombre de marqueurs, et les distances sont cette fois exprimées en paires de bases. Pour réaliser cette prouesse, l'équipe du Dr Cohen a utilisé une technique d'empreinte (ou fingerprinting) de chromosomes artificiels de levure (Yeast Artifical Chromosomes, ou YACs), en étudiant le chevauchement de chacun de ces clones grâce à des séquences STS (sequence tag site).

Le projet de séquençage du génome humain a été amorcé en 1985, sous l'impulsion de trois chercheurs (Renato Dulbecco, Robert Sinsheimer et Charles DeLisi) (Sinsheimer, 1989). C'est le 12 Février 2001 que l'annonce officielle du séquençage de 95 % du génome humain a été faite, d'un côté par le Consortium international de recherche publique (Human Genome Project Consortium, ou HGP Consortium)

Figure 4 : La « Une » du journal Nature du 14 Mars 1996 sur la carte génétique du génome humain publiée par le Généthon, et celles du journal Nature du 15 Février 2001 et du journal Science du 16 Février 2001 sur le séquençage du génome humain.
D'après Dib et al., 1996; IHGSC et al., 2001; Venter et al., 2001.

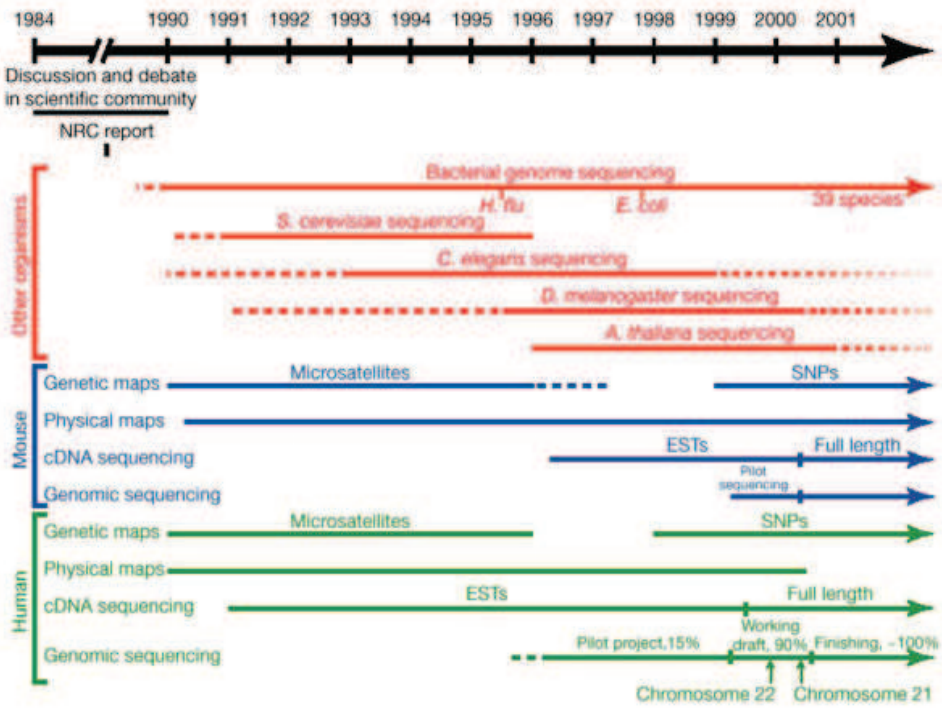


Figure 5 : Chronologie des analyses génomiques à grande échelle, ayant conduit au séquençage du génome humain.
Sont présentés plusieurs composants des travaux sur des organismes modèles non-vertébrés (rouge), la souris (bleu) et l'homme (vert) depuis les années 1990. Le projet de séquençage du génome humain (ou Human Genome Organization, HUGO) a été lancé à la fin de l'année 1990. Après 1995, le projet avance considérablement grâce à la construction de cartes génétiques et physiques du génome de la souris, et également grâce aux séquençages du génome de la levure *S. cerevisae* et du nématode *C. elegans*.
SNPs: Single Nucleotide Polymorphisms
ESTs: expressed sequence tags.
Extrait de IHGSC et al., 2001.

(Lander et al., 2001), de l'autre par leur concurrent privé Celera Genomics Corp (Venter et al., 2001) (Figures 4 & 5). Pour arriver à ce résultat, le HGP Consortium a utilisé une stratégie plus chronophage que celle de la société Celera, en ajoutant une étape supplémentaire de cartographie des clones BAC (chromosomes artificiels bactériens) avant d'en sélectionner une sous-partie qui sera séquencée par séquençage global aléatoire (SGA), plus connu sous le nom de whole-genome shotgun sequencing, la stratégie directement utilisée par Celera (Figure 6). Néanmoins, l'obtention des résultats par C. Venter a pu être si rapide grâce à une utilisation massive de séquences de clones BAC assemblées, produites par le HGP et déposées dans des banques de données.

Bien qu'inachevée et brute, cette première version du génome humain a permis de révolutionner la recherche en génétique. Tout d'abord par la découverte du faible nombre de gènes que possède le génome humain : de 30 000 à 35 000, contrairement aux 120 000 parfois attendus (Liang et al., 2000). Enfin, le séquençage de 5 individus d'origine différente par la société Celera a permis d'identifier 0,1 % de variations génétiques entre individus (Venter et al., 2001). Trois ans plus tard, le Consortium international public publiera une séquence plus complète, couvrant plus de 99 % des séquences euchromatiques, c'est-à-dire les séquences transcrites du génome humain, mais comportant encore plus de 340 trous (ou gaps) (International Human Genome Sequencing Consortium, 2004).

L'amélioration des techniques de séquençage, par la robotisation et l'analyse en parallèle des séquences, et l'introduction des puces à ADN a permis de faire des avancées majeures dans le domaine de la génétique humaine, et de mettre en évidence l'importance des variations interindividuelles (Ku et al., 2010). Actuellement, environ 2500 individus ont été séquencés ou génotypés par le projet 1000 Génomes (1000 Genomes Project Consortium et al., 2012) (http://www.1000genomes.org/). Un des problèmes majeurs reste l'interprétation du nombre croissant de variants ainsi découverts.

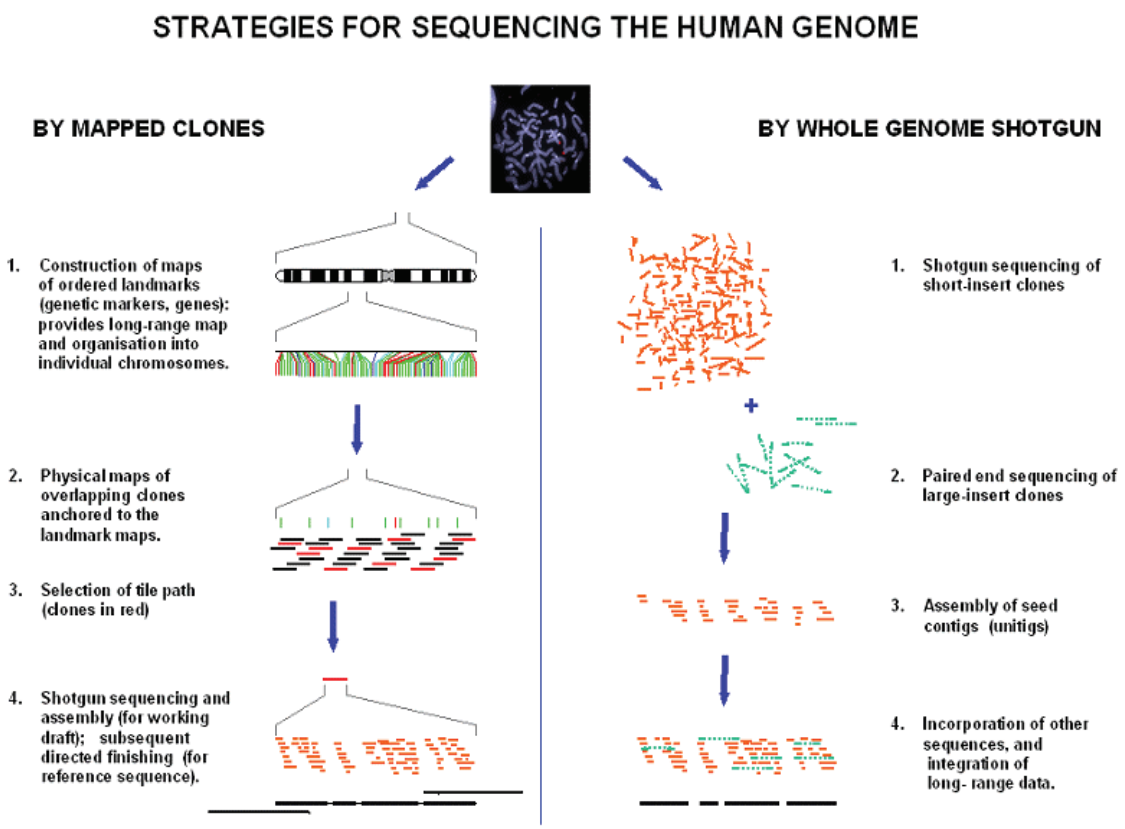


Figure 6 : Les différentes stratégies utilisées pour séquencer le génome humain.
A. Par séquençage de clones BACs préalalement mappés (HGP Consortium).
B. Par Whole-Genome Shotgun Sequencing (Celera Genomics).
D'après https://gene-tics.wikispaces.com/20.+DNA+Technology+and+Genomics

1.1.2 Les Définitions des Variations Génétiques

Au cours des 50 dernières années, chaque avancée technologique dans le monde de la génomique a été accompagnée par la découverte d'une nouvelle catégorie de variants. Ces découvertes progressives ont compliqué la mise en place d'une classification universelle et homogène. Ainsi, mis à part pour les polymorphismes d'un seul nucléotide (Single Nucleotide Polymorphism, ou SNP), il est difficile de trouver un consensus dans la littérature sur la définition, la taille de leur unité de base, la taille totale du locus, ou encore leur importance au sein du génome humain. Ainsi, les courtes insertions et délétions peuvent être définies comme des séquences de 1 à 10 000 paires de bases (pb) (Mullaney et al., 2010) ou comme inférieures à 50 pb (Montgomery et al., 2013), les modifications impliquant plus de 50 pb étant alors regroupées sous l'appellation de variants structuraux (Alkan et al., 2011).

Pour les séquences répétées en tandem, la classification originelle de Jeffreys permet de distinguer les microsatellites (1-5 pb) (Jeffreys et al., 1985; Tautz and Schlötterer, 1994), les minisatellites (6-100 pb) (Bois et al., 1998; Jeffreys et al., 1995; Tautz, 1993; Vergnaud and Denoeud, 2000), les midisatellites (101-400 pb) et les macrosatellites (> 400 pb) (Jeffreys et al., 1985). Depuis, la taille du motif de base des minisatellites a fait l'objet de définitions variées : 6 ou 10 nucléotides pour la taille minimale du motif, et de 60 à 100 nucléotides pour la taille maximale. Par ailleurs, la définition des microsatellites se recoupe avec celle, plus récente, des petites répétitions en tandem (Short Tandem Repeats, STRs, ou Simple Sequence Repeats, SSRs). Les microsatellites et minisatellites sont communément regroupés sous l'appellation de répétitions en tandem variables en nombre (Variable Number of Tandem Repeats, ou VNTR).

Des études postérieures ont reconsidéré la taille de l'unité des macrosatellites de 101 à 999 pb (Gondo et al., 1998), et ainsi conduit à l'oubli du terme *midisatellite*, et défini une nouvelle classe de variants, les mégasatellites (> 1000 pb) (Gondo et al., 1998; Saitoh et al., 2000). Récemment, le terme *mégasatellite* a également été abandonné, au profit du terme *macrosatellite*, ces séquences macrosatellites étant

Tableau 1 : Classification des variations génétiques du génome humain, en fonction de la taille de leur motif de bases.

Type de variation		Taille du motif	Taille du locus
SNP		1 pb	
Courtes insertions/délétions	Micro-insertion/délétion	1 – 9 pb	0.5 – 15 kb
	Mini-insertion/délétion	10 – 99 pb	10 – 1000 bp
	Midi-Insertion/délétion	100-999 bp	
CNV (Copy Number Variation)		> 1 kb	< 5 Mb
Variant microscopique		ND	> 5 Mb

Tableau 2 : Classification des variations génétiques répétées en tandem du génome humain, en fonction de la taille de leur motif de bases.

Type de variation	Taille du motif
Microsatellite	2 – 9 pb
Minisatellite	10 – 99 pb
Midisatellite	100 – 999 pb
Macrosatellite	> 1 kb

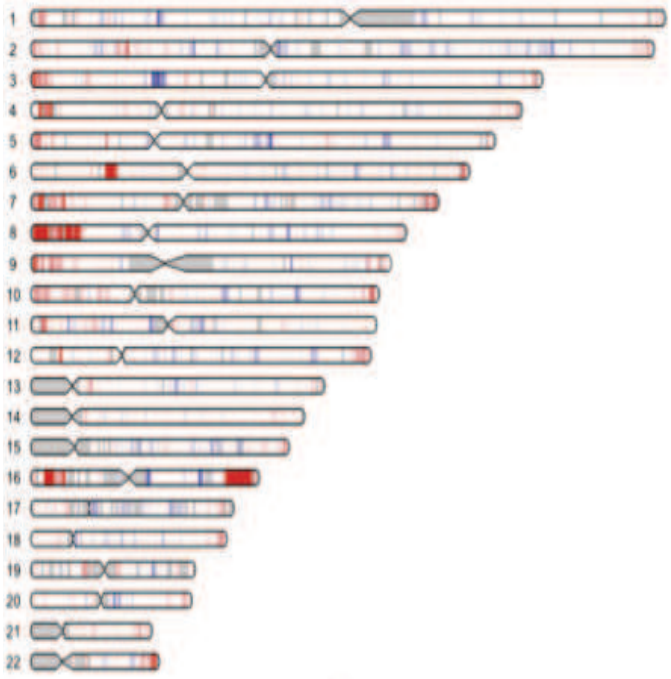


Figure 7 : Distribution de la densité en SNP (Single Nucleotide Polymorphism) sur les autosomes.
Rouge: haute densité, Bleu: faible densité (sur 1 Mb).
Extrait de 1000 Genomes Project Consortium et al. , 2010.

définies comme englobant plusieurs dizaines de kilobases (kb), plutôt que sur la taille de leur motif de bases.

Il me semble important, avant de lister et détailler ces variations, de proposer des définitions claires qui seront employées tout au long de ce manuscrit. Par analogie avec les séquences répétées en tandem, les courtes insertions et délétions (short indels) regroupent les micro-indels (1 – 9 pb), les mini-indels (10 – 99 pb) et les midi-indels (100 – 999 pb) (Tableau 1). Les CNVs (Copy Number Variants) sont définis comme des insertions et ou délétions d'une séquence d'au moins 1 kb. Lorsque la taille totale du locus variant en nombre de copies est supérieure à 3 mégabases (Mb), on parle alors de variant microscopique : il peut s'agir d'une insertion ou délétion d'une grande séquence unique d'ADN, ou d'évènements multiples d'insertions ou délétions d'une séquence de taille moyenne. Cette même classification sera utilisée pour les répétitions en tandem, qui sont un cas particulier de chaque catégorie de variations génétiques de par leur structure très caractéristique et leur mode de génération (Tableau 2).

1.1.2.1 Les Polymorphismes d'un seul nucléotide (SNP)

Plusieurs milliers de polymorphismes d'un seul nucléotide ont été mis en évidence par l'analyse moléculaire par Southern blot des polymorphismes de longueur des fragments de restriction (Restriction Fragment Length Polymorphism, ou RFLP) (Narayanan, 1991). On estime actuellement que le génome humain comporte plus de 10 à 15 millions de SNPs dont la fréquence de l'allèle mineur est supérieur à 1 % (Figure 7) (International HapMap Consortium, 2003, 2005; Kruglyak and Nickerson, 2001; Sachidanandam et al., 2001). Ainsi, on trouve un SNP tous les 100 à 300 paires de base.

Grâce au séquençage d'un millier de génomes, de nombreux SNPs ont été identifiés, localisés et génotypés dans différentes populations humaines, et sont référencés par le projet HapMap. L'objectif du projet international HapMap est de développer une carte haplotypique du génome humain, décrivant les variations génétiques communes (http://hapmap.ncbi.nlm.nih.gov/). Un autre objectif de ce projet

Figure 8 : Diagramme des SNPs associés à un phénotype humain avec une p-value ≤ $5{,}0 \times 10^{-8}$, jusqu'au mois de Mars 2013.
D'après le catalogue http://www.genome.gov/gwastudies

Table 1.
INDEL discovery in human populations and personal genomes

Study	Year	No. of INDELs	Individual(s)	Method	INDEL size range (bp)	Validation study[a]	Rate (%)
Human populations							
Mullikin *et al.* (23)	2000	NR[b]	31	ABI trace mapping	NR	ND	N/A
Dawson *et al.* (24)	2001	2180	9[c]	BAC overlap	NR	PCR/RFLP/invader	92
Weber *et al.* (13)	2002	2000	NR	Various	2–55	PCR	58
Bhangale *et al.* (14)	2005	2393	137	PCR/sequencing	1–543	ND	N/A
Bhangale *et al.* (25)	2006	1126	ENCODE	PCR/sequencing	1 to 30	Manual inspection	100
Mills *et al.* (15)	2006	415 436	36	ABI trace mapping	1–9989	PCR	97
Kidd *et al.* (22)	2008	796 273	8	ABI trace mapping	1–100	ND[d]	96
R.E. Mills *et al.* (submitted for publication)	2010	1.96 million	79[e]	ABI trace mapping	1–10 000	PCR[f]	97
Personal human genomes							
Levy *et al.* (1)	2007	823 396	Venter	ABI	1–82 711	PCR	84–100[g]
Wheeler *et al.* (2)	2008	222 718	Watson	454	2–38 896	PCR	70[h]
Wang *et al.* (3)	2008	135 262	Han Chinese	Illumina/SOAP	1–3	PCR	90–100
Bentley *et al.* (4)	2008	400 000[i]	Yoruban	Illumina/ELAND/MAQ	1–16	ND	N/A
Ley *et al.* (5)	2008	726[h]	AML	Illumina/Custom	1–30	PCR	93[j]
Ahn *et al.* (7)	2009	342 965	Korean (SJK)	Illumina/MAQ	1–26	PCR	100[h,k]
Kim *et al.* (6)	2009	170 202	Korean (AK1)	Illumina/Alpheus	1–29	Re-sequencing	100
Schuster *et al.* (8)	2010	NR	African	N/A	N/A	N/A	N/A

Figure 9 : Nombre d'indels découverts dans les populations humaines et les génomes personnels.
D'après la table 1 de Mullaney et al., 2010.

est d'identifier des SNPs marqueurs (ou tagSNPs) situés dans des blocs de déséquilibre de liaison, permettant ainsi de diminuer le nombre de variants à génotyper.

La cartographie de ces SNPs a permis de tester leur implication dans l'apparition de phénotypes particuliers ou maladies complexes, grâce aux nouvelles technologies de génotypage à grande échelle (genome-wide association study, ou GWAS). Ces études pan-génomiques ont d'ores et déjà mis en évidence un nombre considérable de loci associés à l'apparition de maladies communes (Figure 8) (Hindorff et al., 2009).

1.1.2.2 Les courtes insertions et délétions (ou short indels)

Bien que très abondantes au sein du génome humain, les petites insertions et délétions, plus communément appelées indels (ou short indels), ont reçu moins d'attention que les SNPs ou les variants de plus grande taille. Ces variations, couvrant de 1 à 999 pb, sont particulièrement difficiles à détecter, valider et génotyper (Mullaney et al., 2010). Les premiers efforts pour les cartographier ont été conduits sur le chromosome 22, en comparant la couverture de différents clones BACs adjacents (Dawson et al., 2001; Mullikin et al., 2000). Le développement d'outils personnalisés, tels que l'algorithme PolyPhred, a permis d'automatiser la détection de cette classe de variations génétiques, à partir de données de séquençage nouvelle-génération par exemple (Bhangale et al., 2006). Aujourd'hui, environ 1 à 2 million de short indels ont été identifiés dans les populations humaines modernes (Figure 9) (Mills et al., 2006; Montgomery et al., 2013). A partir de données de séquençage de 179 individus, Montgomery et ses collaborateurs ont identifié 1,6 million d'indels, dont une grande partie (23,6 %) sont des courtes répétitions en tandem.

On estime que le glissement de la polymérase lors de la réplication, un mécanisme mutationnel bien connu, est à l'origine de ¾ des indels (Figure 10) (Levinson and Gutman, 1987; Montgomery et al., 2013; Streisinger et al., 1966; Taylor et al., 2004). Un autre mécanisme, décrit pour expliquer des variants structuraux de plus grande taille, pourrait également être à l'origine de ces polymorphismes : *l'interruption de la fourche de réplication et commutation de la matrice* (fork stalling

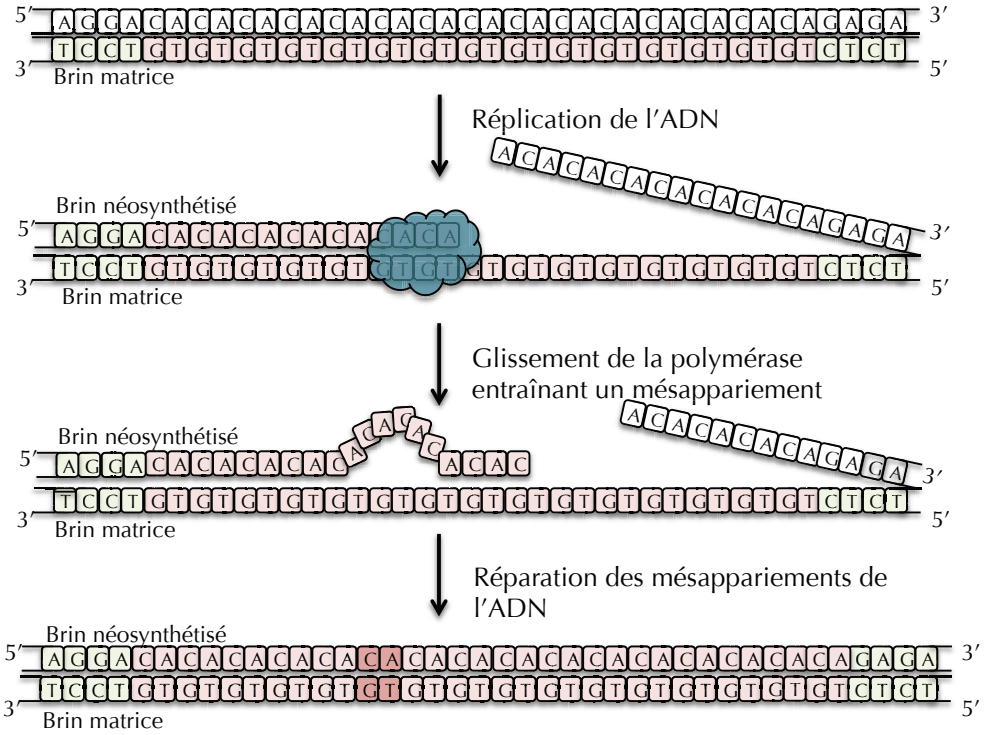


Figure 10 : Mécanisme de glissement de la polymérase lors de la réplication de l'ADN, entrainant un changement du nombre de répétitions d'une séquence microsatellite.

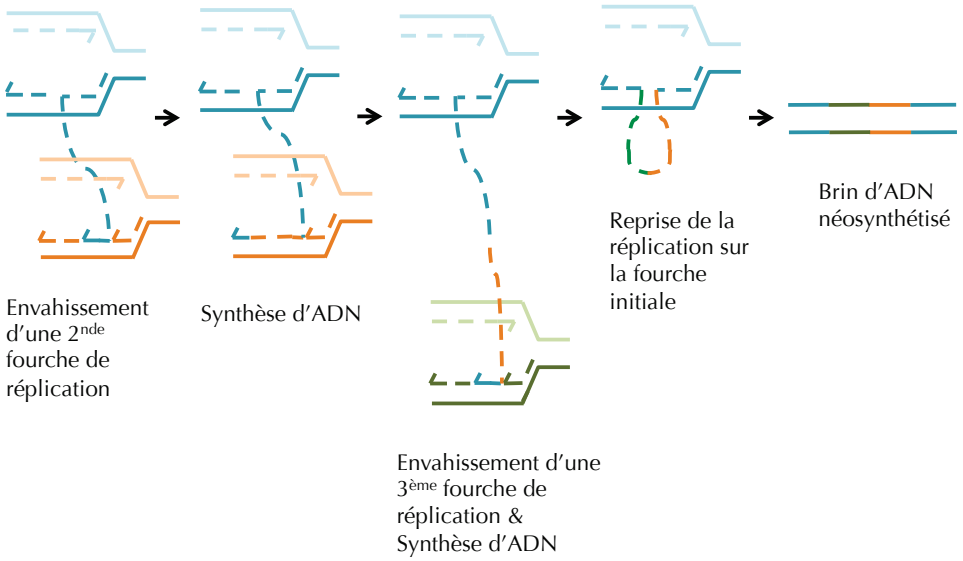


Figure 11 : Modèle proposé du mécanisme FoSTeS (Fork Stalling and Template Switching)/MMBIR (Microhomology-Mediated Break-Induced Replication).
Adapté de Lee et al., 2007.

and template switching, ou FoSTeS) (Figure 11). Ce processus est basé sur un mécanisme permettant de relancer la réplication après un blocage de la fourche de réplication *via* l'invasion d'un ADN double brin par l'extrémité 3' libre. Lorsque le changement de matrice se produit avec peu d'homologie de séquence, on parle d'une réplication induite par cassure médiée par une microhomologie (microhomology-mediated break-induced replication, ou MMBIR) (Hastings et al., 2009; Lee et al., 2007). Plusieurs changements de matrice peuvent ainsi survenir, à l'origine de réarrangements multiples et complexes.

1.1.2.2.1 Les séquences microsatellites

Les séquences microsatellites, également appelées STRs ou SSRs, sont des séquences d'ADN formées par la répétition en tandem d'un motif de 1 à 10 nucléotides. Une séquence microsatellite peut couvrir de 80 à 400 pb, et contient en moyenne de 5 à 40 répétitions. Elles ont été découvertes au cours des années 1980, sont particulièrement abondantes chez les eucaryotes et sont présentes sur l'ensemble du génome humain, plus particulièrement au niveau des séquences non-codantes (Dib et al., 1996). On estime que plus d'un million de séquences microsatellites sont réparties sur l'ensemble du génome, ce qui équivaut à plus de 3 % du génome humain (Lander et al., 2001).

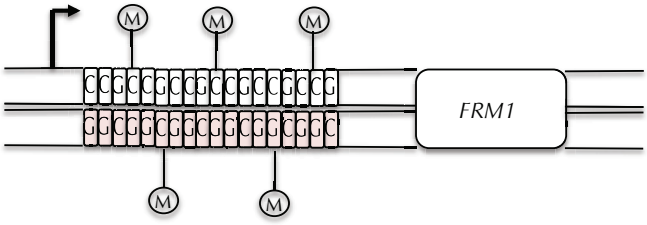
Elles font partie des séquences les plus polymorphes du génome humain de par leur hypermutabilité (Weber, 1990; Weber and Wong, 1993). La plupart de ces séquences microsatellites présentent un fort taux d'hétérozygotie, variant entre 50 % et 80 % (Gulcher, 2012). Elles sont ainsi utilisées comme marqueur moléculaire, puisque facilement amplifiables et caractérisables par PCR (Polymerase Chain Reaction). Une cartographie à haute résolution de plus de 5000 marqueurs microsatellites a été publiée en 2002 par la compagnie deCODE Genetics (Kong et al., 2002).

Parmi ces séquences microsatellites, plusieurs répétitions de trinucléotides, particulièrement des répétitions de polyglutamine (polyQ repeats) localisées au sein de gènes, ont été reliées à l'apparition de maladies humaines, plus communément

Tableau 3 : Caractéristiques génétiques des maladies à expansion de triplets.
AR : récepteur aux androgènes ; CACNL 1A4 : canal calcique ; DM : dystrophie myotonique ; DMPK : myotonin protein kinase ; FMR : protéine de liaison à l'ARN ; SCA : ataxie spinocérébelleuse ; P : paternel ; M : maternel.
Adapté de Sablonnière et al., 2000.

Maladie	Triplet	Transmission	Chromosome	Gène	Nombre de répétitions	
					Allèle Normal	Allèle Muté
Amyotrophie spino-bulbaire (ou maladie de Kennedy)	CAG	Récessif lié à l'X	Xq13-q21	*AR*	7-34	36-68
Maladie de Huntington	CAG	Autosomal dominant	4p16.3	*Huntingtine*	10-35	37-121
Atrophie dentarubro-pallidoluysienne	CAG	Autosomal dominant	12p13	*Atrophine*	5-35	49-85
Ataxie SCA1	CAG	Autosomal dominant	6p23	*Ataxine 1*	6-39	43-82
Ataxie SCA2	CAG	Autosomal dominant	12q24.1	*Ataxine 2*	14-31	35-59
Ataxie SCA3	CAG	Autosomal dominant	14q32.1	*Ataxine 3*	13-44	55-84
Ataxie SCA6	CAG	Autosomal dominant	19q13	*CACNL1A4*	4-16	21-30
Ataxie SCA7	CAG	Autosomal dominant	3p12.13	*SCA7*	4-35	38-220
Ataxie SCA8	CTG	Autosomal dominant	13q21	*SCA8*	16-37	107-127
Syndrome de l'X fragile	CGG	Dominant lié à l'X	Xq27.3	*FMR1*	5-52	230-1000
Retard mental de type FRAXE	GCC	Dominant lié à l'X	Xq28	*FMR2*	7-35	230-750
Dystrophie myotonique	CTG	Autosomal dominant	19q13.3	*DMPK*	5-35	50-2000
Ataxie de Friedreich	GAA	Autosomal récessif	9q13	*Frataxine*	6-22	200-1700

6 à 60 répétitions du triplet CGG -> Transcription du gène *FRM1*



200 répétitions du triplet CGG -> Hyperméthylation du locus CpG
-> Aucune transcription du gène *FRM1*

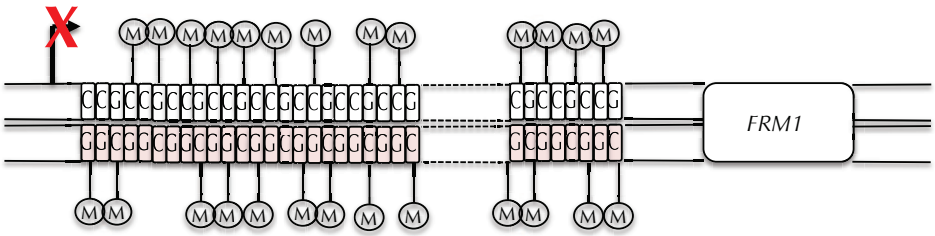


Figure 12 : Mécanisme moléculaire du Syndrome du X fragile.

appelées *maladies à triplets*, telles que le syndrome du X fragile ou bien la maladie de Huntington (Tableau 3) (Gatchel and Zoghbi, 2005; Mandel, 1997). Le nombre de copies de ces microsatellites, variable mais limité au niveau des allèles de la population générale, augmente considérablement dans les allèles mutés. Ces mutations ont été dénommées « mutations dynamiques », puisqu'à partir d'un certain seuil de répétitions, ces loci subissent une expansion dynamique instable du nombre de répétitions par différents mécanismes mutationnels (Richards and Sutherland, 1992). Ce mécanisme mutationnel a été identifié comme responsable d'un nombre croissant de maladies affectant le système nerveux de façon prédominante ou d'expression multisystémique avec une atteinte neuromusculaire majeure (Reddy and Housman, 1997; Robitaille et al., 1997). Dans la plupart de ces maladies, l'augmentation de la taille de l'expansion observée lors de la transmission de la mutation à la descendance s'accompagne d'un âge de début plus précoce de la maladie (phénomène d'anticipation).

Dans le cas du syndrome du X fragile, on observe une augmentation du nombre de répétitions du motif CGG, au sein du premier intron du gène de la frataxin *FMR1*. Lorsque le nombre de répétitions dépasse un seuil critique, le profil de méthylation du promoteur du gène change, entraînant une modification de l'expression du gène (Figure 12) (Bardoni and Mandel, 2002; Warren et al., 1987).

1.1.2.2.2 Les séquences minisatellites

Les séquences minisatellites ont été identifiées en 1985 (Jeffreys et al., 1985). Ce sont des séquences répétées en tandem d'un motif formé de 10 à 100 nucléotides. Un allèle peut contenir jusqu'à 1000 répétitions, une séquence minisatellite peut donc s'étendre de 1 à 10 kb. Présentant un très fort taux de polymorphisme, elles ont été largement utilisées comme marqueurs moléculaires dans des analyses de liaison ou pour des expériences d'empreinte génétique (DNA fingerprinting). Elles ont également été utilisées pour des tests de paternité ou par la police scientifique pour l'identification d'un suspect (Silver, 1989). Depuis, elles ont été délaissées au profit des SNPs et des

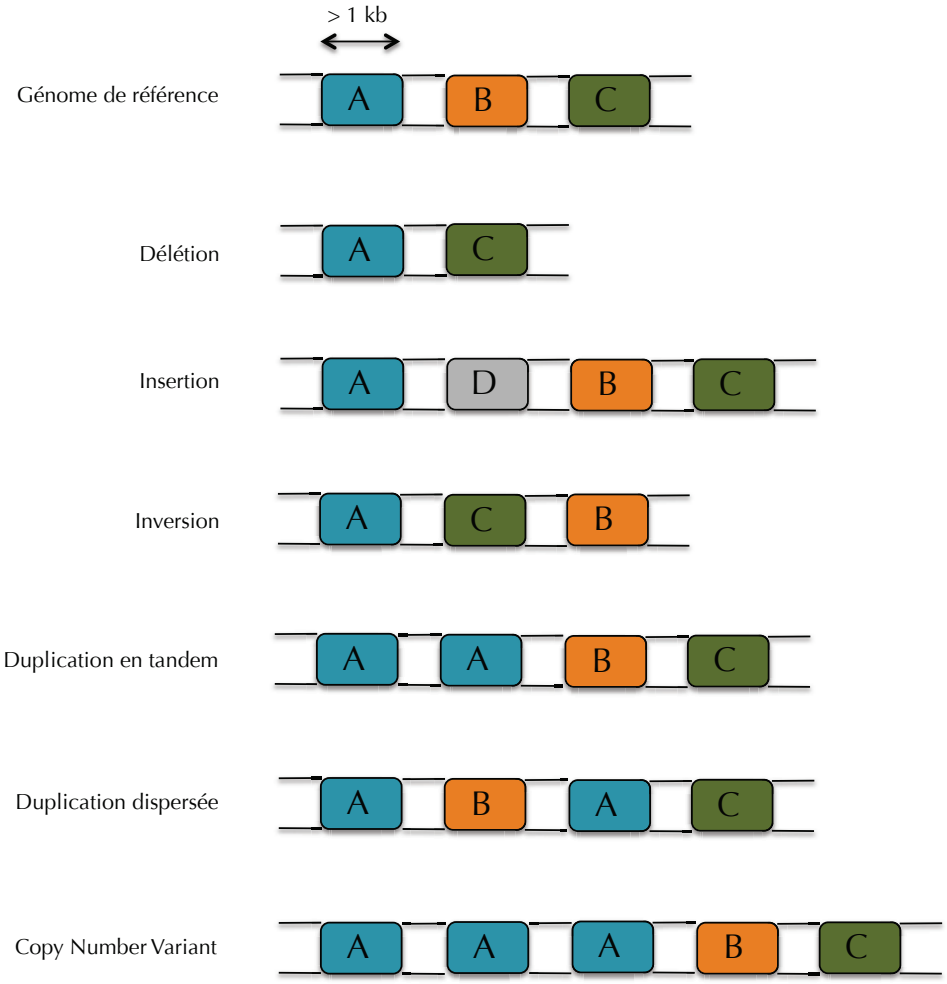


Figure 13 : Les différents types de variants structuraux de grande taille.
Adapté de Baker et al., 2012.

variants structuraux de plus grande taille, bien qu'elles fassent régulièrement l'objet d'études portant sur leur influence sur l'évolution des génomes (López-Flores and Garrido-Ramos, 2012).

1.2 Les différentes catégories de variants structuraux de grande taille

Les variants structuraux de grande taille, définis comme des insertions et/ou délétions d'une séquence d'ADN supérieure à 1 kb, comprennent des formes non balancées de variations (les larges insertions, duplications et/ou délétions, les longues répétitions en tandem), ainsi que des formes balancées (les inversions et les translocations) (Figure 13) (Feuk et al., 2006). Les variants structuraux sont répertoriés dans la base de données des variants génomiques (DGV) et dans la base de données des Variations Structurales Humaines (dbVar) (http://www.ncbi.nlm.nih.gov/dbvar).

Initialement, le terme CNV désignait les séquence d'ADN supérieure à 1kb dont le nombre de copies varie par rapport à un génome de référence ne contenant que 2 copies (1 par allèle) (Feuk et al., 2006). Cette définition excluait de fait les répétitions en tandem. Actuellement, ce terme est utilisé pour désigner toute variation quantitative du génome, incluant aussi bien les duplications et délétions que les répétitions en tandem. Dans ce manuscrit, nous considérerons que dans la plupart des études le terme CNV englobe également les macrosatellites. Cependant, elles représentent un exemple bien particulier de CNV, et feront donc l'objet d'un chapitre plus détaillé.

1.2.1 Les variants structuraux non-balancés : les CNVs (Copy Number Variations)

Ces variations structurales non balancées se différencient des petits évènements d'insertions et/ou de délétions (short indels) par la taille plus élevée de la séquence

variant en nombre (Feuk et al., 2006; Freeman et al., 2006). Lorsque l'évènement d'insertion ou de délétion est présent chez plus de 1 % de la population, la séquence CNV est référencée comme un polymorphisme de nombre de copies (Copy Number Polymorphism, ou CNP). Lorsque le CNV est retrouvé chez moins de 1 % de la population, on parle de variant privé.

1.2.2 Les variants structuraux balancés : inversions & translocations

Une translocation est un échange de fragments chromosomiques entre deux chromosomes non homologues. Une inversion se caractérise quant à elle par le changement d'orientation d'une séquence d'ADN au même locus. Ces réarrangements ne peuvent être identifiés par des techniques classiques d'hybridation sur puce (aCGH : Array Comparative Genome Hybridization), mais peuvent l'être par la technique de séquençage « paired-end » (ou mate-pair). Une étude de 2008 conduite sur 8 génomes diploïdes humains a permis ainsi d'identifier 224 inversions (Kidd et al., 2008). Ces réarrangements peuvent également être caractérisés par PCR (Flores et al., 2007). Actuellement, plus de 800 événements d'inversions sont répertoriés dans la base de données DGV (Database of Genomic Variants).

1.2.3 Les variants structuraux microscopiques

Ces variants structuraux microscopiques désignent des réarrangements dont la taille est supérieure à 5 Mb, et qui peuvent donc être détectés avec des microscopes optiques (Gripenberg, 1964; Tjio and Nichols, 1985). Ces variants peuvent provenir de grands réarrangements au sein d'un chromosome modifiant considérablement la taille de ce dernier, mais ils peuvent également modifier l'intégralité d'un chromosome dans le cas de l'aneuploïdie ou lors d'identification d'un marqueur chromosomique surnuméraire (MAR). Un marqueur chromosomique surnuméraire est un chromosome additionnel de structure anormale pouvant dériver de tous les chromosomes humains. Dans le cas de l'aneuploïdie, la cellule comporte un nombre anormal de chromosomes

dû à une perte (monosomie) ou gain (trisomie, tétrasomie, pentasomie) de certains chromosomes. L'aneuploïdie est une cause courante de maladies génétiques, comme dans le cas de la trisomie 21 (Hassold et al., 1996; Yang et al., 2002).

1.3 Techniques d'identification et importance au sein du génome humain

La découverte et la compréhension des variants structuraux de grande taille sont plus récentes et ont été permises grâce aux avancées technologiques du séquençage à haut-débit et de l'hybridation comparative sur puces. En effet, les SNPs étaient encore considérés il y a une dizaine d'années comme les principaux contributeurs à la variation génétique humaine.

Avant 2004, la plupart des CNVs étaient étudiés grâce à des techniques de cytogénétique classique (Buysse et al., 2009). Le caryotype standard, réalisé sur chromosomes en métaphase, permet de détecter des réarrangements de grande taille (excédant 5 Mb). L'hybridation fluorescente *in situ* (fluorescent *in situ* hybridization, ou FISH), parce qu'elle augmente considérablement la résolution, permet de détecter des évènements dont la taille est supérieure à 100 kb (Bauman et al., 1980). Plus récemment, la technique de FISH sur fibres d'ADN étirées (fiber-FISH) a permis de détecter des changements sub-microscopiques (Bensimon et al., 1994; Florijn et al., 1995). Ces techniques sont encore utilisées pour valider les résultats obtenus sur puces, mais également pour l'étude de translocations balancées (non détectables par aCGH) et de réarrangements plus complexes (répétitions en tandem, ...). D'autres techniques locus-spécifiques sont également utilisées : le Southern Blot, l'électrophorèse en champs pulsé (Pulse Field Gel Electrophoresis, ou PFGE), et plus récemment la qPCR (quantitative PCR). Malheureusement, ces techniques sont souvent lourdes et difficiles à mettre en place, ce qui limite leur utilisation. La plupart des études se concentraient alors sur des gènes spécifiques liés à des pathologies, par exemple les gènes codant pour les globines α et β impliqués dans l'α- et la β-thalassémie, ou bien sur des

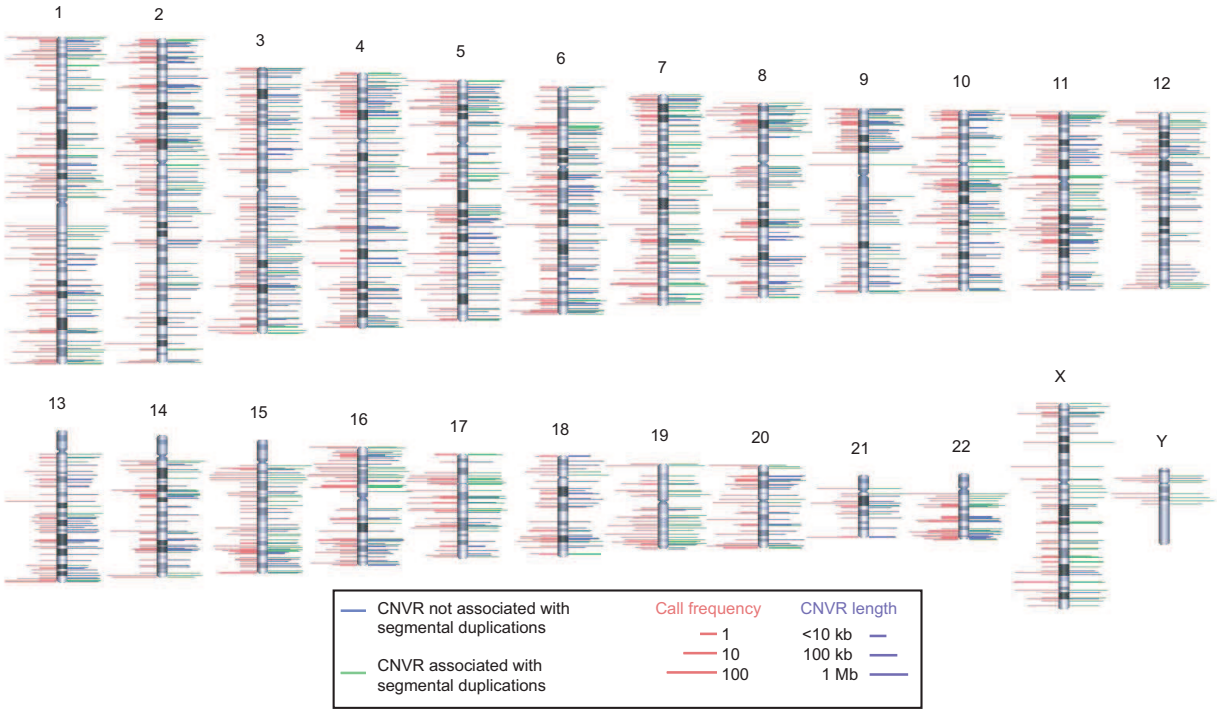


Figure 14 : Distribution sur le génome humain des 1447 régions variable en nombre de copies (CNVR).
Ligne verte: CNVR associée avec une duplication segmentaire. Ligne bleue: CNVR non associée à une duplication segmentaire. Longueur des lignes à droite: fréquence de détection de la CNVR.
Extrait de Redon et al., 2006.

variations structurales liées à l'apparition d'une maladie, telles que des délétions de la région 15q11-q13 impliquées soit dans le syndrome d'Angelman soit dans le syndrome de Prader-Willi (Blunt et al., 1994; Campbell et al., 1990; Gilles et al., 2000; Hollox et al., 2003; Kulski et al., 2002; Riley et al., 2002).

Les techniques moléculaires plus récentes ont permis de passer d'une étude locus-par-locus à une étude des changements au niveau du génome complet (Kallioniemi et al., 1992). La technique la plus utilisée est l'hybridation de génome complet sur puces, permettant de détecter des variations de quelques kb (Carter, 2007). En 2004, deux études sur puces ont mis en évidence un nombre important de CNVs au sein du génome humain de sujets sains (Iafrate et al., 2004; Sebat et al., 2004). 476 CNVs ont été identifiés, dont certains affectant des gènes jouant un rôle majeur dans des fonctions biologiques essentielles (métabolisme, défense contre les pathogènes, …). De nombreuses études se sont ensuite succédées, chacune utilisant des outils différents, mais toutes s'accordant sur la fréquence importante des CNVs au sein du génome humain (Conrad et al., 2006; McCarroll et al., 2006; Redon et al., 2006; Tuzun et al., 2005). En 2006, Redon et al. ont publié une carte des 1447 régions CNVs (CNVRs) identifiées à partir de 270 individus issus de 4 populations humaines différentes (Figure 14) (Redon et al., 2006). Ces CNVRs couvrent plus de 360 Mb d'ADN, et représenteraient ainsi plus de 12 % du génome humain. Actuellement, plus de 38 000 variants structuraux de grande taille, incluant les CNVs, les inversions et les translocations, ont été identifiés. D'après la base de données dbVar, les CNVs couvrent environ 29,7 % du génome humain et 19 % du génome euchromatique. Pourtant, le débat sur la fraction du génome humain touchée par ces variations est encore ouvert. En effet, les interprétations des données d'aCGH varient considérablement en fonction du génome de référence utilisé, ainsi qu'en fonction de la nature des sondes utilisées (BAC, ADNc ou oligonucléotides). Les premières études utilisant des sondes BACs avaient tendance à surestimer la taille des régions génomiques remaniées.

Les études récentes estiment que les CNVs pourraient couvrir entre 5 et 10 % du génome humain (McCarroll et al., 2008; Perry et al., 2008a). Il est actuellement admis que plus de 240 Mb de séquences du génome humain varient en nombre de copies, ce qui représente plus de 12 % de l'euchromatine et environ 6 % de chaque chromosome

(Choy et al., 2010), ce qui en ferait la première source de variations interindividuelles. Les CNVs sont en effet à l'origine d'un plus grand nombre de différences en terme de séquences nucléotidiques entre les individus que les SNPs : 0,5 à 1 % contre 0,1 % (Conrad et al., 2010; Pang et al., 2010; Redon et al., 2006).

Les puces de génotypage de SNPs sont également utilisées pour étudier les CNVs se trouvant dans des blocs de déséquilibre de liaison (Hinds et al., 2005, 2006; Locke et al., 2006; McCarroll et al., 2006). Pour finir, les CNVs peuvent également être détectés directement grâce au séquençage de nouvelle génération (Mills et al., 2006; Zhao et al., 2013), permettant de caractériser plus finement les points de cassure (Schluth-Bolard et al., 2013).

Un des enjeux pour la compréhension de ce type de variants repose sur l'amélioration des techniques d'annotation, ainsi que sur le développement de plateformes accessibles et facilement compréhensibles pour la visualisation et l'interprétation des CNVs ainsi découverts (Gai et al., 2010; Vandeweyer et al., 2011; Zhao and Zhao, 2013).

Ces polymorphismes de nombre de copies ont été observés chez un grand nombre de mammifères, depuis les grands singes jusqu'au rat et à la souris (Adams et al., 2005; Egan et al., 2007; Guryev et al., 2008; Li et al., 2004; Locke et al., 2003; Perry et al., 2006, 2008b; She et al., 2008; Snijders et al., 2005; Wilson et al., 2006). La cartographie des CNVs a également été entreprise chez d'autres organismes modèles, tels que la drosophile (Dopman and Hartl, 2007; Emerson et al., 2008; Zhou et al., 2008) et le zebrafish (Brown et al., 2012).

1.4 Mécanismes de formation

La répartition des CNVs n'est pas uniforme le long des chromosomes. En effet, on observe un enrichissement au niveau des régions péricentromériques et

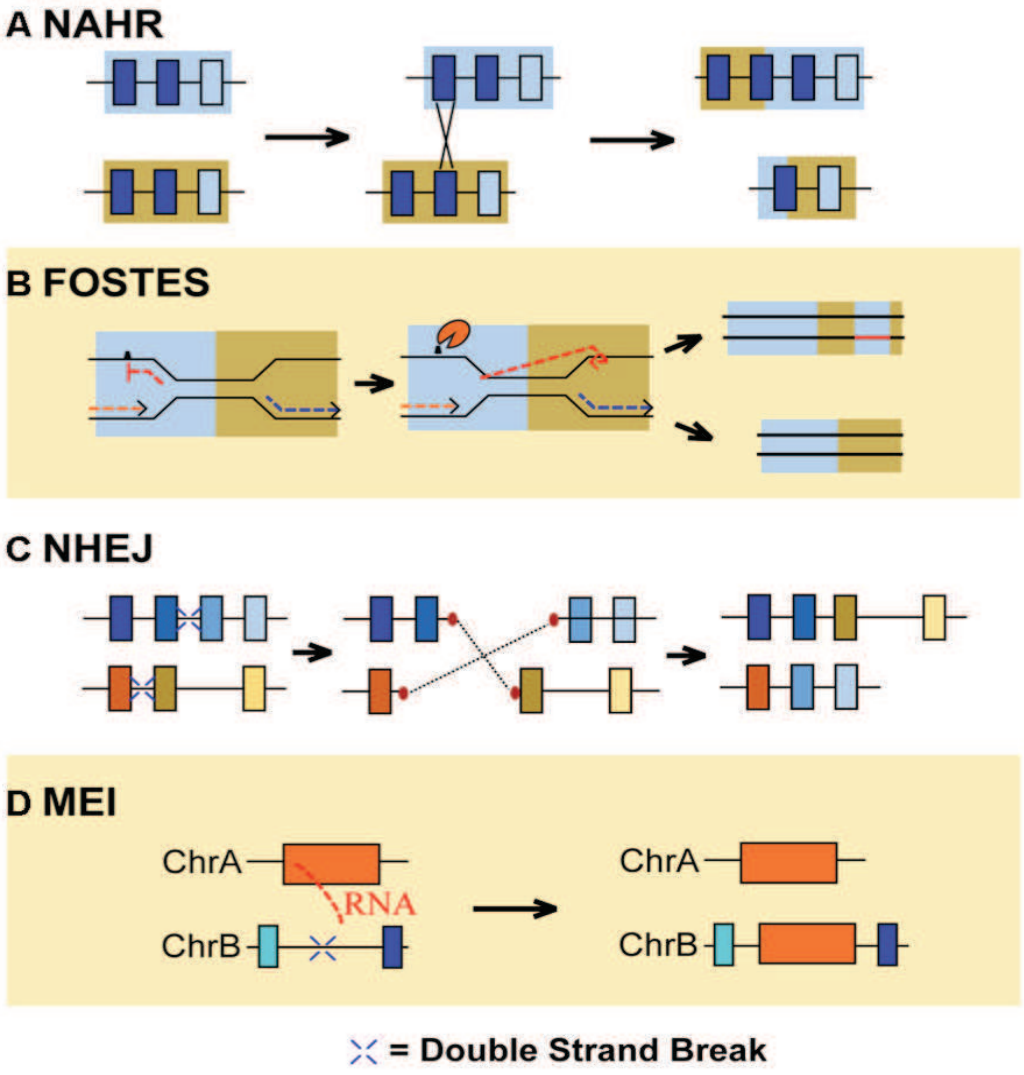


Figure 15 : Mécanisme moléculaire de formation des CNVs.
A. Recombinaison homologue non-allélique (NAHR) entre deux segments non alléliques fortement identiques.
B. Interruption de la fourche de réplication et commutation de la matrice (FoSTeS): permurtation de la fourche de réplication sur une nouvelle séquence hautement similaire.
C. Jonction des extrémités non-homologues (NHEJ) après une cassure double brin.
D. Insertion d'éléments mobiles (MEI): un ARN rétrotranscrit en ADNc est inséré dans le génome.
Extrait de Bickhart and Liu, 2014.

subtélomériques, et une association avec des régions riches en duplications segmentaires (DS, ou segmental duplications SD) (Figure 14) (Sharp et al., 2005). Les DS sont des séquences d'ADN de plus de 1 kb, retrouvées sur plus d'un locus au sein du génome humain, et dont les copies présentent une homologie de plus de 90 % (Bailey et al., 2002). Cette localisation préférentielle des CNVs au niveau des duplications segmentaires s'explique par les mécanismes principalement à l'origine de la variation du nombre de copies : la recombinaison homologue non-allélique (Non-Allelic Homologous Recombination, ou NAHR) et la réparation par jonction des extrémités non-homologues (Non-Homologous End Joining, ou NHEJ), qui sont deux mécanismes de réparation des cassures double brin de l'ADN (CDB, ou Double Strand Break DSB) (Figure 15).

Le mécanisme NAHR génère un CNV quand deux séquences homologues non-alléliques s'alignent et servent de substrat au crossing-over. Ces séquences peuvent être soit des répétitions présentes en peu de copies (Low copy repeats, ou LCR) ou bien des duplications segmentaires de plus de 10 kb. La recombinaison entre deux séquences répétées présentant la même orientation sur un chromosome génère une délétion ou une duplication, tandis que si ces deux séquences présentent une orientation inverse, cela conduira à une inversion de la région concernée. Si les deux séquences répétées se situent sur deux chromosomes différents, la recombinaison génère alors une translocation. Ce mécanisme serait à l'origine de la plupart des CNVs récurrents (Lee and Lupski, 2006), et serait également impliqué dans la formation de réarrangements plus complexes (Hurles, 2005). Les NAHR peuvent survenir pendant la méiose, produisant ainsi des réarrangements chromosomiques *de novo* qui seront transmis à la descendance (Lupski, 2007; Lupski and Stankiewicz, 2005; Turner et al., 2008).

Le mécanisme NHEJ (Non-Homologous End Joining) est un mécanisme alternatif de réparation des CDB, et serait quant à lui médié par des régions de micro-homologie (dont la taille est inférieure à 25 pb) (Lieber et al., 2003). Ce mécanisme, en plus du réarrangement, entraîne souvent la perte ou l'ajout de quelques nucléotides au niveau des points de jonction. Cette signature au niveau des points de cassures est facilement identifiable, et particulièrement retrouvée au niveau des éléments répétés (Alu, LTR, …). Bien que moins fidèle que la recombinaison homologue, le NHEJ est le mode de

réparation des cassures double brin d'ADN prédominant chez les mammifères car il est rapide et peut intervenir quel que soit l'état de réplication de l'ADN.

Plus récemment, un nouveau mécanisme, FoSTeS, lié à la réplication de l'ADN, a été proposé pour expliquer des réarrangements plus complexes et non récurrents (Figures 11 & 15), comme par exemple ceux des délétions et duplications non récurrentes du gène *PLP1* entrainant le syndrome de Pelizaeus-Mersbacher (Lee et al., 2007). Lorsque la fourche de réplication progresse, l'extrémité 3' du brin d'ADN en formation peut changer de matrice pour un autre ADN simple brin situé sur une fourche de réplication voisine.

Pour finir, la rétrotransposition d'éléments L1 (Long interspersed element-1) contribue à la création de certains polymorphismes du nombre de copies. Ces éléments couvrent environ 17 % du génome et représentent la seule classe de transposons encore actifs.

1.5 Les conséquences phénotypiques des CNVs

La plupart des CNVs identifiés à ce jour sont considérés comme bénins : ils n'ont aucun impact phénotypique ou contribuent à certains traits phénotypiques humains non pathologiques. Certains CNVs rares ont quant à eux été reliés à l'apparition de syndromes génétiques. Un des enjeux actuel réside dans la compréhension de la contribution des CNVs communs aux maladies complexes.

La localisation au sein du génome ainsi que le contenu du CNV conditionnent ses éventuelles conséquences phénotypiques. Beaucoup d'études montrent que les CNVs peuvent influencer l'expression de certains gènes en modifiant leur dosage, en affectant, par un effet positionnel, leur régulation ou en interrompant des séquences codantes (Kleinjan and van Heyningen, 2005; McCarroll et al., 2006; Nguyen et al., 2006). Des mécanismes alternatifs ont également été proposés, comme le dévoilement d'une mutation ou d'un SNP fonctionnel lors de la délétion d'un allèle, tandis que

d'autres mécanismes restent bien évidemment à élucider. Ainsi, 15 % de la variation d'expression des gènes pourrait être expliquée par les CNVs (Stranger et al., 2005).

1.5.1 Implication des CNVs dans des traits phénotypiques humains

Des études GWAS avaient mis en évidence l'implication de SNPs dans de nombreux traits humains, tels que la capacité d'élocution (Lai et al., 2001) ou la taille (Gudbjartsson et al., 2008; Lettre et al., 2008). Pour les CNVs, l'exemple le plus connu est celui du gène *AMY1*, codant pour l'amylase salivaire et dont le nombre de copies varie de 2 à 15 en fonction des populations humaines. Il a été montré que le niveau d'expression de l'amylase salivaire était directement corrélé avec le nombre de copies du gène (Perry et al., 2007). Ainsi, due à la sélection positive, le nombre moyen de copies est plus élevé dans les populations consommant plus d'amidon. Un autre exemple bien connu, mais encore largement discuté, est celui de l'effet du nombre de copies du gène *CCL3L1*, codant pour un récepteur aux chémokines, sur la prédisposition à l'infection par le VIH (Virus de l'Immunodéficience Humaine) (Gonzalez et al., 2005; Kuhn et al., 2007). D'autres études ont aussi relié le nombre de copies du gène de l'opsine à l'acuité visuelle (Cooper et al., 2007), et celui du gène *CYP22A6*, un composant du cytochrome P450, au métabolisme de la nicotine (Rao et al., 2000). Cependant, l'implication des CNVs dans des traits quantitatifs plus communs, tels que la taille, n'est pas encore élucidée, bien que de nombreuses études l'envisagent (van Duyvenvoorde et al., 2013; Zahnleiter et al., 2013).

1.5.2 Implication des CNVs dans l'apparition de maladies

Des délétions ou des duplications de régions d'ADN peuvent être à l'origine de syndromes génétiques rares. Ces maladies, causées par des réarrangements génomiques, ont par la suite été regroupées sous l'appellation de « maladies génomiques » (Lupski,

2007; Stankiewicz and Lupski, 2002). Plus de 59 syndromes et CNVs impliqués sont actuellement recensés dans la base de données DECIPHER (Database of Chromosomal Imbalances using Ensembl Resources). En 1991, la duplication à l'origine de la maladie de Charcot-Marie-Tooth fut l'une des premières décrites. Cette duplication en tandem d'une séquence de 1,5 Mb au niveau du locus 17p11.1-p12, médiée par des duplications segmentaires flanquantes, entraîne un problème de dosage de la protéine PMP22 impliquée dans le système nerveux périphérique. A l'inverse, la perte d'une copie du gène *PMP22* conduit à un autre phénotype : une neuropathie héréditaire avec sensibilité à la pression (Hereditary Neuropathy with Liability to Pressure Palsy, ou HNPP) (Lupski et al., 1991). Le syndrome de Prader-Willi, le syndrome de DiGeorge, le syndrome d'Angelman, ou encore le syndrome de Williams-Beuren sont d'autres exemples de ces maladies génomiques. Depuis, de nombreux CNVs rares ont également été associés à l'apparition de tableaux cliniques variés : des maladies congénitales (Southard et al., 2012), des maladies neurologiques (Sebat et al., 2007), ou encore l'apparition de cancer (Kuiper et al., 2010; Ledet et al., 2013; Shlien et al., 2008).

Les récentes études d'associations pan-génomiques ont mis en évidence la contribution des CNVs communs dans l'apparition des maladies multifactorielles. C'est le cas pour certaines maladies auto-immunes ou infectieuses, comme la maladie de Crohn (Bentley et al., 2010) ou des infections pulmonaires chroniques, pour la sensibilité à l'infection par certains pathogènes, comme la malaria (Hedrick, 2011) et le VIH (Gonzalez et al., 2005), et également dans l'apparition de certaines maladies neurologiques, comme l'autisme et la schizophrénie (Levinson et al., 2011; Marshall and Scherer, 2012).

Pour faciliter l'interprétation du nombre croissant de CNVs identifiés, Miller et ses collaborateurs ont proposé une table de classification avec différents critères d'évaluation : caractère *de novo*, taille supérieure à 400 kb, contenu en gènes, … (Miller et al., 2010). Néanmoins, il est parfois encore difficile, par exemple lors de réarrangements complexes, d'établir une corrélation génotype-phénotype fiable (Boutry-Kryza et al., 2012). Ainsi, pour 10 % des CNVs identifiés par aCGH, l'interprétation est difficile. L'implication des CNVs multialléliques est ainsi

particulièrement difficile à tester (Hollox, 2008), puisqu'il ne faut pas prendre en compte le nombre absolu de copies, mais la variation par rapport à la moyenne de la population, une donnée souvent manquante.

1.6 Un seul génome de référence ?

Le séquençage du génome humain a permis de révolutionner la recherche en génétique et la biologie moléculaire, nous faisant entrer dans l'ère de la génomique, en s'affranchissant des étapes longues et fastidieuses que représentaient le clonage et le séquençage de chaque gène potentiellement impliqué dans un phénotype particulier. L. Stein fait ainsi le constat que les biologistes ne vont maintenant plus à la paillasse mais directement sur internet pour chercher dans des bases de données la réponse à leurs questions (Stein, 2004). En moins de 20 ans, nous sommes passés de l'étude *gène-par-gène* à des études pan-génomiques réalisées sur des milliers d'individus, expliquant en partie ou complètement de nombreux traits phénotypiques humains ou des pathologies. En effet, de nombreux traits humains ont montré une forte héritabilité, et ces études pan-génomiques ont permis d'identifier le ou les gènes impliqué(s). Par ailleurs, l'identification des 10 % du génome variant en nombre de copies n'aurait pas pu être possible sans la première ébauche du génome humain (Carter, 2004).

A ce jour, ces études pan-génomiques n'ont permis de rendre compte que d'une faible partie de l'héritabilité observée pour des maladies ou des caractères multigéniques (environ 10 %). 90 % de l'influence génétique reste donc inexpliqué, ce qu'on appelle couramment l'héritabilité manquante. Considérant le nombre croissant d'études pan-génomiques réalisées au cours des dernières années, comment et par quoi expliquer la part d'héritabilité manquante (Maher, 2008) ? Il se peut que la réponse ne se trouve justement pas dans les bases de données ou dans les séquences déjà explorées. La plupart de ces études pan-génomiques utilisent comme seul point de comparaison le génome de référence. Or, ce génome est encore incomplet et imparfait, et ne provient que de la synthèse de quelques génomes. Puisque les génomes de deux individus divergent de 0,5 %, avec en moyenne 600 à 900 régions variant en nombre

de copies (Korbel et al., 2008), on peut s'interroger sur la nécessité de disposer de plusieurs génomes de référence. De plus, des séquences entières du génome humain sont exclues des études pan-génomiques (Eichler et al., 2004), puisque manquantes au sein du génome de référence. Parmi ces séquences difficiles à séquencer ou à assembler par les méthodes bioinformatiques classiques, on trouve les séquences répétées qui représenteraient plus de deux tiers du génome humain (de Koning et al., 2011; Levy et al., 2007).

Un des enjeux pour la compréhension et la caractérisation de nouveaux CNVs est l'amélioration des techniques de séquençage haut-débit et d'assemblage du génome humain, ainsi que la comparaison de génomes de plus d'individus.

2 Le cas particulier des répétitions macrosatellites : un défi pour le séquençage du génome humain

2.1 Définition

2.1.1 Répétitions en tandem

Les répétitions en tandem sont définies comme des copies consécutives parfaites ou légèrement imparfaites d'un motif d'ADN de longueur variable (Charlesworth et al., 1994). Habituellement, on parle de répétitions en tandem dès qu'on observe deux copies contigües d'un motif de nucléotide pouvant être dégénéré (Benson, 1999), chaque répétition pouvant contenir des mutations qui lui sont propres. Elles peuvent avoir un rôle dans la régulation de l'expression des gènes, en interagissant avec des facteurs de transcription (Transcription Factors, ou TFs) ou en modifiant la structure de la chromatine (Benson, 1999; Hamada et al., 1984; Lu et al., 1993; Pardue et al., 1987; Richards et al., 1993; Yee et al., 1991). Une étude a également montré qu'elles jouent un rôle important dans le développement des cellules du système immunitaire (Benson, 1999).

2.1.2 ADN satellite

Les répétitions en tandem présentant un polymorphisme sont référencées comme *satellite* lorsqu'après ultracentrifugation sur un gradient de chlorure de césium, elles apparaissent comme des bandes secondaires (dites *satellites*) dans les tubes de centrifugation, séparées du reste de l'ADN génomique (Campbell et al., 1999;

Charlesworth et al., 1994). Cette sédimentation différentielle est due à leur composition en nucléotides fortement biaisée par rapport au reste du génome.

Les satellites représentent 6 % du génome, et sont ainsi les séquences répétées en tandem les plus abondantes (Lee et al., 1997). Ils sont divisés en six familles : les satellites 1, 2, 3, α, β, et ϒ. Les plus étudiés sont les satellites α, retrouvés au niveau des centromères de tous les chromosomes et composés d'un motif unitaire de 171 pb. Ces motifs sont également organisés en motifs répétés d'ordre supérieur (higher order repeat, ou HOR), alignés de manière ininterrompue dans la même orientation sur des distances allant de 100 kb à 4 Mb.

2.1.3 Répétitions macrosatellites

Les répétitions macrosatellites sont le plus souvent définies comme s'étendant sur plusieurs kb, plutôt que sur la taille de leur unité répétée (qui selon les définitions est supérieure à 100 pb ou 1 kb). Elles s'étendent le plus souvent sur 50 à 100 kb. De nombreuses appellations existent dans la littérature : longues répétitions en tandem (long tandem repeat) (Hannan, 2010), séquences répétées en tandem un nombre variable de fois (VNTR), segments longs répétés en tandem (Long Segment Tandem Repeats, ou LSTR, pour le projet deCODE), CNVs multi-alléliques, CNVs à grande échelle (Large-scale Copy Number Variations, ou LCVs) (Iafrate et al., 2004), mégasatellites (Gondo et al., 1998; Saitoh et al., 2000), … Ces répétitions en tandem de grande taille sont présentes chez tous les organismes vivants (Näslund et al., 2005).

Elles présentent un fort taux de polymorphisme au sein d'une même population, le nombre de répétitions variant habituellement entre quelques répétitions et plus de 100. Elles sont le plus souvent spécifiques d'un ou deux loci. Certaines contiennent une séquence codante. La plupart des macrosatellites identifiés à ce jour sont localisés au niveau des centromères ou des régions péricentromériques, mais quelques-uns ont été identifiés à des loci non-centromériques, comme par exemple les répétitions *D4Z4* qui seront décrites par la suite.

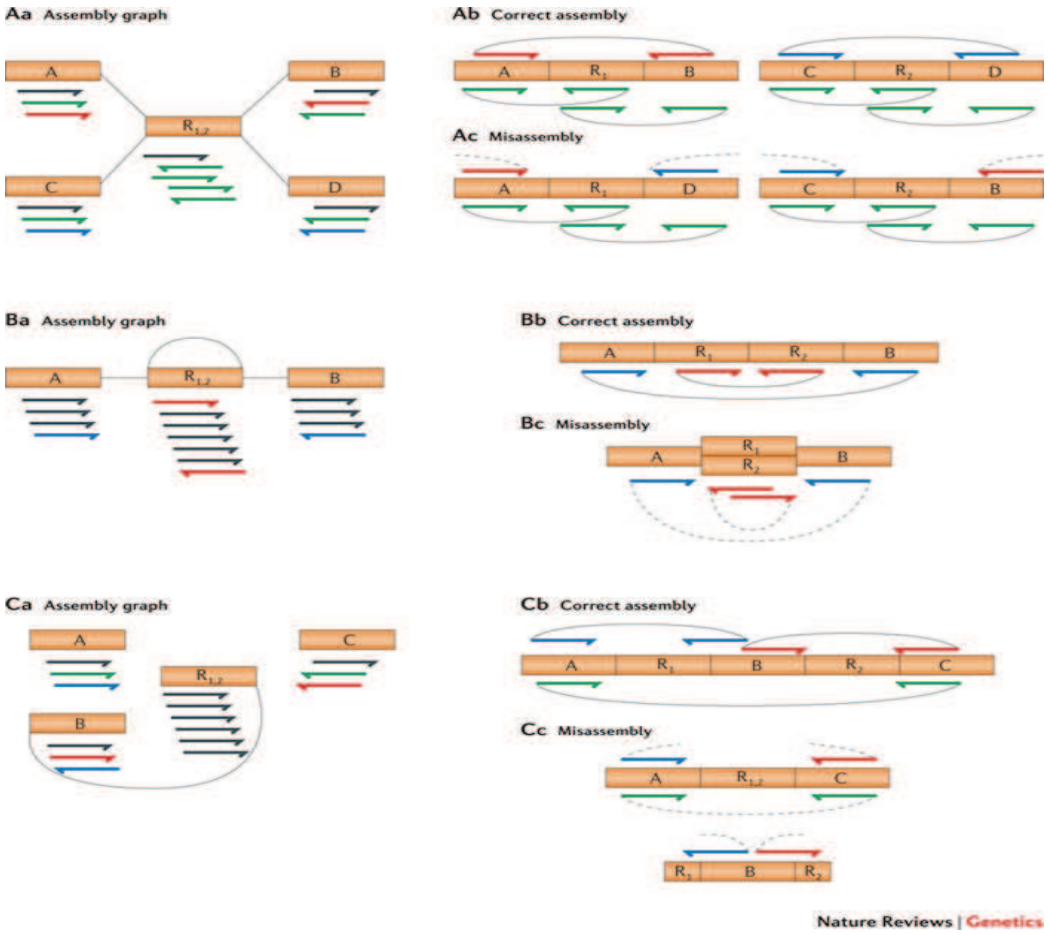


Figure 16 : Erreurs d'assemblage causées par les répétitions d'ADN.
Aa. Graphe d'assemblage contenant 6 contigs, dont 2 identiques (R_1 et R_2), et les lectures alignées sur ces contigs.
Ab. Assemblage correct, respectant les contraintes du mate-pair.
Ac. Deux assemblages incorrects
B. Répétition en tandem.
Ba. Graphe d'assemblage contenant 4 contigs dont 2 identiques (R_1 et R_2)
Bb. Assemblage correct, respectant les contraintes de mate-pair et les distances entre elles.
Bc. Mauvais assemblage, diminuant les distances mate-pair.
C. Répétition dispersée.
Ca. Graphe d'assemblage contenant 5 contigs dont 2 identiques (R_1 et R_2).
Cb. Assemblage correct
Cc. Mauvais assemblage, avec perte de la séquence unique entre les 2 répétitions (localisée sur un autre contig et flanquée de répétitions partielles).
Extrait de Treangen and Salzberg, 2012.

2.2 Un défi pour l'assemblage et une sous-estimation au sein du génome humain

La difficulté à assembler les régions répétées en tandem avait été évoquée dès la publication de la première ébauche du génome humain : « le HGP s'est focalisé en premier lieu sur les séquences euchromatiques, en délaissant les régions hétérochromatiques composées de longues séries d'ADN répété particulièrement difficiles à séquencer et/ou assembler sans erreur ». De fait, les répétitions en tandem contenant les gènes codant pour les unités ribosomales 28S, 18S et 5,8S n'étaient tout simplement pas présentes (Lander et al., 2001).

Les régions répétées génèrent des ambiguïtés dans l'alignement des lectures de séquençage (Figure 16). Ces ambiguïtés sont d'autant plus fortes que les répétitions ont un degré d'homologie élevé : le logiciel d'assemblage peine à placer correctement les lectures les unes par rapport aux autres, et aura tendance à sous-estimer le nombre de répétitions, voire à considérer la séquence comme unique. Plusieurs protocoles d'alignement privilégient encore la stratégie qui consiste à ignorer ces lectures multiples. Cette stratégie limite l'analyse aux régions uniques du génome, mettant ainsi de côté les familles multigéniques et les séquences répétées en tandem, qui pourraient avoir une signification biologique (Tucker et al., 2011). Encore aujourd'hui, les macrosatellites sont souvent absents du génome de référence ou localisés au niveau de régions faiblement ou incorrectement assemblées (Burrows et al., 2010; Iafrate et al., 2004; Phillippy et al., 2008; Warburton et al., 2008), et ne font pas l'objet d'une recherche systématique ou d'un effort d'annotation particulier dans les différents projets de génomes (Delgrange and Rivals, 2004). L'étude pan-génomique de Warburton, identifiant 96 macrosatellites (dont certains préalablement décrits dans les bases de données de CNVs), a ainsi montré que les macrosatellites sont plus abondants qu'attendu (Warburton et al., 2008).

Néanmoins, des progrès considérables ont été réalisés ces dernières années. De nombreux programmes d'assemblage ont été développés afin de pallier ces difficultés

Tableau 4 : Caractéristiques des principaux macrosatellites connus à ce jour.

	TAF11-like ou MSR5p	***SST1***	***PRR20 ou FLJ40296***	***ZAV***	***RS447***	***RNU2***	***CT47***	***D4Z4 (4q)***	***D4Z4 (10q)***	***DXZ4***	***Gor1***	***CT45***
Locus	5p15.1	4q28.3 (19q13.12)	13q21.1	9q32	4p16.1 (8p23)	17q21-22	Xq24	4q35.2	10q26.3	Xq23	8p21.2	Xq26.3
Taille de l'unité	3,4	2,4 - 2,5	6,6	5,3	4,7	6,1	4,8	3,3	3,3	3	12,2	19.9
Nombre de répétitions au sein du génome de référence	17	17 (17 + 15)	5	5	9	0	12	7	6	17	6	4
Nombre d'individus et techniques utilisées	Schaap et al. PFGE 210 individus	Tremblay et al. PFGE 22 individus	Schaap et al. PFGE 210 individus		Schaap et al. PFGE 210 individus	Schaap et al. PFGE 210 individus	Schaap et al. PFGE 210 individus	Schaap et al. PFGE 210 individus	Schaap et al. PFGE 210 individus	Schaap et al. PFGE 210 individus		
Nombre de répétitions	8 - 131	14-154	2 - 30	3 - 31	8-113	5-63	4-17	10 - > 150	2 - 105	18-120	50-150	
Nombre d'allèles	54	134	20	41		> 53				> 50		
Variation de taille du macrosatellite (kb)	34-335	35-387	34 - 136	20 - 168	37 - 531	30 - 386	19 - 82	33 - >495	7 - 347	54 - 360	610 - 1830	
Taille moyenne du macrosatellite (kb)	181	161	76	71	281	134	100	103	83	177	167	108,1
Contenu en GC	50%	64%	50%	60%	50%	56%						
ORF	*TAF11-like*: RNA Polymerase like	-	*PRR20*	-	*USP17 / DUB3*	*U2*	cancer testis antigen 47	*DUX4* et *DUX4-like*		long non coding RNA	*Gor1*	-
Expression	Testicule, Cerveau	Ubiquitaire	Testicules	Testicules, Cerveau	Cœur, foie, pancréas	Ubiquitaire	Testicules, placenta, cerveau				ND	
Instabilité méiotique	Oui	Non	Oui (0,8%)	Non	Oui	Oui (0,8%)	Non	Non		Oui	Non	
Instabilité mitotique	Oui (0,4%)	Oui	Oui (0,7%)	Non	Oui (0,4%)	Oui (1,5%)	Oui (0,7%)	Oui (0,7%)	Oui (0,7%)	Oui (2,2%)		
Conséquence phénotypique	Schizophrénie ?	-	-	-	Parkinson ?	-	Expression faible dans le cancer du poumon	< 10 copies : FSHD		Inactivation chromosome X		

(Huang and Madan, 1999; Tammi et al., 2003; Treangen and Salzberg, 2012) et plusieurs outils ont été développés pour détecter ces variants structuraux à partir de données de séquençage nouvelle-génération (Next-Generation Sequencing, ou NGS), comme par exemple l'outil Tandem Repeat Finder (TRF) (Benson, 1999). Ainsi, un nombre croissant de macrosatellites a pu être découvert et cartographié avec précision. Actuellement, la base de données des répétitions en tandem (Tandem Repeat Database, ou TRDB) (Gelfand et al., 2007) répertorie 162 répétitions en tandem avec une unité répétée supérieure à 1kb, dont le nombre de copies moyen varie entre 1,8 et 83. Malheureusement, cette base est très incomplète : aucune répétition avec une unité supérieure à 2,1 kb n'est répertoriée, alors que plusieurs ont été largement décrites et étudiées (Schaap et al., 2013; Tremblay et al., 2010, 2011).

Par ailleurs, lorsque les macrosatellites sont correctement localisés au sein du génome de référence, ils sont souvent retrouvés au niveau de régions pauvrement assemblées (présentant des trous au sein des répétitions ou directement avant ou après), et le niveau de polymorphisme indiqué ne reflète pas ou peu la réalité (Schaap et al., 2013; Tremblay et al., 2010; Warburton et al., 2008). Par exemple, au sein du génome de référence, seulement 6 copies du locus *Gor1* sont présentes, séparées par un trou dans l'assemblage. On retrouve ainsi 5 répétitions du côté centromérique, un trou de 87 kb, et environ 1,5 répétitions du côté télomérique. Warburton *et al*. ont montré, par électrophorèse en champ pulsé, que ce locus contient en fait de 50 à 150 répétitions d'une unité de 12 kb (Warburton et al., 2008). Un des enjeux dans l'étude des macrosatellites est la détermination précise du nombre moyen, mais également du nombre minimum et maximum de répétitions au sein de chaque population. L'étude locus par locus avec des techniques moléculaires, certes plus fastidieuse, reste la seule à fournir ces informations et donc la seule façon de caractériser finement ces loci.

L'étude pan-génomique la plus complète à ce jour montre que parmi les 96 macrosatellites identifiés, la taille d'une unité répétée varie de 1,5 à 38,8 kb, le nombre de répétitions variant lui entre 2 et 36 (Warburton et al., 2008). Néanmoins, peu de ces régions ont été caractérisées finement par des techniques moléculaires. Elles sont présentées dans le Tableau 4 (Schaap et al., 2013; Tremblay et al., 2010, 2011). Certains macrosatellites ont été particulièrement bien documentés, soit à cause de leur

découverte historique ou parce qu'ils ont été reliés à l'apparition d'une maladie. Ils seront présentés dans la section suivante.

2.3 Les macrosatellites connus

2.3.1 Le locus *D4Z4* et le syndrome FSHD

Le macrosatellite *D4Z4* a été décrit en 1995 comme des répétitions en tandem complexes d'un motif de 3,3 kb (Lyle et al., 1995). Par analyse de liaison, il a été localisé dans un premier temps en 4q35 (van Geel et al., 1999; Lemmers et al., 2004; Lyle et al., 1995; Wijmenga et al., 1991). Par la suite, un second locus contenant des répétitions *D4Z4* a été identifié en 10q26. Le nombre de copies du locus *D4Z4*, localisé en 4q35, varie dans la population générale entre 11 à plus de 150 copies.

Ce macrosatellite a été relié à l'apparition de la myopathie facio-scapulo-humérale (Fascioscapulohumeral Muscular Dystrophy, ou FSHD), une maladie touchant environ 1 individu sur 20 000 (van Geel et al., 1999). Il s'agit d'une dystrophie musculaire de transmission autosomique dominante, caractérisée par une atrophie musculaire progressive impliquant la face, la ceinture scapulaire, les membres supérieurs, les membres inférieurs et enfin la hanche (Padberg et al., 1991; Tawil et al., 1998). La sévérité et l'âge d'apparition de la maladie sont variables, mais la pénétrance de la maladie est quasiment complète après l'âge de 20 ans (Lunt and Harper, 1991; Lyle et al., 1995). Chez 95 % des patients FSHD, une contraction du locus *D4Z4* est observée, celui-ci ne contenant plus que de 1 à 10 copies (Lunt, 1998). Il existe généralement alors une corrélation inverse entre le nombre de répétitions *D4Z4* et la sévérité de la maladie : plus le nombre de répétitions est bas, plus le début de la maladie est précoce et les symptômes accentués (Lemmers et al., 2004; Lunt et al., 1995; Tawil et al., 1996; Zatz et al., 1995). La contraction du locus *D4Z4* sur le chromosome 10q n'a quant à elle aucune conséquence pathologique.

Cette contraction ne semble pas perturber la structure d'un gène spécifique.

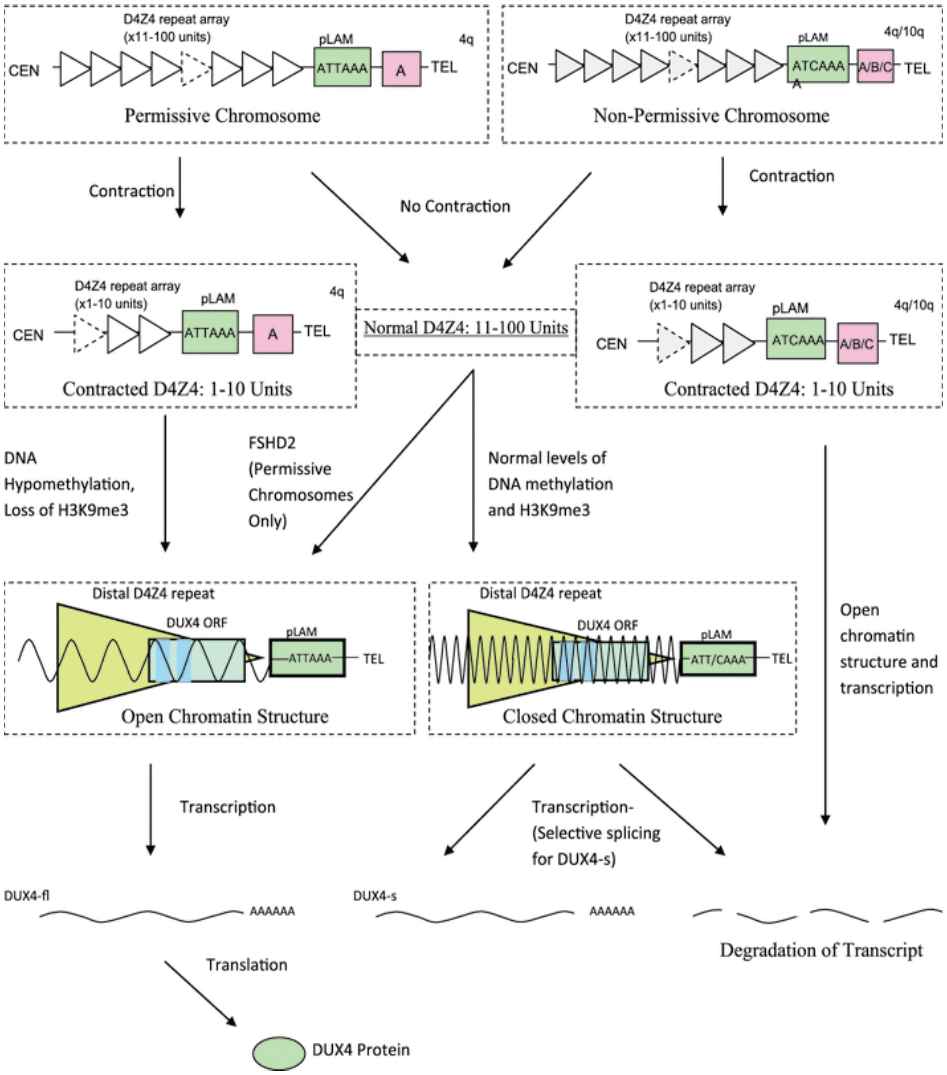


Figure 17 : Mécanismes moléculaires dans le syndrome de la dystrophie facio-scapulo-humérale (FSHD).

Les contractions *D4Z4* sur les allèles permissifs ou non permissifs conduisent à une hypométhylation de l'ADN et la perte des H3K9Me3, signes d'une structure chromatinienne plus ouverte. Ceci permet la transcription de *DUX4*. Pour les allèles non permissifs, l'absence de signal de poly-adénylation (polyA) entraîne la dégradation du transcrit. Pour les allèles permissifs, deux transcrits sont synthétisés: DUX4-s et DUX4-fl.

Extrait de Richards et al., 2012.

L'hypothèse d'un mécanisme épigénétique complexe, impliquant une modification de la conformation de la chromatine au locus 4q35 et conduisant à une régulation inappropriée de gènes cibles en *cis* et/ou en *trans,* a donc été formulée afin d'expliquer la maladie (de Greef et al., 2008; van der Maarel and Frants, 2005). En effet, le locus contracté présente une perte partielle de méthylation sur l'ADN, et de certaines marques épigénétiques répressives sur les histones (H3K9Me), conduisant à une structure chromatinienne plus ouverte (van Overveld et al., 2003). Cette relaxation de la chromatine entraîne l'augmentation de l'expression du gène *DUX4*, contenu dans chacune des répétitions et codant pour la protéine à double homéobox 4. L'augmentation de l'expression de ce facteur de transcription a été observée dans les myoblastes des patients atteints de FSHD. Il a également été montré qu'un site de polyadénylation fonctionnel aboutit à des transcrits stables *DUX4*, résultant en l'expression de la protéine DUX4. L'expression de *DUX4* dans le muscle entraîne une activation des cellules souches primaires, et la transcription de gènes impliqués dans des dommages aux cellules musculaires (Lemmers et al., 2010). Ces allèles porteurs du site de polyadénylation sont considérés comme permissif (Figure 17).

Récemment, une étude a montré que la contraction des répétitions *D4Z4* entraîne la transcription d'un ARN non-codant (long ncRNA), DBE-T. Ce dernier serait responsable de l'activation des gènes voisins, comme *DUX4* (Cabianca et al., 2012). Encore aujourd'hui, plusieurs hypothèses physiopathologiques, ne s'excluant pas forcément, sont envisagées : boucle chromatinienne, insulateur, localisation nucléaire, ... Toutes s'accordent sur l'importance du remodelage de l'architecture chromatinienne au locus 4q35, provenant directement de la diminution du nombre de copies.

2.3.2 Le locus *DXZ4* et l'inactivation du chromosome X

De même, le locus *DXZ4*, contenant de 20 à 120 copies d'une unité de 3 kb (Schaap et al., 2013), est largement étudié à cause de son implication dans le processus d'inactivation du chromosome X et comporte plus de 50 allèles différents. Il a été

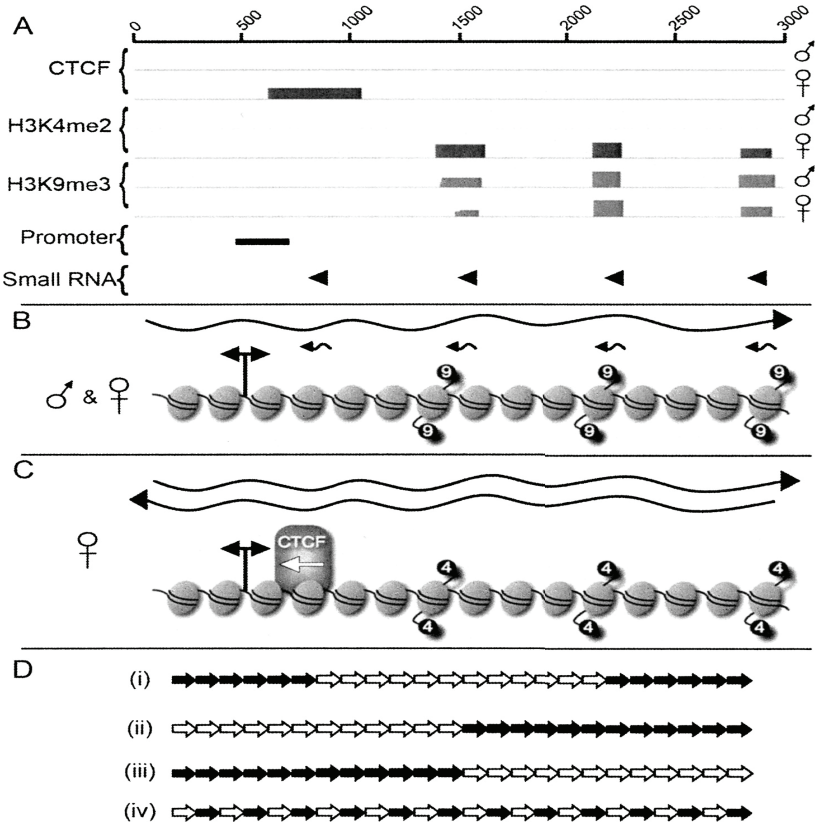


Figure 18 : Organisation chromatinienne et transcription des répétitions *DXZ4* localisées sur les chromosomes X actifs et inactifs.
A. Résumé des résultats de ChIP-chip, luciférase et Northern Blot.
B. Chromatine et transcrits chez les mâles Xa. Flèche en vague droite: transcrit sens de 3kb. Flèche gauche: transcrit anti-sens.
C. Chromatine et transcrits pour le Xi.
D. Représentation schématique d'une possible conformation des monomères H3K4Me2 (blanc) et H3K9Me3 (noir) sur le Xi.
Extrait de Chadwick, 2008.

montré par FISH que le locus *DXZ4* est conservé chez les trois grands singes (chimpanzé, orang-outan et gorille) (Samonte et al., 1999). Cependant, une étude plus récente évoque une conservation uniquement des séquences promotrices (McLaughlin and Chadwick, 2011).

Chez les mammifères, la femelle possédant deux chromosomes X contre un seul chez le mâle, un mécanisme d'inactivation d'un de ces deux chromosomes se met en place. Ce processus a pour but d'équilibrer le niveau d'expression des gènes liés à l'X entre la femelle et le mâle. Ainsi, au cours du développement embryonnaire précoce, l'un des deux chromosomes X des femelles passe donc d'un état actif (Xa) à un état inactif (Xi) et hétérochromatique, connu sous le nom de corpuscule de Barr.

L'étude de McLaughlin a montré que le macrosatellite *DXZ4* arbore des marques épigénétiques différentes sur le chromosome X actif (Xa) et le chromosome X inactif (Xi) (McLaughlin and Chadwick, 2011). Bien qu'aucune phase ouverte de lecture n'ait pu être identifiée au sein de chaque monomère, ils ont montré que le facteur de liaison CCCTC (CTCF : zinc finger protein CCCTC binding factor) se lie aux répétitions *DXZ4*. De plus, l'état chromatinien des répétitions est différent de celui des séquences environnantes, suggérant un rôle de ce macrosatellite dans l'organisation du Xi impliquant la protéine CTCF (Figure 18) (Chadwick, 2008).

2.3.3 Le locus *TAF11-like* : un rôle potentiel dans la schizophrénie ?

Les répétitions *TAF11-like*, localisées en 5p15, contiennent une courte séquence codante pour une protéine ressemblant à un facteur associé aux protéines de liaison de la boîte TATA. La plupart des tissus n'expriment pas cette protéine, mis à part les testicules et le cerveau (Tremblay et al., 2010).

Une étude par CGH-array, puis par PFGE, a révélé la présence d'allèles de petite taille du locus *TAF11-like* (< 21 copies) chez plusieurs individus atteints de schizophrénie (Bruce et al., 2009). Ces allèles de petite taille co-ségrégent partiellement avec la schizophrénie dans un nombre restreint de familles, bien que les résultats d'une

étude d'association comparant la longueur des allèles chez 406 cas de schizophrénie et 392 témoins ne montre pas de différence significative. Cependant, la mesure du nombre de copies lors de cette analyse de liaison a été faite par qPCR : seule la somme des deux allèles est prise en compte, masquant potentiellement l'effet d'un des allèles de petite taille.

2.3.4 *RS447* et la maladie de Parkinson ?

Le macrosatellite *RS447* a été découvert en 1997 (Kogi et al., 1997), et a été localisé majoritairement en 4p16.1 par FISH (Okada et al., 2002) ainsi qu'au niveau d'un locus mineur en 8p23 (Gondo et al., 1998; Okada et al., 2002). Il comporte de 50 à 70 répétitions d'une unité de base de 4,7 kb. L'unité de base ainsi que l'organisation en tandem du locus *RS447* sont conservées chez le singe, la vache, le lapin, le porc et le porc-épic (Gondo et al., 1998).

Le nombre de copies et le profil de méthylation de ce macrosatellite pourraient influencer la structure chromatinienne et ainsi l'expression des gènes voisins (Okada et al., 2002). Chaque répétition contient une séquence codante de 1590 pb, le gène *USP17*, codant pour une enzyme de déubiquitination USP17 (Ubiquitin-specific protease 17) (Saitoh et al., 2000). Cette enzyme a pour fonction de cliver l'ubiquitine des protéines ubiquitinées. Une mutation faux-sens dans le gène de l'hydrolase carboxyterminale d'ubiquitine L1 (*UCHL1*) a été reliée à la maladie de Parkinson dans une famille allemande (Leroy et al., 1998; McNaught et al., 2001). L'effet de cette mutation pourrait être relié à *RS447*, bien que des études supplémentaires soient nécessaires.

2.3.5 Autres macrosatellites potentiellement impliqués dans des traits humains

Le gène *CT47* fait partie de la famille des gènes des cancers testiculaires (CT : cancer/testis). Le locus de ce gène, localisé en Xq24, contient de 4 à 17 répétitions d'une unité de 4,8 kb (Schaap et al., 2013). Dans les lignées cellulaires de carcinome du poumon à petites cellules, il a été montré que l'expression de CT47 est réactivée, bien que très faiblement, suite à la perte des marques épigénétiques répressives (Balog et al., 2012). Ainsi, l'expression de *CT47* pourrait être utilisée comme biomarqueur de certains cancers du poumon et de l'œsophage (Chen et al., 2006).

Le locus *Gor1*, également connu sous le nom de *REXO1L1*, contiendrait lui en moyenne 173 répétitions par génome diploïde (Brahmachary et al., soumis). Bien que la fonction de la protéine GOR ne soit pas connue, plusieurs études suggèrent un lien avec l'infection par le virus de l'Hépatite C et des maladies du foie autoimmunes (Löhr et al., 1994; Michel et al., 1992).

D'autres études associent également la variation du nombre de copies de gènes présents en multiples copies avec certains traits humains, sans que ces loci ne présentent un taux particulièrement élevé de polymorphisme. L'augmentation du nombre de copies du gène *C4*, répété entre 2 et 6 fois dans la population générale, protégerait du lupus érythémateux systémique (Yang et al., 2007). La difficulté à génotyper précisément ces macrosatellites aboutit parfois à une divergence sur leur contribution phénotypique, comme on peut le voir pour le locus *DEF4B* qui, selon les études, serait associé ou non à la maladie de Crohn (Aldhous et al., 2010; Bentley et al., 2010).

2.4 Une source majeure de l'héritabilité manquante ?

Etant donné le degré de polymorphisme particulièrement élevé des macrosatellites jusqu'à présent, il apparait indispensable de génotyper un grand nombre d'individus pour les caractériser. Ainsi, pour les loci étudiés sur un faible nombre d'individus, il est fort probable que de nouveaux allèles, certes plus rares, restent à découvrir. A ce jour, trois études confirment donc que les répétitions macrosatellites font sûrement partie des structures les plus polymorphes du génome humain (Schaap et al., 2013; Tremblay et al., 2011; Warburton et al., 2008).

Pendant longtemps, les séquences répétées ont été considérées comme de l'ADN *poubelle* (junk DNA). Elles suscitent maintenant un intérêt nouveau, avec la prise de conscience des conséquences fonctionnelles qu'elles peuvent avoir du fait de leur organisation en tandem très caractéristique. Elles pourraient s'avérer être des sources majeures de l'héritabilité manquante. Les séquences répétées en tandem, et particulièrement les macrosatellites, sont un outil formidable pour étudier les relations entre la structure de l'ADN et ses fonctions, en se penchant par exemple sur l'implication de ces loci dans l'architecture génomique, la recombinaison ou encore l'évolution (Saitoh et al., 2000).

Pourtant, la majorité des études pan-génomiques actuelles n'arrivent pas à analyser l'effet potentiel de ces macrosatellites. En effet, une multitude d'allèles existent pour ces loci, contrairement aux SNPs qui sont bialléliques. Il est donc peu probable, voire impossible, de trouver un SNP marqueur expliquant la totalité de la variabilité pour chacun de ces loci. Warburton fait donc le constat que l'implication des CNVs multialléliques, et plus particulièrement des macrosatellites, dans les phénotypes complexes, est une question qui reste encore ouverte (Warburton et al., 2008).

Bien que peu de choses soient connues sur leurs fonctions exactes et leurs comportements, les macrosatellites ont d'ores et déjà été connectés à d'importantes fonctions biologiques chez l'Homme et chez d'autres eucaryotes. C'est le cas, par exemple, des répétitions du gène *AMY1* et de la capacité de digestion de l'amidon, un exemple que nous avons déjà évoqué. Plusieurs études soulignent également le rôle

prépondérant de l'ADN satellite dans la formation et la maintenance de structures chromatiniennes au niveau du centromère.

3 Le macrosatellite *RNU2*

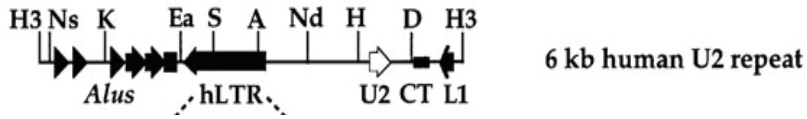
3.1 Découverte et caractérisation du locus

3.1.1 Organisation en tandem et niveau de polymorphisme

En 1981, Westin et ses collaborateurs apportent la première preuve de l'existence de plusieurs gènes *RNU2* au sein du génome humain. En isolant plusieurs clones provenant de librairies d'ADN, dont trois qu'ils étudieront plus en détails (U2/4, U2/6 et U2/7), ils émettent l'hypothèse de l'existence d'au moins 30 copies du gène au sein du génome (Westin et al., 1981). Cette hypothèse sera confirmée en 1984, lorsque Van Arsdell et Weiner montrent que les gènes *RNU2* sont contenus dans un fragment d'ADN de 6 kb, répété en tandem entre 10 et 20 fois par génome haploïde. Contrairement à la famille multigénique *U1*, ces répétitions sont retrouvées dans un seul cluster, et seulement quelques pseudogènes existent au sein du génome humain (Van Arsdell and Weiner, 1984). Dans le même temps, Westin et ses collaborateurs confirment et affinent ces observations : le locus *U2/6,* qui contient le gène *U2 bona fide*, est décrit comme une succession de répétitions en tandem d'une unité de 6,2 kb, comportant de 3 à environ 20 répétitions par génome haploïde (Westin et al., 1984).

Ce n'est que bien des années plus tard que le niveau de polymorphisme de ce locus est étudié sur un grand nombre d'individus issus de différentes populations humaines. Par PFGE conduite sur plus de 40 individus (80 chromosomes), provenant de 8 populations humaines différentes, Liao montre que le nombre de répétitions varie de 5 à plus de 30 (Liao et al., 1997), et donc que le locus *RNU2* peut s'étendre sur 30 à plus de 200 kb. La taille des allèles étudiés se répartit de la façon suivante : 57 % se situe entre 100 et 200 kb, 32 % entre 40 et 100 kb, et 11 % au dessus de la limite de résolution de la technique (200 kb).

A.



B.

+	score	% div.	% del.	% ins.	query sequence	position in query- begin	end	(left)	C	matching repeat	repeat class/family	-position in repeat- (left) begin	end	begin (left)	linkage id/graphic
+	191	30.4	7.3	3.5	UnnamedSequence	221	329	(5803)	+	L2	LINE/L2	270	382	(3037)	1
+	538	15.9	5.6	1.8	UnnamedSequence	384	490	(5642)	+	FRAM	SINE/Alu	54	164	(12)	2
+	16	15.5	0.0	5.0	UnnamedSequence	493	532	(5600)	+	(AACA)n	Simple_repeat	1	40	(0)	3
+	854	19.3	3.4	0.6	UnnamedSequence	552	727	(5405)	+	AluJb	SINE/Alu	129	309	(3)	4
+	1021	21.9	4.9	2.9	UnnamedSequence	801	1105	(5027)	+	ERV3-16A3_I-int	LTR/ERVL	4719	5029	(197)	5
+	1220	16.8	0.0	0.0	UnnamedSequence	1231	1420	(4712)	+	AluSx	SINE/Alu	117	306	(6)	6
+	2349	6.4	2.7	0.3	UnnamedSequence	1421	1718	(4414)	+	AluSx	SINE/Alu	2	306	(6)	7
+	2112	9.3	0.7	2.0	UnnamedSequence	1729	2030	(4102)	+	AluSp	SINE/Alu	9	306	(7)	8
+	463	25.7	1.4	0.2	UnnamedSequence	2040	2177	(3955)	C	MER33	DNA/hAT-Charlie	(194)	130	3	9
+	5417	17.5	1.9	4.5	UnnamedSequence	2324	3355	(2777)	C	LTR13	LTR/ERVK	(0)	1007	1	10
+	264	19.4	9.1	11.1	UnnamedSequence	3365	3474	(2658)	C	MER96B	DNA/hAT-Tip100	(309)	108	1	11
+	208	30.0	0.0	7.5	UnnamedSequence	3479	3550	(2582)	+	MIRb	SINE/MIR	153	219	(49)	12
+	423	32.6	6.2	11.6	UnnamedSequence	3799	4152	(1980)	C	L2c	LINE/L2	(1)	3386	3050	13
+	1764	0.0	0.0	0.0	UnnamedSequence	4882	5068	(1064)	+	U2	snRNA	1	187	(1)	14
+	104	6.8	1.4	0.7	UnnamedSequence	5381	5520	(612)	+	(TC)n	Simple_repeat	1	142	(0)	15
+	13	36.0	0.0	0.0	UnnamedSequence	5522	5579	(553)	+	(TCCCCC)n	Simple_repeat	1	58	(0)	16
+	3172	10.4	0.0	0.4	UnnamedSequence	5580	6063	(69)	C	L1PA8	LINE/L1	(23)	6149	5668	17

Figure 19 : Composition de l'unité de base du locus *RNU2*.
A. Schéma d'après *Pavelitz et al., 1995* de l'unité de base arbitrairement délimitée par deux sites de restriction HindIII (H3).
U2: séquence codante du gène *RNU2*.
CT: microsatellite formé de répétitions des dinucléotides (CT).
Alus, LTR, L1: séquences d'ADN répétés.
B. Contenu en séquences répétées de l'unité répétée après une analyse avec le logiciel RepeatMasker.

Ces données ont été de nouveau affinées par l'étude plus récente, également par PFGE, de 210 individus non-apparentés issus du projet HapMap (Schaap et al., 2013). Les individus ainsi analysés sont issus de différentes populations : des résidents de l'UTAH ayant des ancêtres issus de l'Europe du Nord et de l'Ouest (CEPH European from Utah, ou CEU), des Yoruba du Nigeria (Yoruba in Ibadan Nigera, ou YRI), et une population asiatique (East Asian, ou ASN) regroupant des chinois Han (CHB) et des japonais de Tokyo (JPT). Les auteurs n'ont pas observé de différence de distribution de la taille des allèles en fonction de la population. Le nombre allélique moyen est de 22,79 pour la population asiatique (ASN), 22,99 pour la population européenne (CEU) et 21,72 pour la population africaine (YRI). Le nombre allélique de répétitions du macrosatellite *RNU2* varie de 6 à plus de 63 répétitions.

3.1.2 Contenu de l'unité répétée, évolution concertée et profil de méthylation

L'unité de base, définie arbitrairement par deux sites de restriction de l'enzyme HindIII, a été séquencée par la méthode de Sanger pour la première fois en 1995 (Numéro d'accession : L37793) (Figure 19A) (Pavelitz et al., 1995). Cependant, à cause des limitations technologiques de l'époque, seulement 5734 pb ont été correctement lues et assemblées : une séquence Alu de 300 pb, localisée au sein de la queue polyA d'une autre séquence Alu, était manquante. Cette séquence a par la suite été corrigée en 1996 et déposée en juillet 2000 sous le numéro d'accession U57614 dans les bases de données. L'unité de base contient 6132 pb, et elle est composée de 56,21 % de GC et de 66,89 % de séquences répétées, d'après une analyse avec le programme Repeat Masker (Figure 19B). Elle ne contient qu'une seule séquence codante de 188 bp, le gène *RNU2* codant pour le petit ARN nucléaire U2 (small nuclear RNA U2, ou ARNsn U2). On retrouve deux séquences promotrices : un promoteur proximal (Proximal Sequence Element, ou PSE) de 20 pb, centré en position -50 en amont du site d'initiation de la transcription et ayant des fonctions proches de la boîte TATA, et un promoteur distal (Distal Sequence Element, ou DSE), centré en -235, ayant des

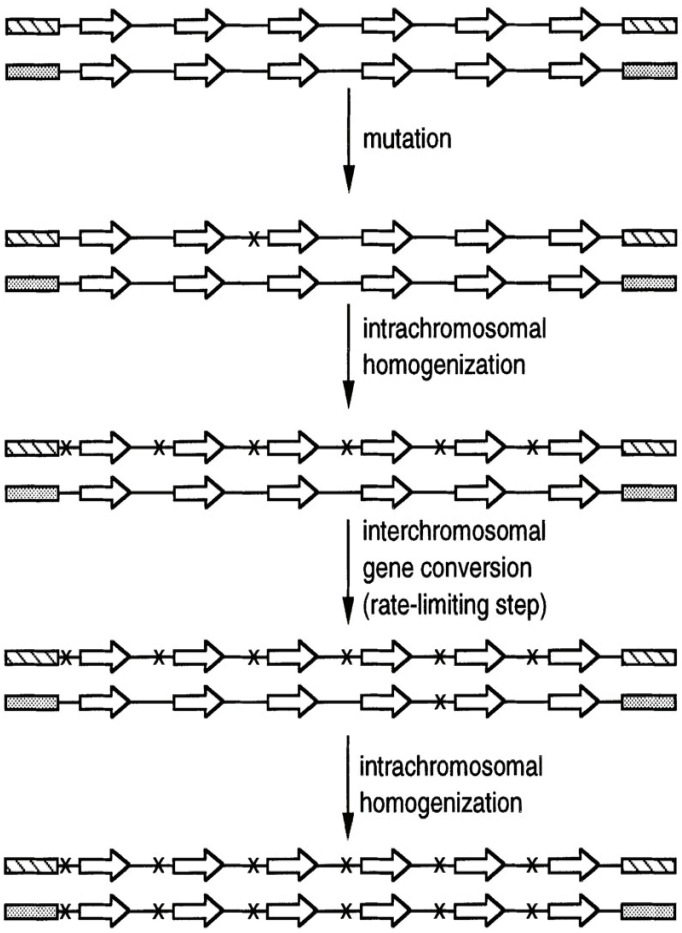


Figure 20 : Modèle proposé pour l'évolution concertée des répétitions en tandem des familles multigéniques, dont le locus *RNU2*.

Deux groupes de répétitions (ou allèles), ainsi que les séquences d'ADN flanquantes, sont illustrées. Les séquences codantes sont représentées avec des flèches blanches encadrées de noir, et les séquences intergéniques avec des lignes. Les séquences flanquantes sont annotées différemment.

Une mutation (croix noire) survient dans une répétition sur un premier allèle, et se propage rapidement sur toutes les répétitions de cet allèle par le mécanisme d'homogénéisation intrachromosomique. La mutation est ensuite transmise sur le second allèle de répétitions par conversion génique interchromosomique, sans qu'aucun échange des séquences flanquantes n'intervienne. Ainsi, la mutation est fixée sur toutes les répétitions du second allèle, par de nouveaux cycles d'homogénéisation intrachromosomique.

Extrait de Liao, 1999.

fonctions d'enhancer et possédant des sites de fixation pour les facteurs de transcription Sp1 et Oct1 (Ares et al., 1987; Pavelitz et al., 1995; Sadowski et al., 1993).

L'analyse de plusieurs polymorphismes au sein de l'unité répétée du locus *RNU2* a montré que les répétitions d'un même allèle sont entièrement homogènes (Liao, 1999; Liao et al., 1997). Ainsi, le macrosatellite *RNU2* est soumis à un processus d'évolution concertée, qui a également été observé pour la plupart des autres familles multigéniques répétées en tandem (rRNA, tRNA, gènes codant pour les histones, …) (Matera et al., 1990; Westin et al., 1984). L'évolution concertée conduit à une homogénéisation des séquences des répétitions entre chromosomes homologues et également entre chromosomes non homologues. On observe alors un plus grand degré d'homologie de séquence entre des répétitions paralogues dans la même espèce qu'avec des répétitions orthologues de différentes espèces (Gonzalez and Sylvester, 2001; Hillis and Dixon, 1991; Liao, 1999). Cette homogénéisation de la variation au sein d'une famille de gènes dupliqués en tandem se traduit, lorsqu'une mutation apparaît sur l'une des répétitions, par sa perte ou bien sa fixation sur l'ensemble des répétitions d'un allèle dans un premier temps (Figure 20). Ensuite, une seconde phase d'homogénéisation a lieu entre les deux allèles. Plusieurs mécanismes moléculaires ont été proposés : la conversion génique, des crossing-over inégaux entre chromatides sœurs, ou bien une amplification génique (pertes et gains fréquents dans une famille) (Liao, 1999). Dans le cas du locus *RNU2*, le mécanisme principal d'évolution concertée serait la conversion génique (consistant en des échanges non réciproques entre séquences homologues) ou des échanges entre chromatides sœurs, puisque Liao et ses collaborateurs n'ont pas observé d'échanges entre les marqueurs flanquants (Liao et al., 1997).

Parmi les polymorphismes décrits au sein de l'unité répétée, le plus décrit est le microsatellite (CT)•$(GA)_n$ car il pourrait jouer un rôle dans l'homogénéisation du locus en promouvant la conversion génique (Pavelitz et al., 1995). En effet, ce microsatellite n'est pas retrouvé dans la séquence des pseudogènes *RNU2* et colocalise avec un site de clivage de la nucléase S1 (site SNS1) (Htun et al., 1985). Ce microsatellite pourrait favoriser les cassures double brin grâce à l'adoption d'une conformation de l'ADN différente de celle en hélice droite (Majumdar and Patel, 2002; Wells, 1988). Il a été

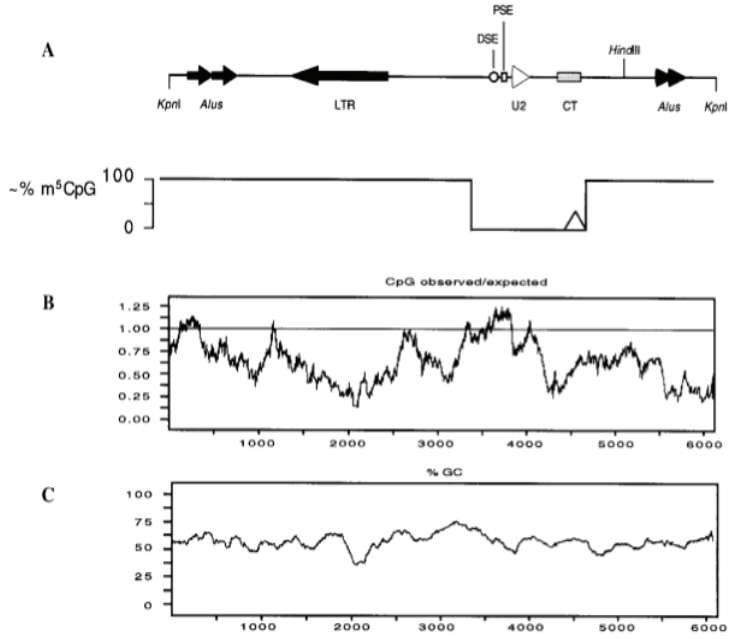


Figure 21 : Représentation schématique du profil de méthylation bimodal de l'unité répétée du locus *RNU2*.

A. Représentation des niveaux de méthylation, estimés par des digestions enzymatiques et séquençage de l'ADN traité au bisulfite. Le triangle indique les variations de niveau de méthylation observée pour le microsatellite CT.
PSE: Proximal Sequence Element
DSE: Distal Sequence Element
B. Fréquence des dinucléotides CpG au sein de l'unité répétée.
C. Densité en GC au sein de l'unité répétée.
Extrait de Jiang et al., 1999.

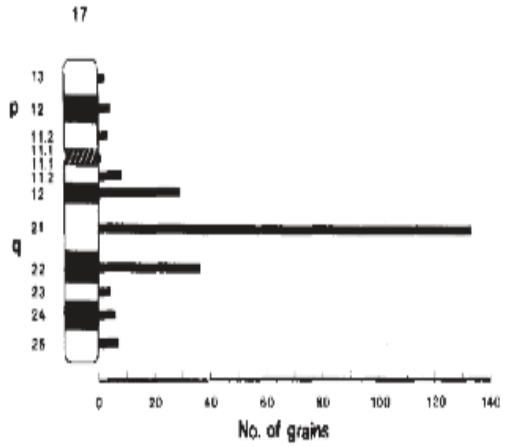


Figure 22 : Expérience ayant permis la localisation du locus *RNU2* en 17q21 en 1985.
Représentation du chromosome 17, montrant la distribution des grains d'argent après hybridation d'une sonde *U2* radioactive sur des chromosomes provenant de deux donneurs humains.
Les barres horizontales indiquent le nombre de grains observés sur les bandes correspondantes sur 158 étalements de chromosomes en métaphase.
Extrait de Lindgren et al., 1985.

montré chez la levure que la formation de cassures double brin pouvait conduire à la conversion génique.

Le profil de méthylation des répétitions pourrait également jouer un rôle dans la maintenance du locus *RNU2* comme une entité homogène. En 1999, Jiang et Liao montrent en effet que chaque répétition présente un profil de méthylation bimodal : une région de 1,5 kb couvrant les séquences promotrices ainsi que la région codante *U2* jusqu'au microsatellite (CT)•$(GA)_n$ est déméthylée, tandis que le reste de la répétition est fortement méthylé (Figure 21) (Jiang and Liao, 1999). Il est admis que l'hypométhylation est souvent corrélée à une instabilité. Ce profil bimodal a déjà été observé pour d'autres répétitions en tandem : les gènes codant pour l'ARNsn U1 (Lund and Dahlberg, 1984), les gènes ribosomaux (Brock and Bird, 1997), et pour le macrosatellite *RS447* (Okada et al., 2002). Ce profil pourrait faciliter le recrutement de la polymérase sur les séquences promotrices, favorisant ainsi un fort taux de transcription, tout en maintenant l'intégrité du locus.

3.1.3 Localisation au sein du génome humain

3.1.3.1 Données présentes dans la littérature

Le locus *RNU2* a été localisé par des expériences de FISH à un endroit unique du génome, en 17q21-22 (Figure 22) (Hammarström et al., 1984; Lindgren et al., 1984, 1985a), et quelques années plus tard en position télomérique par rapport à *BRCA1* (Neuhausen et al., 1994; Pavelitz et al., 1995).

Les jonctions droite (416 pb dont 36 pb de l'unité répétée) et gauche (92 pb dont 47 pb de l'unité répétée), désignées comme telles arbitrairement, ont été décrites et séquencées en 1995 (Pavelitz et al., 1995), puis ordonnées en 1997 (Liao et al., 1997). La région s'organiserait ainsi : Centromère - *BRCA1* – Jonction gauche – locus *RNU2* – Jonction droite – télomère.

Tableau 5 : Les pseudogènes *RNU2*.

Noms	N° d'accession	Localisation	Taille	Autres noms / Remarques
RNU2-1	NR_002716	17q21	188 pb	*Bona fide* gene
RNU2-2P	NR_002761	11q12.3	138 pb	*RNU2B*, *RNU2-2*
RNU2-3P	NG_001258	15q26.2	266 pb	*RNU2P1*
RNU2-4P	NG_001257	17q21.31	289 pb	*U2/7*, *RNU2P2*
RNU2-5P	NG_001256	9q21.12	177 pb	*RNU2P3*
RNU2-6P	NG_032605	13 (not annotated)	190 pb	
RNU2-7P	NG_032604	13 (not annotated)	178 pb	

Tableau 6 : Degré d'homologie et de mésappariements entre les pseudogènes *RNU2*.

	RNU2-1	***RNU2-2P***	***RNU2-3P***	***RNU2-4P***	***RNU2-5P***	***RNU2-6P***	***RNU2-7P***
RNU2-1	100% / 0%	96% / 0%	93% / 0%	93% / 0%	86% / 5%	94% / 0%	80% / 7%
RNU2-2P	-	100% / 0%	90% / 0%	89% / 0%	82% / 8%	88% / 1%	82% / 0%
RNU2-3P	-	-	100% / 0%	88% / 0%	81% / 5%	88% / 0%	79% / 6%
RNU2-4P	-	-	-	100% / 0%	85% / 6%	88% / 1%	74% / 7%
RNU2-5P	-	-	-	-	100% / 0%	83% / 6%	70% / 13%
RNU2-6P	-	-	-	-	-	100% / 0%	76% / 6%
RNU2-7P	-	-	-	-	-	-	100% / 0%

3.1.3.2 Données de l'assemblage de référence

Le macrosatellite *RNU2*, ainsi que la séquence codante *RNU2-1*, ne sont pas présents dans les dernières versions de l'assemblage du génome de référence (Hg18 et Hg19). Dans la base de données du NCBI, pour la version 36 de l'assemblage du génome de référence, il est indiqué que le locus a été identifié dans un contig non-mappé, établi à partir d'un BAC non assemblé (AC087365.3). Un trou dans l'assemblage, d'environ 100 kb, était indiqué entre les clones RP11-242D8 (N° d'accession: AC060780) et CTD-3014M21 (AC109326) (Zody et al., 2006). Dans les données supplémentaires de cette étude, il est indiqué que ce trou est « flanqué de duplications segmentaires et contient plusieurs copies en tandem d'un élément HERV de 6 kb ». Dans la version 37, ce BAC a été enlevé de l'assemblage, mais il est indiqué que le locus *RNU2* est susceptible de se situer à proximité du contig AC109326.11 sur le chromosome 17. Un pseudogène, *RNU2-4P*, est présent dans l'assemblage, non loin de la position attendue du macrosatellite.

Cette absence du locus *RNU2* du génome de référence s'explique, comme nous l'avons vu, par la complexité d'assembler les régions contenant des unités répétées en tandem et de nombreuses séquences répétées, mais il est néanmoins regrettable que l'unité répétée ne soit pas présente au moins une fois au sein du génome de référence.

3.1.4 Les pseudogènes *RNU2*

Comme pour d'autres familles multigéniques, des pseudogènes *RNU2* existent au sein du génome humain, dont seuls 6 sont actuellement référencés au sein des bases de données : *RNU2-2P* à *RNU2-7P* (Tableau 5). Ces pseudogènes ne sont pas (ou plus) soumis au processus d'évolution concertée et présentent des degrés de divergence variés (Tableau 6). Les pseudogènes *RNU2* ont fait l'objet d'études approfondies uniquement au début des années 1980. Ainsi, deux pseudogènes ont été décrits en 1981 (*U2/4* et *U2/7*) en même temps que le gène *bona fide* (*U2/6*) (Westin et al.,

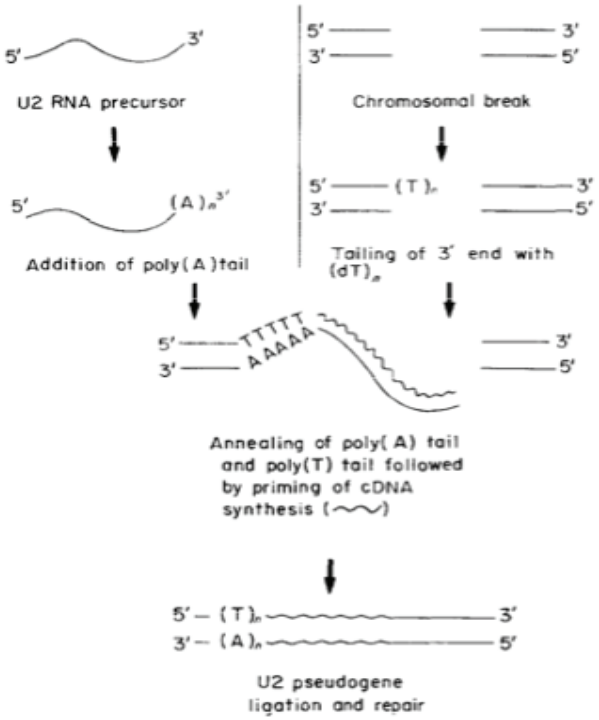


Figure 23 : Modèle de génération d'un pseudogène *U2*, le locus *U2/4*.
Le précurseur de l'ARNsn U2, se terminant bar une queue polyA, se lie à une queue polyT et est ainsi attaché à un brin d'ADN au niveau d'un site de cassure. L'ADNc est ensuite synthétisé par reverse-transcription, suivie d'une étape de réparation et de ligation.
Extrait de Hammarstrom et al., 1984.

1981). En parallèle, Denison isolait 7 phages complémentaires du gène *RNU2* : *RNU2P1* à *RNU2P7* (Denison et al., 1981). Seuls deux ont été précisément décrits. Le pseudogène *U2/4* ne présente aucune homologie avec les séquences en 5′ et en 3′ du gène *RNU2 bona fide*, et provient donc d'après l'hypothèse d'Hammarström sûrement de l'intégration d'un ADNc (ADN complémentaire) sur un autre chromosome (Figure 23) (Hammarström et al., 1984). Le pseudogène *U2/7* présente une homologie parfaite des séquences retrouvées en 5′ et 3′. Cette homologie s'étend jusqu'au microsatellite (CT)•$(GA)_n$. Il a été proposé que cette séquence fasse originellement partie du locus *RNU2* et qu'elle ait dérivée suite à l'insertion de nombreuses séquences Alu. Ce pseudogène *U2/7* (dont le symbole officiel est maintenant *RNU2-4P*) est localisé sur le chromosome 17 en 17q21.31, entre les nucléotides 41,464,597 et 41,464,885. En plus de l'homologie de séquences entre le gène et le pseudogène, on retrouve trois régions d'homologie entre les séquences entourant le pseudogène et la séquence de l'unité répétée du locus *RNU2*. Ces observations confirment vraisemblablement l'hypothèse d'Hammarström.

3.2 Stabilité du locus *RNU2*

3.2.1 Conservation au cours de l'évolution

La séquence codante du gène *RNU2* (ou celle de l'ARNsn U2) a été très conservée au cours de l'évolution, depuis le blé jusqu'à la souris (Nohga et al., 1981; Nojima and Kornberg, 1983; Skuzeski and Jendrisak, 1985), en passant par le rat (Reddy et al., 1981), le poulet et le faisan (Branlant et al., 1982) et la grenouille (Mattaj and Zeller, 1983). La taille de la séquence codante a également été conservée chez les eucaryotes (184 pb chez le xénope, 188 pb chez l'Homme), exceptée chez la levure *Saccharomyces cerevisae* (1175 pb). Cette différence de taille n'a pas été retrouvée chez la levure *Schizosaccharomyces pombe* et chez la bactérie *Escherichia coli* (Brennwald et al., 1988).

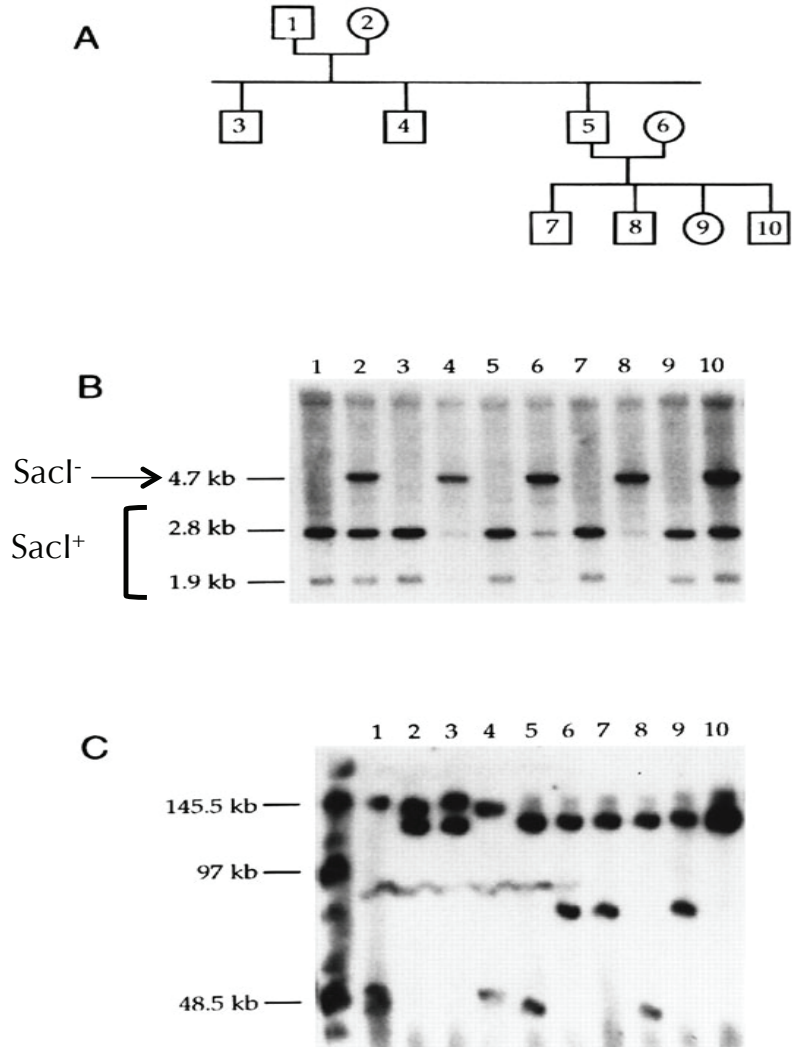


Figure 24 : Analyse de la transmission mendélienne du nombre de copies du locus *RNU2* dans une famille Amish par électrophorèse en champs pulsés (PFGE).

A. Arbre généalogique de la famille

B. Analyse par électrophorèse en champs pulsés du polymorphisme de restriction SacI au sein de l'unité répétée du locus *RNU2*

C. Analyse par électrophorèse en champs pulsés du locus *RNU2* après digestion par EcoRI.

Extrait de Liao et al., 1997.

L'organisation en tandem des gènes *RNU2* a été observée chez de nombreux organismes comme l'oursin (Card et al., 1982), les rongeurs (Dahlberg and Lund, 1988; Nojima and Kornberg, 1983) et les prosimiens (Matera et al., 1990). Les répétitions sont également organisées en tandem chez les singes de l'Ancien Monde (babouin, macaque) et les singes du Nouveau Monde (gibbon, orang-outan, chimpanzé et gorille) (Matera et al., 1990; Pavelitz et al., 1995). Comme chez l'Homme, la taille de l'unité répétée est de 6 kb chez les singes du Nouveau Monde, mais dépasse 11 kb chez les singes de l'Ancien Monde (Matera et al., 1990). Cette différence de 5 kb est due à l'excision d'une séquence pro-virale, contenue entre deux LTRs (Long Terminal Repeat). Le locus *RNU2* a conservé le même contexte chromosomique chez les primates, présumant d'une évolution concertée depuis plus de 35 millions d'années (Pavelitz et al., 1995). Cependant, plusieurs réarrangements et divergences ont pu être observés au niveau des jonctions du locus (Pavelitz et al., 1999).

3.2.2 Stabilité méiotique et mitotique du locus

En 1997, l'équipe de Pavelitz a étudié par PFGE la stabilité du locus chez 10 individus issus d'une famille Amish, et a ainsi montré que le nombre de répétitions, ainsi que le polymorphisme Sac I au sein de chaque répétition, sont transmis de façon mendélienne (Figure 24) (Liao et al., 1997).

En étudiant 60 trios père-mère-enfant, issus de populations caucasienne et africaine, Schaap *et al.* n'ont constaté qu'un seul changement du nombre de copies au cours d'une génération, entre une mère africaine et son enfant (Schaap et al., 2013). Le père porte respectivement deux allèles avec 6 et 29 copies, la mère deux allèles avec 14 copies et l'enfant un allèle avec 6 copies et un allèle avec 15 copies, laissant penser à une expansion *de novo* d'une seule copie (de 86 kb à 94 kb, soit de 14 à 15 copies). Le taux d'instabilité méiotique du locus serait donc d'environ 0,8 % par génération.

Schaap et ses collaborateurs ont observé un mosaïcisme chez 4 individus sur les 210 étudiés : une chinoise (20 copies, 22 copies et 25 copies), une japonaise (deux allèles à 7 copies et un allèle à 28 copies), un enfant africain (24, 25 et 37 copies) et un

père africain (8, 12 et 13 copies). D'après leur étude, l'instabilité mitotique de ce locus serait donc d'environ 1,5 %.

3.3 Le locus *RNU2*, un site fragile du génome humain

Chaque chromosome du génome humain contient des régions instables qui subissent préférentiellement des cassures, le plus souvent suite à une exposition à une grande variété d'agents infectieux ou de drogues. Parmi eux, le locus *RNU2* a été décrit comme étant un site de fragilité suite à une infection par l'Adénovirus 12 (Durnam et al., 1988; Li et al., 1993; Lindgren et al., 1984). Trois autres sites de fragilité suite à l'infection par l'Adénovirus 12 ont été décrits et colocalisent avec des larges répétitions en tandem codant pour des petits ARN structuraux abondants : le locus *RNU1* en 1p36, le locus *RN5S* en 1q42-43 (codant pour les ARN ribosomaux, ou ARNr, 5S) et le locus *PSU1* en 1q21-22 (contenant des pseudogènes de *RNU1*) (Bernstein et al., 1985; Lindgren et al., 1985b; Schramayr et al., 1990; Sørensen et al., 1991).

La fragilité du locus *RNU2* a été montrée comme étant dépendante de la présence des séquences promotrices (Bailey et al., 1995; Gargano et al., 1995; Li et al., 1993), mais indépendante de la présence des séquences flanquantes du locus (Liao et al., 1997; Pavelitz et al., 1995). Par la suite, il a été montré que la fragilité du locus *RNU2* pouvait également être induite suite à l'expression transitoire de la protéine E1B 55kDa de l'Adénovirus 12 (Liao et al., 1999), à des mutations dans la protéine CSB (Cockaine Syndrome B), et à un traitement par des agents entraînant des dommages à l'ADN tels que l'actinomycine D (un inhibiteur de la transcription) (Yu et al., 1998) et l'arabinoside cytosine (ou araC, un inhibiteur des polymérases) (MacArthur et al., 1997). De manière très intéressante, il a été montré que la transcription du locus *RNU2* est requise pour sa fragilité (Bailey et al., 1995; Gargano et al., 1995), bien qu'indépendante du nombre de copies (Bailey et al., 1995). L'activité de la protéine p53 est également requise pour la fragilité du locus (Li et al., 1998a, 1998b). Il a été proposé que la protéine CSB, en plus de son rôle dans la réparation couplée à la transcription, pourrait intervenir comme facteur d'élongation lors de la transcription de

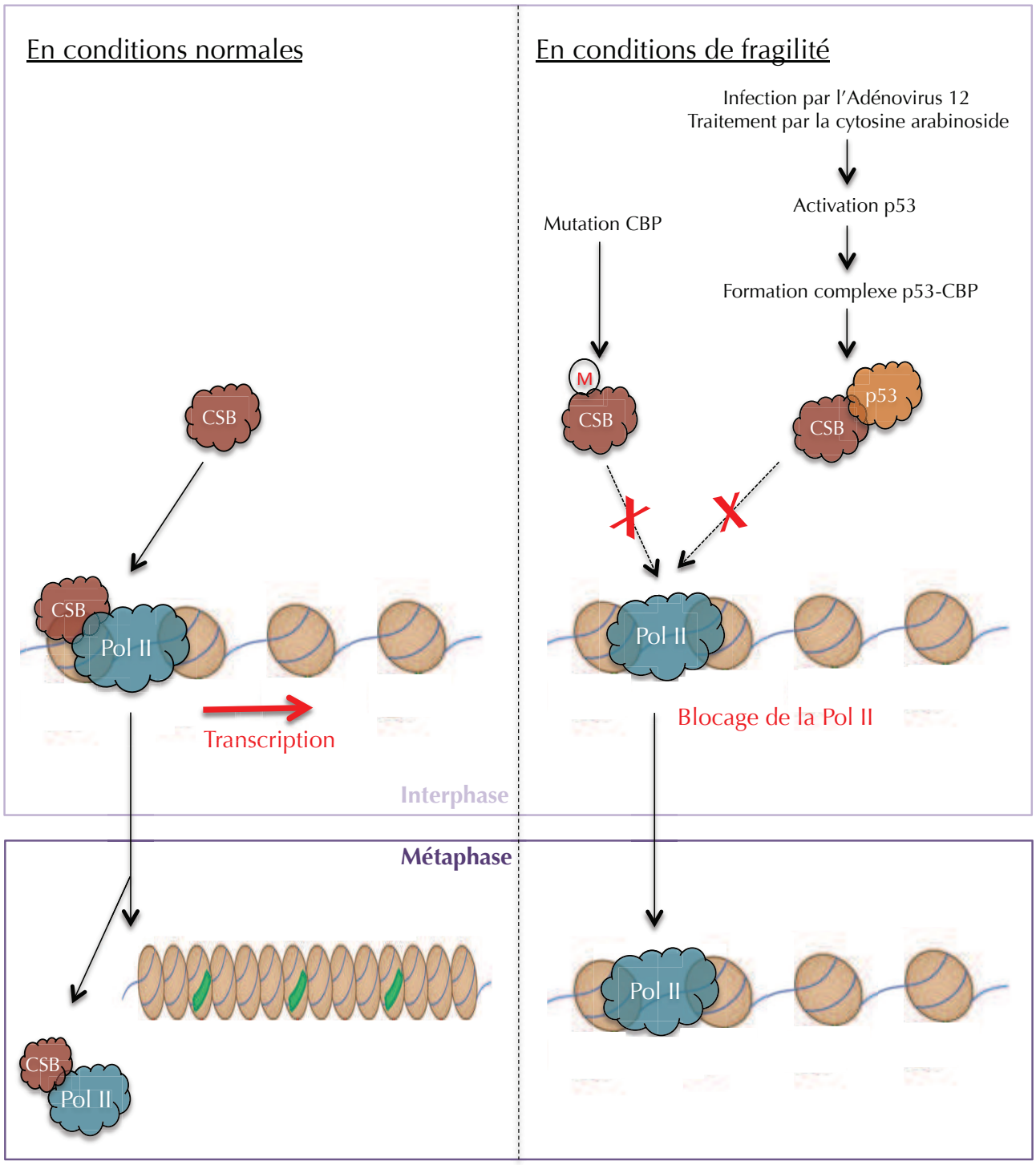


Figure 25 : Modèle proposé pour la fragilité métaphasique au locus *RNU2*.
En conditions normales, CBP se lie à la Polymérase II (Pol II) et agit comme un facteur d'élongation, empêchant ainsi la Pol II de stagner sur le locus *RNU2* fortement structuré. En métaphase, la chromatine est condensée.
Lors d'une infection par l'Adénovirus 12, un traitement par la cytosine arabinoside ou une mutation sur CBP, CBP ne se lie plus à la Pol II, qui reste bloquée sur le locus, et empêche ainsi la condensation de la chromatine en métaphase. Ceci se traduit par une fragilité du locus.
Adapté de Yu et al., 2000.

petits ARNs hautement structurés et transcrits par la polymérase II (pol II) et la polymérase III (pol III). L'infection par l'Ad12 ou le traitement par l'actinomycine D ou l'araC entraîneraient l'activation de la protéine p53, conduisant à la formation d'un complexe entre p53 et CSB. L'interaction entre ces deux protéines a été constatée *in vivo*. La formation de ce complexe provoque la perte de la fonction hélicase de la protéine CSB. Par ailleurs, une mutation de la protéine CSB a également été associée à une fragilité constitutive du locus *RNU2* (Yu et al., 2000). Ces données ont permis d'élaborer le modèle selon lequel cette mutation de CSB engendrerait le blocage de la polymérase sur le locus *RNU2*, interférant avec la condensation de la chromatine pendant la métaphase, et induisant de ce fait une fragilité du locus. En appui de ce modèle, Pavelitz et ses collaborateurs ont récemment montré que la protéine SNAPc, nécessaire à la transcription des ARNsn, reste anormalement fixée au niveau du PSE (Proximal Sequence Element) pendant la métaphase en conditions de fragilité du locus (Pavelitz et al., 2008).

Selon le modèle proposé, l'ARNsn U2 étant un composant essentiel de la cellule requis en grande quantité, les séquences promotrices montrent un état 'ouvert' permanent, afin de faciliter l'initiation rapide de la transcription à ce locus (Figure 25). Cet état d'ouverture est perdu transitoirement pendant la métaphase, afin de permettre la condensation de la chromatine. Cette transition entre ouverture et compaction représente un compromis entre une initiation rapide de la transcription et une éventuelle fragilité métaphasique. Cependant, dans certaines conditions (*i.e.* infection par l'Ad12, traitement par l'actinomycine D, surexpression de p53, perte de BRCA1 ou BRCA2, mutation de CSB,…), la polymérase stagne au locus, empêchant ainsi la condensation de la chromatine et entraînant une fragilité. Tout comme les locus *RNU1* et *RNU5S*, le locus *RNU2* est un site fragile induit par un défaut de la transcription, et non suite à un défaut de réparation pendant la réplication comme d'autres sites bien connus tels que *FRAXA* impliqué dans le syndrome du X fragile (Kremer et al., 1991).

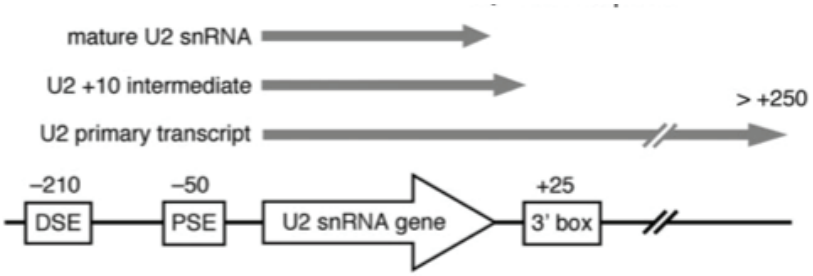


Figure 26 : Eléments régulateurs de la transcription de l'ARNsn U2.
Les positions des éléments en amont de l'ARNsn U2 mature sont indiqués en nombres négatifs, tandis que celles des éléments en aval en nombres positifs. La transcription du gène U2 produit dans un premier temps un transcrit primaire d'environ 1 kb, qui sera par la suite maturé en un transcrit mature de 188 pb.
DSE: Distal Sequence Element
PSE: Proximal Sequence Element
Extrait de Jacobs et al., 2004.

3.4 Transcription du gène *RNU2*

L'ARNsn U2 est fortement exprimé et de façon ubiquitaire (Egloff et al., 2008; Hernandez, 2001). Bien que le nombre de copies du gène *RNU2* varie d'un individu à l'autre, le niveau d'expression de l'ARNsn U2 ne varierait pas entre les individus. Ainsi, le niveau d'expression de l'ARNsn U2 ne serait pas corrélé au nombre de copies du locus (Bailey et al., 1995; Mangin et al., 1985). Chaque allèle semble être soumis à une compensation de dosage : seulement quelques copies sont exprimées, les autres ne sont pas transcrites.

Pour arriver à une abondance proche de celles des ARNr sachant que probablement peu de gènes *U2* sont actifs, le taux de transcription de chaque gène doit être incroyablement élevé. Le taux d'initiation de la transcription pour *U2* serait de 1,2/s contre 0,2/s pour les ARNr (Pavelitz et al., 2008).

Le transcrit mature de l'ARNsn U2 comporte 188 pb, et n'est ni épissé ni polyadénylé comme les autres ARNsn. La transcription du gène *RNU2* se fait grâce à la polymérase II (Pol II), et nécessite le recrutement du large complexe multiprotéique Integrator (Baillat et al., 2005). Deux séquences activatrices ont été décrites : le PSE, qui mime les fonctions de la boîte TATA en fixant par exemple les facteurs de transcription spécifiques des ARNsn (SNAPc : snRNA-activating protein complex), et la séquence DSE qui fixe les facteurs Sp1 et Oct1 (Rincon et al., 1998) (Figure 26). Il a été montré que la transcription continue après la fin de la séquence du gène *RNU2*, sur environ 800 nt déplétés en nucléosomes (Egloff et al., 2009), aboutissant ainsi à un transcrit primaire d'environ 1 kb (Cuello et al., 1999) (Figure 26).

La terminaison de la transcription de l'ARNsn U2, au niveau d'un signal en 3′ hautement conservé, communément appelé boîte 3′ (3′-box) (Yuo et al., 1985) est un processus finement régulé et est couplé à la maturation en 3′. Un facteur spécifique, PTF, participe à la déplétion en nucléosomes de la région, favorisant ainsi le passage de la pol II, tandis que le facteur d'élongation négatif, NELF, promeut la terminaison de la transcription à la fin du gène lorsque le niveau de nucléosomes augmente (O'Reilly et

al., 2014). Ceci suggère un rôle de la structure chromatinienne dans la régulation de la terminaison de la transcription *in vivo*. L'activité endonucléase du complexe Integrator est requise pour l'étape de clivage (O'Reilly et al., 2014). Par ailleurs, il a été montré que la terminaison de la transcription et la maturation de l'extrémité 3' de l'ARsn U2 sont bloquées par de faibles doses d'actinomycine D ou une irradiation aux UVs (Jacobs et al., 2004).

Le processus de maturation de l'extrémité 3' se produit dans le cytoplasme, où le transcrit intermédiaire est exporté (Eliceiri and Sayavedra, 1976; Huang et al., 1997). Après ajout d'une coiffe triméthylée en 5' ainsi que l'arrivée des protéines Sm (Mattaj, 1986; Mattaj and De Robertis, 1985), la petite ribonucléoprotéine nucléaire (small nuclear ribonucleoprotein, ou snRNP) U2 est importée dans le noyau (Huber et al., 1998), où des modifications de bases surviennent (Darzacq et al., 2002).

3.5 L'ARNsn U2 : un composant de la machinerie d'épissage

3.5.1 Rôle de l'ARNsn U2 lors de l'épissage des ARNm humains

L'ARNsn U2 est un composant majeur du spliceosome, également appelé particule d'épissage. Cette machinerie est responsable de l'épissage d'environ 90 % des pré-ARN messagers humains. Le spliceosome est constitué principalement de snRNPs, provenant de l'association entre les ARNsn et 6 à 10 protéines.

L'épissage des pré-ARNm se fait selon les étapes suivantes :

1) Appariement des ARNsn U1 et U2 avec le pré-ARNm : recrutement de l'ARNsn U1 au niveau de la jonction exon/intron du pré-ARNm, et liaison de l'ARNsn U2 avec le point de branchement au sein de l'intron ;
2) Appariement des ARNsn U4 et U6, et association avec l'ARNsn U5 ;
3) Association du complexe U4/U6/U5 avec le complexe U1/U2/pré-ARNm ;

4) Remaniements des snRNP : libération des ARNsn U1 et U4, fixation de l'ARNsn U6 sur l'extrémité 5' du site de clivage, glissement de l'ARNsn U5 sur l'intron ;

5) Première réaction de transestérification, catalysée par U2 et U6 ;

6) Liaison des deux exons, formation de la structure en lasso et libération de l'intron ;

7) Dissociation du complexe.

3.5.2 Implication dans des maladies neurodégénératives ?

Chez la souris, on retrouve un cluster de 5 gènes codant pour l'ARNsn U2 : de *Rnu2-6* à *Rnu2-10*. Tous sont identiques, mis à part un SNP dans la séquence de *Rnu2-6*. L'équipe de Jia a observé qu'une délétion de 5 nucléotides au sein d'une région hautement conservée dans une seule de ces répétitions entraîne une neurodégénération et de graves défauts d'épissage (Jia et al., 2012). Cette délétion au sein de *Rnu2-8* cause la perte de la séquence de reconnaissance du site de branchement de l'ARNsn U2. En comparant le niveau d'expression de *Rnu2-8* chez la souris porteuse de la délétion (*Rnu2-8Δ)* et chez la souris sauvage (WT), les auteurs ont remarqué une plus forte expression de cet ARNsn mutant dans le cerebellum. Ils ont également montré que le niveau d'expression de cet ARNsn variait au cours du développement post-natal, en atteignant son niveau le plus haut au cours de la maturation des neurones granulaires.

Cette étude est la première à s'intéresser à l'effet sur l'épissage des pré-ARNm d'une mutation dans l'une des répétitions du locus *RNU2*, et également aux conséquences de cette mutation sur la transcription de cette copie. Les auteurs montrent pour la première fois qu'il existe une régulation indépendante, aussi bien temporelle que spatiale, de l'expression de chaque copie au sein d'une répétition de gènes.

Ce travail ouvre de nombreuses perspectives pour identifier de nouvelles mutations liées à l'apparition de maladies au sein de répétitions de gènes, en étudiant à la fois la variation du nombre de copies du locus mais également des variants génétiques au sein de chaque copie.

3.5.3 L'ARNsn U2, un biomarqueur diagnostique pour certains cancers ?

Il a été montré que des ARN non codants peuvent générer différents produits, d'une part suite à leur dégradation, mais également par différentes étapes impliquant des protéines telles que Dicer (Cole et al., 2009; Fu et al., 2009). Quatre études récentes ont mis en évidence que l'ARNsn U2 ou des fragments dérivés de cet ARN pourraient constituer des biomarqueurs pour le dépistage de plusieurs types de cancer (Baraniskin et al., 2014, 2014; Kuhlmann et al., 2014; Mazières et al., 2013).

Dans un premier temps, Baraniskin et ses collaborateurs ont identifié un fragment dérivé de U2 (RNU2-1f) retrouvé de manière stable dans le sérum et le plasma. Le taux de RNU2-1f circulant permet de discriminer des patients atteints d'un cancer colorectal ou d'un cancer pancréatique d'individus témoins (Baraniskin et al., 2013). Dans un second temps, Mazières et ses collaborateurs se sont penchés sur les taux d'ARN circulant chez les patients atteints de cancer du poumon et chez des individus témoins. Cette étude a montré que l'épissage alternatif produit effectivement des petits ARN régulateurs (RNU2-derived smRNAs). Parmi ces dérivés, le miR-U2 (19-22 nt) est fortement exprimé dans le tissu pulmonaire et exporté dans le sang. Une sur-expression de miR-U2 a été retrouvée chez les patients atteints d'un cancer du poumon (Mazières et al., 2013). Récemment, Baraniskin et ses collaborateurs ont également montré que le taux d'ARNsn U2 dans le fluide biliaire était un excellent biomarqueur pour le cholangiocarcinome, une tumeur des cellules épithéliales des voies bilaires (Baraniskin et al., 2014). Enfin, Kuhlman et ses collaborateurs ont montré que le taux de RNU2-1f dans le sérum était plus élevé chez les femmes atteintes d'un cancer de l'ovaire. De plus, la persistance d'un taux élevé de RNU2-1f après la chirurgie et la chimiothérapie permettrait de distinguer un sous-groupe de patientes à haut risque de rechute et avec un mauvais pronostic (Kuhlmann et al., 2014).

Les taux d'ARNsn U2 ou d'ARNs dérivés de U2 circulant pourraient ainsi être utilisés dans le cadre d'une stratégie non-invasive de détection précoce de certains types de cancer.

4 *BRCA1* et la prédisposition génétique au cancer du sein

4.1 Du gène à la protéine BRCA1

4.1.1 Le gène *BRCA1*

4.1.1.1 Structure du gène

Le gène *BRCA1*, un gène suppresseur de tumeur, est localisé en 17q21 et constitué de 23 exons répartis sur 81 kb d'ADN. La séquence codante est de 5,6 kb. La taille des exons varie entre 36 pb (pour l'exon 18) et 3426 pb (pour l'exon 11). L'exon 11 est l'un des plus grands exons chez les mammifères et représente à lui seul 60 % de la séquence codante de *BRCA1*. Les séquences introniques sont de tailles variées (de 403 pb à 9,2 kb) et représentent 91 % du gène.

4.1.1.2 Région régulatrice et gènes du bloc de déséquilibre de liaison

Le promoteur de *BRCA1* s'étend sur environ 2000 pb, dont une région minimale située dans la région intergénique de 56 pb et ne contenant pas de boîte TATA. Plusieurs séquences régulatrices ou éléments régulateurs ont été identifiés en amont ou au sein du gène *BRCA1* :

- Une séquence RIBS contenant un site de fixation pour le facteur de transcription GABPα/β (GA-Binding Protein) constitue un activateur transcriptionnel de *BRCA1*, dont l'activité varie en fonction des lignées cellulaires (Atlas et al., 2000).

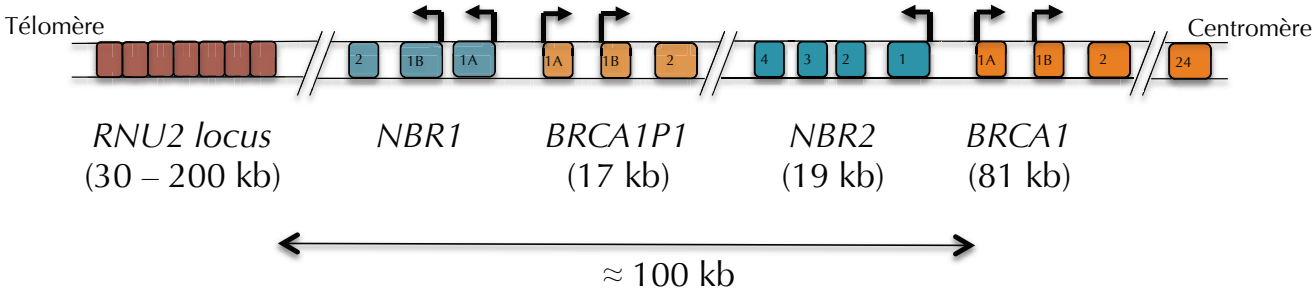


Figure 27 : Organisation de la région en amont de *BRCA1* chez l'Homme.
NBR1: Next *BRCA1* gene 1
NBR1: Next *BRCA1* gene 2
BRCAP1: pseudogène *BRCA1*
Adapté de Liu and Barker, 1999.

- Une région PRR (Positive Regulatory Région) de 36 pb, riche en purines/pyrimidines et contenant une séquence proche de la séquence consensus de liaison CREB (Cyclic AMP Regulatory Element Binding Protein), joue un rôle pour l'expression constitutive du promoteur (Thakur and Croce, 1999).
- Des élements, partiels ou imparfaits, de réponse aux estrogènes (Estrogen Response Element, ou ERE) et à la progestérone (PRE : Progesterone Response Element) ont été mis en évidence au sein des introns 2 et 7 ainsi que 2 kb en amont du site d'initiation de la transcription (Atlas et al., 2001; Smith et al., 1996).
- Un large îlot CpG de 2,7 kb situé entre l'intron 1 de *NBR2* et l'exon 1B de *BRCA1* pourrait jouer un rôle de répresseur transcriptionnel par sa méthylation (Magdinier et al., 2000).

L'organisation de la région en amont de *BRCA1* est extrêmement complexe, suite à la duplication partielle de la région 5' de *BRCA1* (Figure 27). Ainsi, le pseudogène *BRCA1* (*BRCA1P1),* localisé environ 30 kb en amont de *BRCA1,* contient les exons 1A, 1B et 2 de *BRCA1*. Ce pseudogène est placé « tête-bêche » avec le gène *NBR1* (Neighbor of *BRCA1* gene 1), s'étendant sur 41 kb, avec lequel il partage un promoteur bidirectionnel (Brown et al., 1996). Cette duplication a également conduit à la création en amont de *BRCA1* du gène *NBR2* (Neighbor of *BRCA1* gene 2), contenant les exons 1A, 1B et 2 du gène *NBR1*. *NBR2* est localisé « tête-bêche » avec *BRCA1* et partage avec lui un promoteur bidirectionnel de 218 pb (Xu et al., 1997).

Plusieurs études ont montré que le gène *BRCA1*, ainsi que les gènes voisins (*NBR2*, *BRCA1P1, NBR1*, *TMEM106A*), étaient localisés dans un bloc de déséquilibre de liaison d'environ 290 kb (Bonnen et al., 2002; Freedman et al., 2005; Liu and Barker, 1999). Ces régions en déséquilibre de liaison sont caractérisées par une faible occurrence d'événements de recombinaison. On observe en conséquence une association forte entre différents marqueurs de la région, permettant de définir des blocs haplotypiques.

4.1.1.3 Mutations germinales de *BRCA1*

Plus de 1700 mutations germinales de *BRCA1* ont été identifiées à ce jour, dont 55 % sont uniques, et sont recensées dans différentes bases de données, dont la base de données BIC (The Breast cancer Information Core database) et la base française UMD-BRCA1 (Universal Mutation Database). Quelques mutations fondatrices ont été identifiées dans certains groupes ethniques : les mutations 185delAG et 5283insC sur *BRCA1* et 6174delT sur *BRCA2* dans la population juive ashkénaze, et la mutation 999del5 sur *BRCA2* chez les islandais. Ces mutations sont réparties sur l'ensemble du gène *BRCA1*, sans qu'aucun point chaud mutationnel n'ait pu être mis en évidence à l'exception d'un point chaud créé par la duplication ayant conduit à la création du pseudogène (Puget et al., 2002).

La plupart des mutations identifiées sont des mutations ponctuelles ou des petites insertions / délétions. Cependant, environ 10 % des mutations sont des grands réarrangements, tels que des pertes ou des duplications d'exons (Mazoyer, 2005). Environ une trentaine de réarrangements est répertoriée, le plus connu étant la duplication de l'exon 13 (ins6kbEx13) (The BRCA1 Exon 13 Duplication Screening Group, 2000). Bien que pouvant être silencieuse, faux-sens, non-sens, ou entraînant une modification d'épissage, la plupart des mutations identifiées à ce jour entraîne un décalage du cadre de lecture (frameshift mutations). Elles résultent alors en un codon stop prématuré, et ont pour conséquence une protéine tronquée lorsque le variant n'est pas pris en charge par le NMD (Nonsense Mediated mRNA Decay) (Perrin-Vidoz et al., 2002). Ces mutations sont donc essentiellement à l'origine d'une perte de fonction de la protéine BRCA1.

4.1.2 Le transcrit *BRCA1*

La transcription de *BRCA1* est initiée à partir de deux sites séparés par 277 pb, permettant d'inclure alternativement les exons 1A et 1B (Xu et al., 1995). Les deux

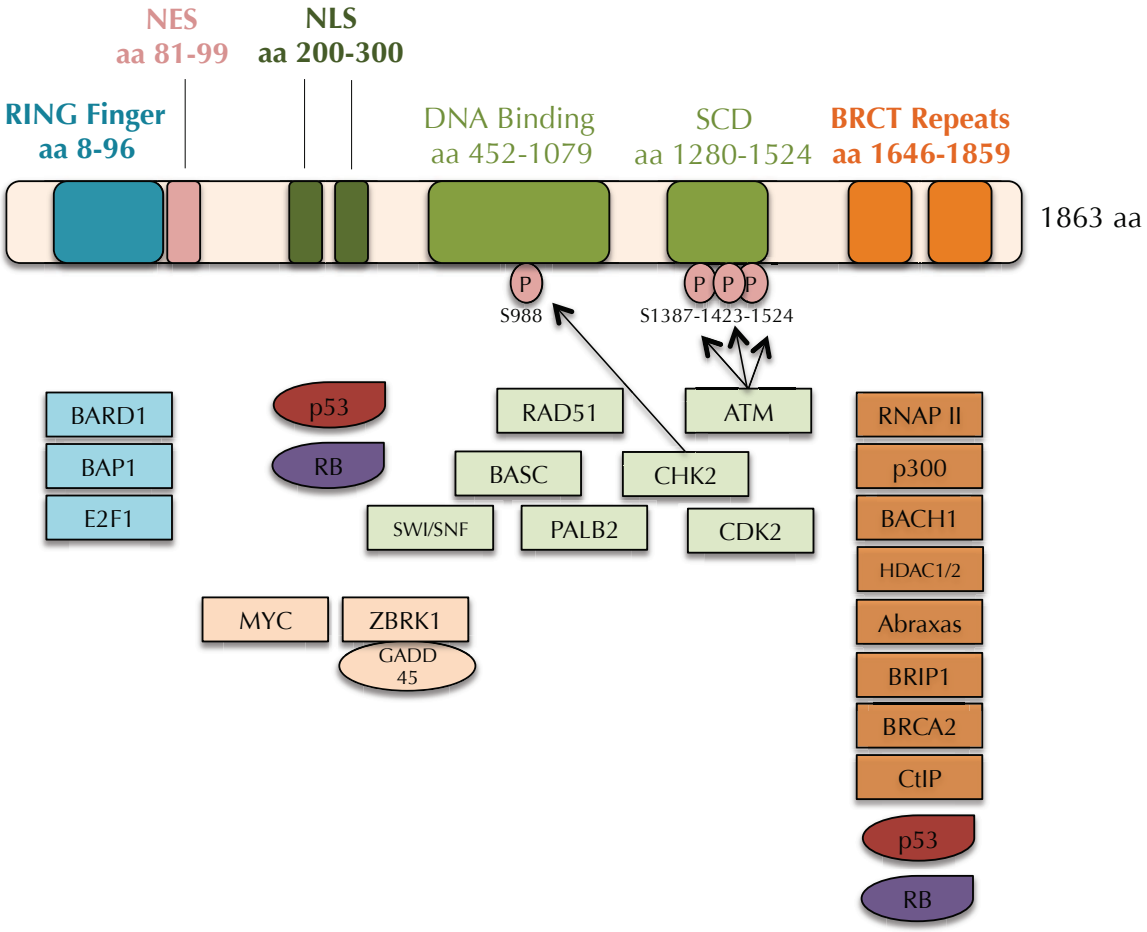


Figure 28 : Domaines fonctionnels et interactants de BRCA1.
NLS: Nuclear Localisation Signal
NES: Nuclear Export Signal
SCD: Serine Containing Domain
BRCT: BRCA1 C-terminus Domain
S: Sérine
Adapté de Narod and Foulkes, 2004.

transcrits sont ubiquitaires, bien que le transcrit 1A est majoritaire dans la glande mammaire et le 1B dans le placenta (Xu et al., 1995).

Le transcrit *BRCA1* canonique s'étend sur 7,8 kb. Après épissage, il est constitué de 22 exons et correspond à un cadre de lecture ouvert (ORF) de 5589 pb. L'extrémité 5'UTR (Untranslated Region) comprend l'exon 1 non codant ainsi que le début de l'exon 2. L'extrémité 3'UTR est constituée de 1396 pb et comprend pratiquement tout l'exon 24. A ce jour, plusieurs variants d'épissage, présents dans différents tissus, ont été identifiés. Les plus étudiés sont les transcrits *BRCA1Δ11*, *BRCA1Δ11b*, *BRCA1Δ9,10*, et *BRCA1Δ9,10,11b* et *BRCA1-IRIS*. Toutefois, les rôles potentiels de ces variants d'épissage sont encore peu connus.

4.1.3 La protéine BRCA1

4.1.3.1 Domaines structuraux

Le transcrit sauvage est traduit en une protéine multifonctionnelle ubiquitaire de 1863 acides aminés (aa) (220 kDa) (Figure 28). Les régions N- et C-terminales sont fortement conservées chez les mammifères (Abel et al., 1995; Szabo et al., 1996), et contiennent les deux domaines structuraux actuellement caractérisés.

En N-terminal, le domaine RING (Really Interesting New Gene) s'étend des acides aminés 1 à 109 (Meza et al., 1999). Il est composé d'un domaine zinc finger entouré de deux hélices alpha antiparallèles. Ces hélices permettent l'interaction de BRCA1 avec BARD1, conférant une activité d'E3-Ubiquitine ligase à BRCA1 (Hashizume et al., 2001; Mallery et al., 2002).

En C-terminal, le domaine BRCT (BRCA1 C-terminal) est composé d'un tandem de modules BRCT connectés par une charnière de 23 aa. Les tandems BRCT ont la capacité d'interagir avec des phosphoprotéines (Manke et al., 2003; Yu et al., 2003).

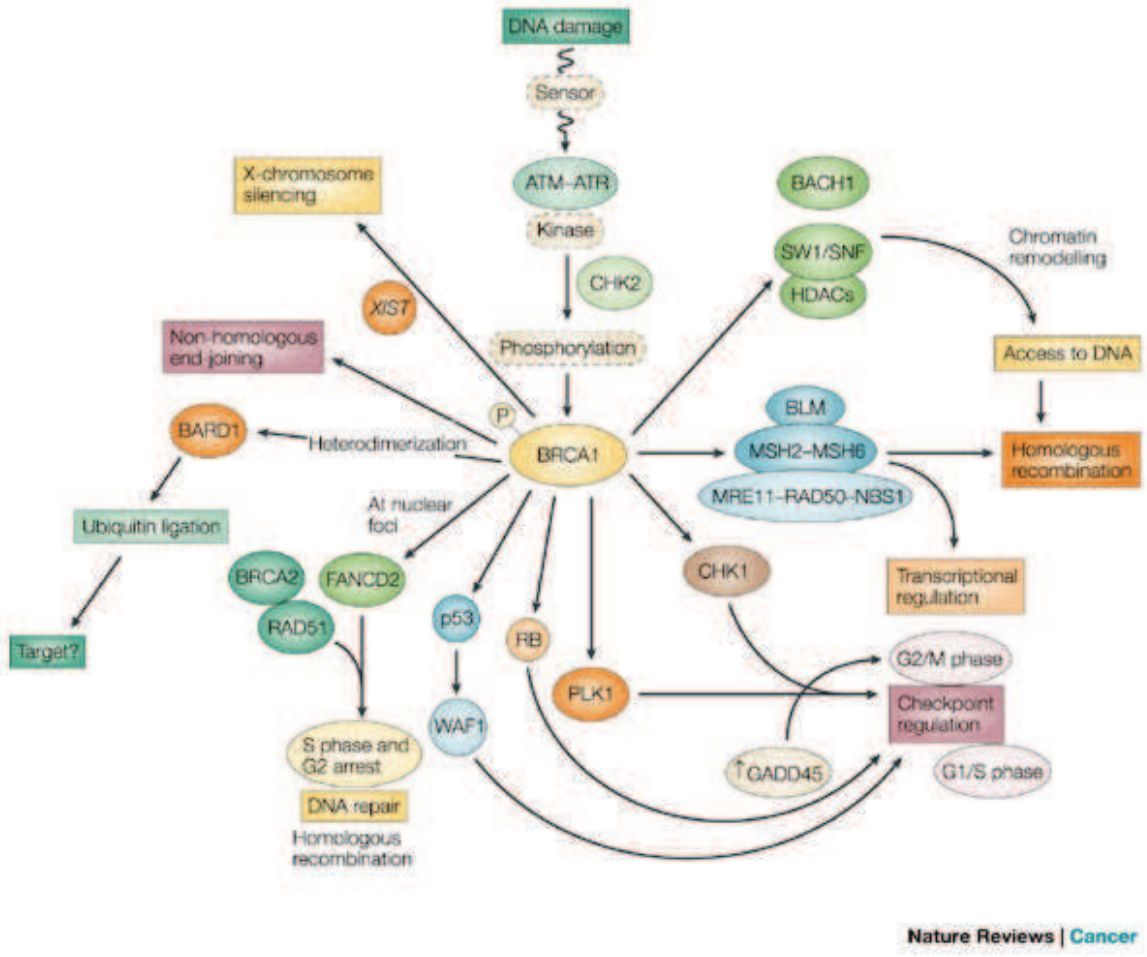


Figure 29 : Interactants et fonctions de BRCA1.
Extrait de Narod and Foulkes, 2004.

4.1.3.2 Localisation subcellulaire

Un signal d'import nucléaire (Nuclear Localisation Signal, ou NLS) et un signal d'export nucléaire (Nuclear Export Signal, ou NES) ont pu être identifiés au sein de la région centrale de BRCA1 (Henderson, 2005; Miki et al., 1994).

Longtemps considérée comme exclusivement nucléaire et bien qu'elle soit principalement localisée dans le noyau, la protéine BRCA1 est également retrouvée dans le cytoplasme et au niveau de la mitochondrie (Coene et al., 2005).

4.1.3.3 Fonctions cellulaires de BRCA1

Plusieurs études chez la souris et des études *in vitro* chez l'Homme ont souligné le rôle important de BRCA1 dans la différenciation de la glande mammaire et dans l'embryogenèse tant chez la souris que chez l'Homme (Gowen et al., 1996; Hakem et al., 1996; Liu et al., 1996; Magdinier et al., 1999).

BRCA1 est une protéine multifonctionnelle possédant de nombreux interactants protéiques (Figure 28). BRCA1 est impliquée dans de nombreuses voies cellulaires, particulièrement dans le contrôle de l'intégrité cellulaire dont elle est un acteur majeur (Figure 29). De fait, BRCA1 interagit, *via* ses multiples domaines fonctionnels, avec de nombreuses protéines impliquées dans la réparation des dommages à l'ADN ou dans la régulation du cycle cellulaire.

BRCA1 agit comme un senseur et un médiateur entre les protéines de détection des cassures double brin (CDB) et les protéines impliquées directement dans la réparation. BRCA1 est principalement impliquée dans la voie de réparation par recombinaison homologue (HR) et également dans la réparation par NHEJ, mais pourrait intervenir dans d'autres voies.

Par ailleurs, de nombreuses études suggèrent que BRCA1 participe à la stimulation de la transcription (Anderson et al., 1998; Krum et al., 2003; MacLachlan et al., 2002; Ouchi et al., 1998; Schlegel et al., 2000; Scully et al., 1997; Zhang et al.,

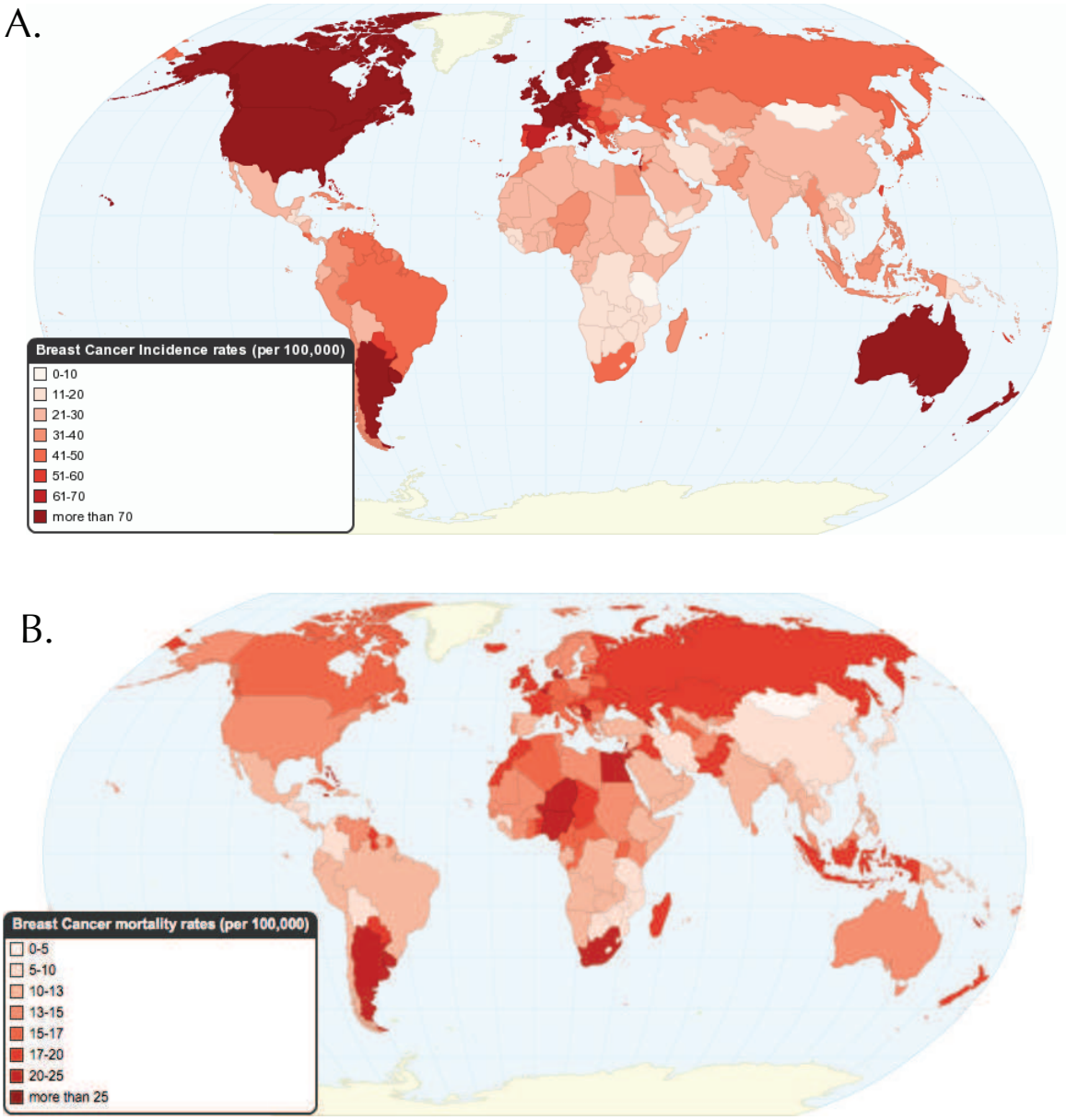


Figure 30 : Données épidémiologiques sur le cancer du sein.
A. Taux d'incidence
B. Taux de mortalité
Extrait de Ferlay et al., 2012.

1998). BRCA1 est également impliquée dans les deux activités enzymatiques permettant le remodelage de la chromatine : l'acétylation des histones par les acétyl-transférases (HAT) (Aprelikova et al., 1999; Pao et al., 2000) et le remodelage de la chromatine par le complexe SWI/SNF (Bochar et al., 2000).

BRCA1 a également été impliquée dans la régulation du nombre et de la fonction des centrosomes, dans la balance vie/mort cellulaire (notamment en réponse au stress), dans la lipogenèse (Brunet et al., 2008; Ray et al., 2006), dans la biogenèse des microARNs (Kawai and Amano, 2012), dans la motilité cellulaire (Yasmeen et al., 2008) et dans la régulation de la traduction (Dacheux et al., 2013).

4.2 Le Cancer du sein : données épidémiologiques et facteurs de risque

4.2.1 Données épidémiologiques

En 2012, en France, le cancer du sein est le cancer le plus fréquent chez la femme avec 48 800 nouveaux cas par an, devant le cancer colorectal (18 900 nouveaux cas) et le cancer du poumon (11 300 nouveaux cas) (Binder-Foucard et al., 2013). C'est également la première cause de décès par cancer chez la femme (11 886 décès en 2012). Au niveau mondial, on recense 1,7 millions de nouveaux cas de cancer du sein par an. C'est le cancer le plus fréquemment diagnostiqué chez les femmes dans 140 des 184 pays couverts par GLOBOCAN, et celui entraînant le plus grand nombre de décès chez les femmes (522 000 décès par an) (Figure 30) (Bray et al., 2013; Ferlay et al., 2013). On estime qu'une femme sur 9 à 12 femmes développe un cancer du sein au cours de sa vie.

4.2.2 Facteurs de risque hormonaux et environnementaux

- Genre, âge et ethnie

Le cancer du sein est un cancer quasiment exclusivement féminin, moins d'un cancer du sein sur 100 étant développé par un homme. L'âge est ensuite le facteur de risque le plus important du cancer du sein, l'âge moyen de diagnostic étant, en France, de 60 ans. Environ 10 % des cas de cancer du sein se manifestent chez les femmes âgées de moins de 35 ans et près de 20 % avant 50 ans. Près de 50 % des cancers du sein sont diagnostiqués entre 50 et 69 ans et environ 28 % sont diagnostiqués après 69 ans.

Un risque plus important de développer un cancer du sein a été observé pour des individus d'origine ethno-géographique particulière, le plus souvent à cause d'un effet fondateur d'une mutation sur les gènes *BRCA1/2*. C'est le cas pour les populations juive ashkénaze, islandaise, et également polonaise (ainsi que dans une moindre mesure les populations mexicaine, québécoise et finlandaise). On observe également une grande variabilité dans l'incidence et la mortalité du cancer du sein par pays, reflétant les différences de modes de vie et d'exposition à l'environnement (Figure 30) (Bray et al., 2013; Ferlay et al., 2013).

- Histoire familiale

Après l'âge, l'histoire familiale est le second facteur de risque. En effet, la probabilité de développer un cancer du sein augmente lorsqu'une parente au premier ou deuxième degré a été atteinte. Ainsi, près de 20 à 30 % des cancers du sein sont diagnostiqués chez des femmes ayant des antécédents familiaux de cancers du sein.

Lorsque plusieurs cas d'un ou plusieurs type(s) de cancer(s) ont été diagnostiqués à un âge variable à l'intérieur d'une même famille, on parle alors de cancers familiaux.

Souvent, aucun signe de transmission mendélienne d'une prédisposition n'est apparent, et ces cancers familiaux ne présentent pas les caractéristiques classiques des syndromes de cancers héréditaires.

- Facteurs hormonaux et vie reproductive

Les œstrogènes (les hormones ovariennes) ont un rôle promoteur dans le cancer du sein. Ainsi, une durée prolongée d'exposition aux hormones endogènes déterminée par une puberté précoce (< 12 ans) et une ménopause tardive (> 54 ans) augmente sensiblement le risque de cancer du sein. Par ailleurs, une première grossesse avant 30 ans diminue le risque de cancer du sein de 25 % (Ewertz et al., 1990). Plusieurs grossesses ainsi que l'allaitement auraient un effet protecteur (Collaborative Group on Hormonal Factors in Breast Cancer, 2002). Par ailleurs, l'utilisation d'hormones de synthèse pourrait contribuer à augmenter le risque de cancer du sein, bien que seulement 1% des cancers du sein peut être attribué à la prise de contraceptifs oraux (Centre International de Recherche sur le Cancer et al., 2007).

- Autres facteurs de risque (densité du sein, radiation, histoire personnelle,…)

De nombreux autres facteurs apparaissent comme des indicateurs émergents de risque de cancer du sein : la densité osseuse (Buist et al., 2001), le poids de naissance, la taille à l'âge adulte (Glade, 1999), la densité mammaire (McCormack and dos Santos Silva, 2006), ainsi que l'exposition à une irradiation thoracique (Travis et al., 2005).

Par ailleurs, le mode de vie peut également augmenter le risque de développer un cancer du sein : le surpoids et l'obésité à l'âge adulte (Reeves et al., 2007), la consommation d'alcool (Allen et al., 2009), ainsi que le tabagisme actif ou passif (Luo et al., 2011). A l'inverse, la pratique d'une activité physique régulière semble diminuer le risque de développer un cancer du sein (Friedenreich and Cust, 2008).

4.2.3 Facteurs de risque génétiques

La première étude rapportant un nombre anormalement élevé de cancers du sein dans une famille a été faite par Broca (Broca, 1866). L'augmentation du risque de développer un cancer du sein lorsque de nombreux proches sont atteints est d'autant plus importante dès lors que ces cancers ont été développés à un âge précoce. Aujourd'hui, on estime que 5 à 10 % des cancers du sein sont associés à une mutation constitutionnelle héritée, dont 4-5 % dûs à des gènes de forte pénétrance. La composante génétique est donc le facteur de risque le plus important.

Les facteurs de prédisposition au cancer du sein identifiés jusqu'alors peuvent être classés en trois catégories selon le risque qu'ils confèrent : des gènes de forte pénétrance, des gènes de pénétrance intermédiaire, et des allèles de faible pénétrance.

4.2.3.1 Facteurs de risque de forte pénétrance

A ce jour, trois gènes majeurs de forte pénétrance ont été identifiés : *BRCA1* et *BRCA2* par des analyses de liaison (Claus et al., 1991; Hall et al., 1990; Newman et al., 1988; Wooster et al., 1995) et par clonage positionnel en 1994 (Miki et al., 1994) et 1995 (Tavtigian et al., 1996; Wooster et al., 1994) respectivement, et *TP53* par une approche de gène candidat (Malkin et al., 1990; Srivastava et al., 1990).

Les protéines BRCA1 et BRCA2 ont chacune un rôle important dans la maintenance de l'intégrité et de la stabilité du génome, en intervenant dans la réparation des CDB. Environ une femme sur 500 à 1000 est porteuse d'une mutation sur *BRCA1*, ce qui représente ainsi de 7 à 10 % des cancers du sein familiaux. Les mutations *BRCA2* expliqueraient quant à elles 10 % des cancers familiaux. Une mutation dans l'un de ces deux gènes confère un risque relatif de 10 à 20 fois de développer un cancer du sein. La plupart des mutations germinales détectées sur ces deux grands gènes relève de mutations perte-de-fonction : mutations non-sens, mutations altérant le cadre de lecture, mutations affectant les sites d'épissage, ou

encore des réarrangements (insertion ou délétion). Ces mutations sont répertoriées par un effort collaboratif international dans la base de données du BIC.

La protéine p53 est un facteur de transcription ayant un rôle central dans de nombreuses voies cellulaires et dans le contrôle du cycle cellulaire, d'où son nom de « gardien de l'intégrité du génome », et elle est fréquemment mutée dans les cancers. Des mutations germinales de *TP53* ont été reliées à l'apparition du syndrome de Li-Fraumeni, un syndrome rare, associé dans 30 % des cas au développement d'un cancer du sein avant 30 ans (Malkin et al., 1990). Ainsi, une mutation sur *TP53* confère une augmentation de risque de développer un cancer du sein de 18 à 60 fois avant 45 ans. Cependant, les mutations sur *TP53* sont rarement retrouvées chez des familles de cancers du sein non associés au syndrome de Li-Fraumeni (Evans et al., 2002).

D'autres syndromes familiaux ont été associés à un risque élevé de cancer du sein, et de ce fait d'autres gènes : le syndrome de Cowden (mutation dans *PTEN*), le syndrome de Peutz-Jeghers (mutation dans *LKB1*) ou encore les cancers gastriques diffus héréditaires (mutation dans *CDH1*). La pénétrance des mutations de ces gènes est encore incertaine, du fait d'un manque d'études.

4.2.3.2 Facteurs de risque de pénétrance intermédiaire

A ce jour, quatre gènes de pénétrance intermédiaire ont été identifiés par une approche de gènes candidats : *CHEK2*, *ATM*, *BRIP1*, et *PALB2*. CHK2, BRIP1 (BRCA1-interacting protein C-terminal helicase 1, aussi connue sous le nom de BACH1 ou FANCJ) et ATM (ataxia-telangiectasia mutated) sont des protéines impliquées dans la voie de réparation des cassures double brin de l'ADN, impliquant également BRCA1 et p53. PALBP2 a été impliquée dans la localisation et la stabilité de BRCA2.

Ces gènes seraient impliqués dans 2 à 3 % des cancers du sein familiaux (Rahman et al., 2007). Une mutation dans l'un de ces quatre gènes confère un risque relatif de 2 à 4 fois. D'autres gènes de pénétrance intermédiaire ou faible ont été identifiés : *CASP8* (Cox et al., 2007), *TGFB1*, *BARD1*, *RAD50*.

4.2.3.3 Facteurs de risque de faible pénétrance

Actuellement, plus de 76 allèles communs (présents dans 10 à 50 % de la population générale) ou gènes de faible pénétrance ont été identifiés, la plupart grâce à des études pan-génomiques, chacun conférant un risque relatif inférieur à 1,5 (Bojesen et al., 2013; Easton et al., 2007a; Michailidou et al., 2013). Certains de ces allèles pourraient également augmenter le risque de développer un cancer du sein chez les porteuses de mutation *BRCA1/2* (Antoniou et al., 2008). Ainsi, le risque combiné de deux facteurs dépend de la nature de l'interaction entre eux.

4.2.3.4 D'autres facteurs de risque ?

Malgré les multiples stratégies utilisées pour découvrir des facteurs génétiques de prédisposition au cancer du sein, l'origine de la prédisposition génétique n'est pas identifiée chez près de 80 % des familles analysées en routine lors du dépistage moléculaire. Les analyses de liaison génétique et de séquençage de l'exome rejettent l'hypothèse d'un autre gène majeur de forte pénétrance comparable à *BRCA1* et *BRCA2*. La prédisposition pour certaines de ces familles pourrait être polygénique, c'est-à-dire dûe à une combinaison d'allèles rares et de faible pénétrance (COMPLEXO et al., 2013), et pourrait faire suite à des expositions environnementales variées (Antoniou et al., 2002; Pharoah et al., 2002). De plus, chez 20 % des familles pour lesquelles aucune mutation délétère n'est clairement identifiée, de nombreux variants de signification inconnue sont retrouvés sur les séquences des gènes *BRCA1* et *BRCA2* (Easton et al., 2007b). La classification de ces variants reste problématique, puisqu'il est difficile d'évaluer si ils altèrent suffisamment la fonction des protéines pour prédisposer au cancer.

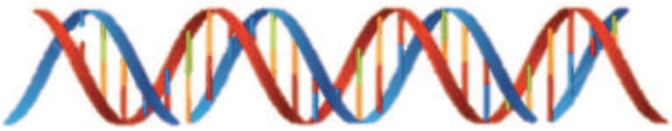
4.2.4 Les cancers du sein héréditaires et le dépistage moléculaire

Les cancers du sein héréditaires, également appelés syndrome du cancer héréditaire du sein et de l'ovaire (Hereditary Breast and Ovarian Cancer, ou HBOC) ne présentent pas de caractéristiques personnelles ou tumorales particulières. Le diagnostic repose sur les éléments suivants : un âge précoce au diagnostic, l'observation de multiples cancers primaires chez un même individu, un nombre important d'apparentés atteints dans la même branche (paternelle ou maternelle), l'observation de cancers bilatéraux, l'association à un cancer de l'ovaire, ou la présence d'un cancer du sein chez un apparenté masculin. Les cancers du sein héréditaires sont souvent associés à une apparition précoce des cancers du sein, et également à un risque élevé de cancer de l'ovaire, du pancréas, de l'estomac, du larynx, des trompes de Fallope et de la prostate.

Des outils comme BOADICEA ou BRCAPRO permettent d'analyser l'histoire familiale des individus atteints d'un cancer du sein pour lesquels une prédisposition génétique est suspectée (Lee et al., 2014; Varesco et al., 2013). Une recherche de mutations *BRCA1/2* dans les familles présentant une forte histoire familiale est alors proposée. Lorsque la mutation à l'origine de la prédisposition est identifiée, un test génétique prédictif est alors proposé aux autres membres de la famille présentant un fort risque et n'ayant pas encore développé la maladie. Un suivi médical personnalisé est alors mis en place pour les individus porteurs de la mutation. Pour les individus non porteurs de la mutation familiale, leur risque de développer un cancer du sein est équivalent à celui de la population générale. Cependant, la majorité des cancers du sein familiaux ne présentent pas de mutations sur *BRCA1/2* et l'origine génétique de la maladie n'est pas clairement identifiée. L'estimation du risque clinique de cancer du sein est alors déterminée empiriquement grâce aux données familiales.

Chapitre 2

Resultats



Mon équipe de recherche, dirigée par le Dr Sylvie Mazoyer, s'intéresse depuis de nombreuses années à la prédisposition génétique au cancer du sein. Comme précédemment énoncée dans l'introduction bibliographique, les gènes *BRCA1* et *BRCA2* sont les deux seuls gènes majeurs de prédisposition au cancer du sein. Malgré l'amélioration des techniques de criblage et la recherche de réarrangements génomiques autrefois négligés, aucune mutation clairement délétère n'est identifiée pour 80 % des familles analysées dans le cadre du diagnostic. Ceci suggère que les gènes *BRCA1/2* pourraient être à l'origine de la prédisposition de certaines familles *via* de nouveaux mécanismes conduisant à l'inactivation ou à la diminution constitutive de leur expression. L'objectif de ma thèse a été de travailler sur l'hypothèse que le CNV *RNU2*, localisé à proximité du gène *BRCA1*, pourrait être impliqué dans la prédisposition génétique au cancer du sein. De nombreuses études ont en effet montré que la présence d'un CNV peut modifier l'expression des gènes dans son voisinage, cet effet pouvant s'étendre jusqu'à 450 kb. Par ailleurs, d'autres répétitions en tandem de tailles variées ont déjà été reliées à l'apparition de maladies complexes : les répétitions *D4Z4* et le syndrome FSHD, les répétitions CAG et la maladie de Huntington, les répétitions CGG et le syndrome du X fragile, les répétitions GAA et l'ataxie de Freidrich.

Afin de tester cette hypothèse, nous avons dû préalablement caractériser plus finement ce locus. Dans un premier temps, nous avons précisé la localisation de ce CNV sur le chromosome 17, à une distance exacte de 124 kilobases en amont de *BRCA1* (Chapitre 2 – Section 1). Dans un second temps, nous avons étudié le niveau de polymorphisme de ce locus sur un grand nombre d'individus et déterminé le taux de mutation de ce CNV (Chapitre 2 – Section 2). Dans un troisième temps, nous avons conduit une étude cas/témoins et mis en évidence un effet potentiel du nombre de copies du CNV *RNU2* sur l'apparition d'un cancer du sein (Chapitre 2 – Section 3). Dans un quatrième temps, nous avons étudié l'effet de la variation du nombre de copies sur la méthylation de la région *BRCA1 – RNU2* (Chapitre 2 – Section 4). L'ensemble de ces résultats est présenté dans ce chapitre.

1 Localisation exacte du CNV *RNU2*

1.1 Introduction

Lorsque nous avons commencé ce travail en septembre 2009, le CNV *RNU2* n'avait plus fait l'objet de publications depuis 2000, et était absent de l'assemblage du génome humain de référence. Cependant, une étude par cartographie génétique avait localisé ce macrosatellite à environ 100 kb en amont de *BRCA1* (Liu and Barker, 1999), et cinq répétitions avaient été correctement assemblées et cartographiées au sein de la première version du génome de référence.

Afin de confirmer les études antérieures et de préciser la localisation de ce CNV au sein du génome humain, j'ai réalisé des expériences de FISH, en collaboration avec l'équipe de Damien Sanlaville (CBPE, Bron), et conduit une étude *in silico* des séquences de contigs BACs non assemblées contenues dans les bases de données. La répétition de base du CNV, préalablement séquencée et déposée sous le numéro d'accession U57614, a pu être retrouvée entièrement ou partiellement au sein de la séquence de nombreux contigs. J'ai étudié uniquement ceux contenant également des séquences uniques localisées à proximité de *BRCA1* (comme par exemple la séquence du gène *NBR1*). En analysant finement ces données et en conduisant des expériences de PCR, nous avons pu reconstituer un assemblage manuel de la région, et ainsi pu proposer la localisation du CNV entre les nucléotides 41,399,577 et 41,401,198 de la séquence du génome de référence.

Pour confirmer ce résultat, j'ai mis au point un *code-barres* (ensemble de sondes fluorescentes) spécifique du CNV *RNU2*, à partir du code-barres *BRCA1* initialement développé par l'équipe de D. Stoppa-Lyonnet (Gad et al., 2001a, 2001b, 2002a, 2002b), puis amélioré par la société Genomic Vision (Cheeseman et al., 2012). En collaboration avec cette société, j'ai complété ce code-barres avec une sonde spécifique de l'unité de base et avec deux sondes bordant le CNV. Grâce à la technique de peignage moléculaire, j'ai validé l'assemblage de la région, mais

également pu déterminer précisément le nombre allélique de copies chez une quarantaine d'individus. L'ensemble de ce travail est présenté dans l'Article 1.

1.2 Article 1 : Direct Visualization of the Highly Polymorphic *RNU2* locus in Proximity to the *BRCA1* Gene

OPEN ACCESS Freely available online

PLOS ONE

Direct Visualization of the Highly Polymorphic *RNU2* Locus in Proximity to the *BRCA1* Gene

Chloé Tessereau[1,2], Monique Buisson[1], Nastasia Monnet[1], Marine Imbert[1], Laure Barjhoux[1], Caroline Schluth-Bolard[3], Damien Sanlaville[3], Emmanuel Conseiller[2], Maurizio Ceppi[2], Olga M. Sinilnikova[1,4], Sylvie Mazoyer[1]*

1 «Genetics of Breast Cancer» team, Cancer Research Centre of Lyon, CNRS UMR5286, Inserm U1052, Université Claude Bernard Lyon 1, Centre Léon Bérard, Lyon, France, **2** Genomic Vision, Bagneux, Paris, France, **3** Service de Génétique, Laboratoire de Cytogénétique Constitutionnelle, Centre de Biologie et de Pathologie Est, Hospices Civils de Lyon and CNRS UMR5292, Inserm U1028, Université Claude Bernard Lyon 1, Equipe TIGER, Lyon, France, **4** Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon/Centre Léon Bérard, Lyon, France

Abstract

Although the breast cancer susceptibility gene *BRCA1* is one of the most extensively characterized genetic loci, much less is known about its upstream variable number tandem repeat element, the *RNU2* locus. *RNU2* encodes the U2 small nuclear RNA, an essential splicing element, but this locus is missing from the human genome assembly due to the inherent difficulty in the assembly of repetitive sequences. To fill the gap between *RNU2* and *BRCA1*, we have reconstructed the physical map of this region by re-examining genomic clone sequences of public databases, which allowed us to precisely localize the *RNU2* array 124 kb telomeric to *BRCA1*. We measured by performing FISH analyses on combed DNA for the first time the exact number of repeats carried by each of the two alleles in 41 individuals and found a range of 6-82 copies and a level of heterozygosity of 98%. The precise localisation of the *RNU2* locus in the genome reference assembly and the implementation of a new technical tool to study it will make the detailed exploration of this locus possible. This recently neglected macrosatellite could be valuable for evaluating the potential role of structural variations in disease due to its location next to a major cancer susceptibility gene.

Citation: Tessereau C, Buisson M, Monnet N, Imbert M, Barjhoux L, et al. (2013) Direct Visualization of the Highly Polymorphic *RNU2* Locus in Proximity to the *BRCA1* Gene. PLoS ONE 8(10): e76054. doi:10.1371/journal.pone.0076054

Editor: Brian P. Chadwick, Florida State University, United States of America

Received June 26, 2013; **Accepted** August 17, 2013; **Published** October 11, 2013

Funding: This work was supported by a grant from the "Fondation ARC pour la Recherche sur le Cancer" The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Chloé Tessereau, Emmanuel Conseiller and Maurizio Ceppi were employed by the commercial company Genomic Vision (Bagneux, Paris, France) at the time this study was conducted. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: sylvie.mazoyer@lyon.unicancer.fr

Introduction

Structural variation in the human genome has gained considerable attention in the recent years as it accounts for much of the variation between human genomes and may represent the main genetic basis of phenotypic differences. These variations may also provide an explanation for the missing heritability of complex diseases. Indeed large deletions, duplications, translocations and inversions have potentially great effects, including the changing of gene structure and dosage, altering gene regulation and exposing recessive alleles [1–2]. CNVs (Copy Number Variations), the most prevalent type of structural variation in the human genome, refer to DNA segments greater than 1 kb in size that are present at variable copy number [2–3]. The assessment of CNV phenotypic and pathologic potency has been made easier recently by the great improvement of CNV maps [4]. According to the high-resolution recent maps, most CNVs in the array-accessible regions of the genome are ancient bi-allelic polymorphisms that are in linkage disequilibrium (LD) with SNPs (Single Nucleotide Polymorphisms). This implies that the contribution of most common CNVs to human phenotypic variation was already detectable in genome-wide association studies (GWAS) as associations to nearby SNPs [5]. However, the question of the implication of multi-allelic CNVs in complex traits remains largely open as most of them cannot be genotyped by array technology, especially macrosatellites, the largest variable number tandem repeats (VNTR) [6]. Some, among which long-published and well documented structural variations, are not even present on the reference genome-assemblies, so their sequence is discarded when alternative genotyping technologies such as next generation sequencing are used [7]. One of such missing CNVs is the *RNU2* locus, which is all the more detrimental that this highly polymorphic macrosatelllite sits next to a major cancer predisposing gene, *BRCA1*.

The *RNU2* gene is transcribed by RNA polymerase II to give the U2 small nuclear RNA (snRNA), an essential component of the spliceosome. In 1984 it was found to lie within a 6.1 kb unit organised as a nearly perfect tandem array of 10 to 20 copies per haploid genome [8–9]. It was subsequently localised on chromosome band 17q21q22 [10] to an adenovirus 12-induced metaphase chromosome fragility site [11], in close proximity to the *BRCA1* gene according to FISH (Fluorescent *In Situ* Hybridization), radiation hybrid, physical and genetic maps [12–17]. The sequencing of the 6,132 bp unit (5,834 bp initially because an Alu sequence was missing in the original Genbank submission) failed to reveal any other coding sequence but showed

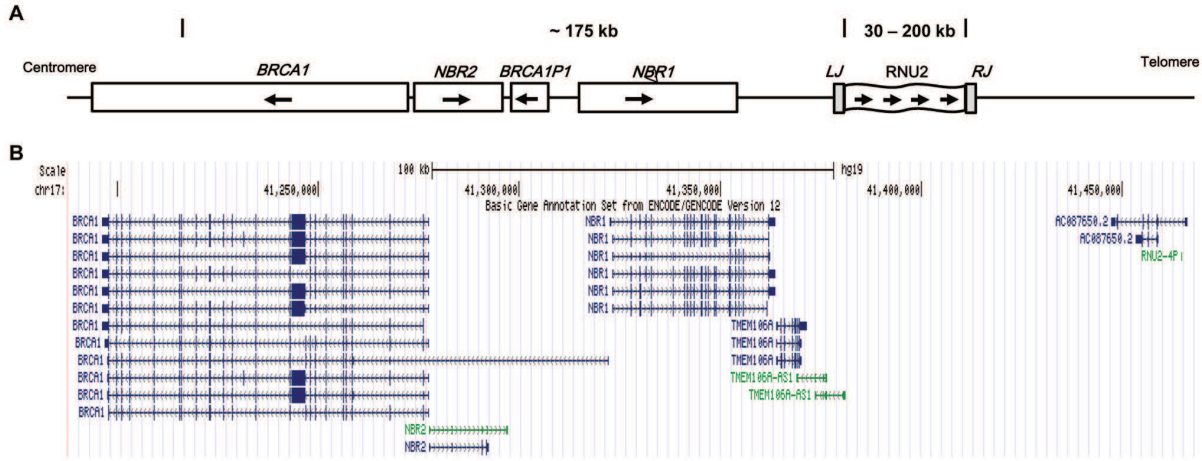


Figure 1. Schematic representation of the chromosome 17q21 region around the *BRCA1* gene. (A) Gene locations and physical map distances as reported in the literature [16,19]. (B) Gene locations within a 300 Kb window as shown in the UCSC Genome Browser. Arrows indicate transcription direction. BRCA1P1: *BRCA1* pseudogene.
doi:10.1371/journal.pone.0076054.g001

a high content of interspersed repeats (comprising 62.87% of the 6.1-kb unit), including notably five Alu and one LTR (Long Terminal Repeat) sequences, this latter suspected to be involved in the origin or maintenance of the *RNU2* array [18] (GenBank accession numbers L37793 and U57614.1). In this study, the regions flanking the *RNU2* locus were also cloned and sequenced, which subsequently allowed the establishment of the gene order on chromosome 17: *BRCA1* – left junction – *RNU2* locus – right junction – chromosome 17 telomere [19].

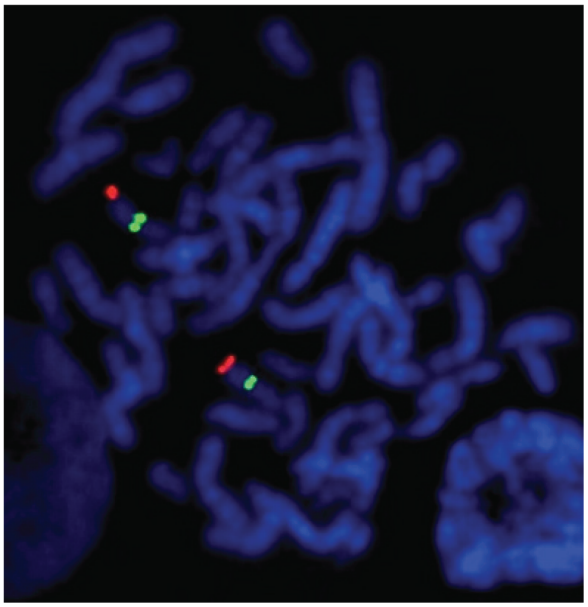


Figure 2. Visualization by FISH on mitotic metaphase chromosome of the *RNU2* locus. Two probes were used, one consisting of the 6.1 kb repeat unit labeled in green and a 17q subtelomeric probe labeled in red.
doi:10.1371/journal.pone.0076054.g002

Field Inversion Gel Electrophoresis (FIGE) analysis of >80 chromosomes from diverse human populations showed that the length of individual *RNU2* tandem arrays varied from ~40 to ~200 kb (~6 to >30 repeats): 57% of them were between 100 and 200 kb (16–30 repeats), 32% were between 40 and 100 kb (6–16 repeats) and 11% were longer than the 200 kb limit of the FIGE conditions used (>30 repeats) [20]. More recently, the study of 210 HapMap individuals with Pulse Field Gel Electrophoresis (PFGE) technique revealed a wider range of allelic size (6 to more than 60 copies) [21].

The first attempts to characterise the *RNU2* array made in the eighties and the nineties were halted before CNVs started to focus the attention of scientists, certainly because its absence from the human genome reference sequence made it disappear into the dustbin of obsolete and discredited sequences. Therefore, this macrosatellite did not benefit at all from the huge acceleration in human knowledge acquisition of the last 15 years resulting from the implementation of new technologies.

Here we present the precise localization of the *RNU2* locus within the chromosome 17 reference assembly and the first direct visualisation of this highly polymorphic CNV by FISH on combed DNA.

Materials and Methods

Ethics Statement

The studied subjects belonged to *BRCA1* families and either carried the *BRCA1* mutation present in the family or were non-carriers [22]. Informed consent was not required as the data were analyzed anonymously. Nevertheless, the subjects belong to a study which has been reviewed and approved by the appropriate ethics committee (Comité de Protection des Personnes Ile de France III, 3 october 2006, agreement n°2373).

Public Access Database Interrogation and Analysis of the Human Chromosome 17 Reference Sequence and of Clones' Sequence

The human genome reference sequence, working draft assemblies, clone sequences and annotations were obtained from the "UCSC Genome Bioinformatics Site" (http://genome.ucsc.edu). Gene-specific

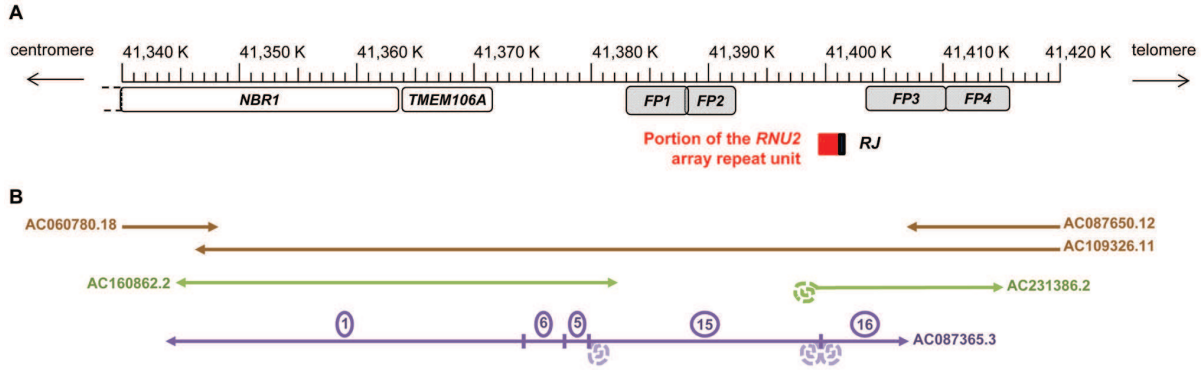


Figure 3. Localisation of the *RNU2* macrosatellite within the chromosome 17 sequence assemblies from NCBI Build 37.p10. (A) Schema of the region surrounding the *RNU2* array. The location of the portion of the *RNU2* repeat unit (not comprising the *RNU2* gene) and of the right junction found in the assemblies are depicted, as well as the probes used in molecular combing experiments that flank the *RNU2* array (FP1-4), and the *NBR1* and *TMEM106A* genes. (B) Clones covering the region. The reference sequence assemblies is based upon the complete sequence of 3 overlapping BACs, RP11-242D8, CTD-3014M21 and RP11-100E5 (AC060780.18, AC109326.11 and AC087650.12 respectively), represented by brown arrows. The complete sequence of the WI2-3095P13 fosmid (AC160862.2, green arrow) matches the reference sequence. The sequence of the ABC10-44487500M2 fosmid (AC231386.2, green arrow) matches the reference sequence up to its centromeric extremity where it contains several *RNU2* repeat units (depicted in a dotted curl). The five unassembled contigs of the working draft sequence of the RP11-570A16 BAC clone (AC087365.3) showing homology with the reference sequence are represented by a purple arrow. Contig 15 has been mis-assembled, as it contains several *RNU2* repeat units (depicted in dotted curls) at both its extremities.
doi:10.1371/journal.pone.0076054.g003

information was obtained from the "Entrez Gene" NCBI's database (http://www.ncbi.nlm.nih.gov/gene). Clone alignments were performed using the BLAST2Seq at the NCBI website (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&SHOW_DEFAULTS=on&BLAST_SPEC=blast2seq&LINK_LOC=align2seq).

Cell Lines

Human lymphoblastoid cell lines (LCLs) established by Epstein-Barr virus immortalization of subject's blood lymphocytes were maintained in RPMI 1640 medium (Life Technologies, Saint Aubin, France) supplemented with 10% fetal calf serum (VWR, Fontenay sous Bois, France) and 1% penicillin– streptomycin (Life Technologies).

Plug Preparation and Molecular Combing of DNA

EBV-immortalized lymphoblastoid cells were embedded in agarose blocks (1.2% NuSieve GTG Agarose, Lonza, Levallois-Perret, France) as previously described [23]. DNA was purified in an ESP solution: EDTA 0.5 M pH 8.0, 1% Sarcosyl (Sigma-Aldrich, Saint Quentin Fallavier, France), 2 mg/mL Proteinase K (Eurobio, Courtaboeuf, France) overnight and then agarose was melted at 68°C for 20 min and digested by 1.5 U of β-agarase (New England Biolabs, Evry, France) overnight in a M.E.S solution (2-N-Morpholino-Ethane sulfonique 500 mM pH 5.5). The resulting DNA solution was incubated with a silanized coverslip (CombiCoverslips, Genomic Vision, Paris, France), which was then removed from the solution at a constant speed of 300 µm/sec with the molecular combing system (MCS, Genomic Vision). This protocol allows maintenance of a constant DNA stretching factor of 2 kb/ µm [24]. Combicoverslips with combed DNA were then baked for 4 hours at 60°C. The quality of combing (linearity and density of DNA molecules) was estimated under an epi-fluorescence microscope equipped with an FITC filter set and a 40× air objective on freshly combed coverslips mounted in 20 µL of a 1 ml ProLong-gold solution containing 1 µL of Yoyo-1 solution (both from Life Technologies).

Metaphase Chromosome Spreading

Metaphase spreads were prepared from patient derived lymphocytes using standard procedures.

Probe Preparation

Probes were obtained by labelling PCR-amplified fragments using primers designed with the Primer3 v.0.4.0 software (http://frodo.wi.mit.edu/primer3/) and synthesized by Eurofins MWG Operon (Ebersberg, Germany). The entire *RNU2* repeat unit was amplified with primers ReRNU2$_{F/R}$ (5′-GCCAAAAGGACGAGAAGAGA-3′ (59°C)/5′-GGAGCTTGCTCTGTCCACTC-3′ (60°C)) for metaphase chromosome FISH experiments. For combed DNA FISH experiments, 2 regions of the repeat unit were chosen and amplified with primers L4$_{F/R}$ and L5$_{F/R}$ in order to include no more than 300 bp of repeat sequences (such as Alu or LTR sequences) according to the Repeat Masker software (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) and 4 regions flanking the *RNU2* array with primers FP1$_{F/R}$, FP2$_{F/R}$, FP3$_{F/R}$ and FP4$_{F/R}$. Long-range PCRs were performed in 20 µL reactions using Long PCR Enzyme Mix (Thermo Fisher Scientific, Illkirch, France), following these cycling conditions: 94°C for 2 min, 10 cycles of (96°C for 20 s, Tm°C for 30 s, 68°C for 45 s/kb), 25 cycles of (96°C for 20s, Tm°C for 30s, 68°C for 45 s/kb+10 s/cycle), 68°C for 10 min. Primer sequences and temperature of annealing (in brackets) were the following: L4$_F$ 5′-GCGGCCCACAAGATAAGATA-3′ (59°C); L4$_R$ 5′-ACGACGCAGTTAGGAGGCTA-3′ (59°C); L5$_F$ 5′-CTACACAGCCCAGGACACG-3′ (59°C); L5$_R$ 5′-GTTGGCCATGCCTTAAAGTG-3′ (59°C); FP1$_F$ 5′-CCAAATTTTCCAAGAGACTGACTT-3′ (59°C); FP1$_R$ 5′-GGAGTGAACAGGTGAGAGGATTAT-3′ (59°C); FP2$_F$ 5′-GAGCCAAAAATGGATACCTAGAGA-3′ (59°C); FP2$_R$ 5′-TGATCCCTGATATCCAATAACCTT-3′ (59°C); FP3$_F$ 5′-TACCCCCTTCCTAGCCCTTA-3′ (59°C); FP3$_R$ 5′-TCATGCAGCCTGGTACAGAG -3′ (58°C); FP4$_F$ 5′-ACCGGGCTGTGTAGAAATTG-3′ (58°C); FP4$_R$ 5′-ACCTCATCCTGGCTTACAGG-3′ (58°C). The sizes of the PCR fragments were 434 bp for L4, 1,959 bp for L5, 4,393 bp for FP1, 4,860 bp for FP2, 7,009 bp for FP3 and

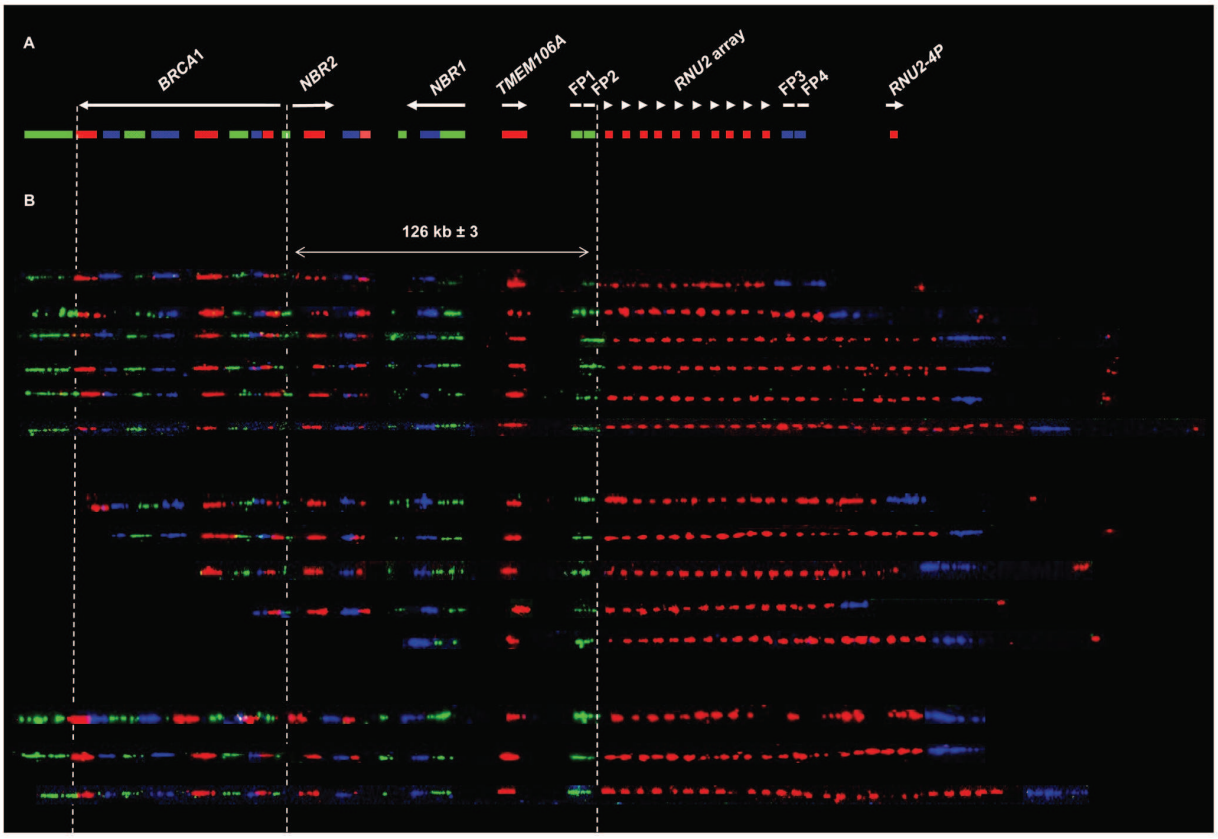


Figure 4. Visualization by molecular combing of the 17q21 region around *BRCA1*. (A) Schematization of the genomic morse code used. The *BRCA1* Genomic Morse Code (GMC) depicted (v4.0) is an improvement of the published code (v1.0) [23]. It covers a genomic region of 200 kb and consist in 17 signals of a distinct color (green, red or blue), each composed of 1 to 3 small horizontal bars corresponding to a single DNA probe. The signals for the flanking probes FP1-4 are each composed of 2 green or blue horizontal bars, while the signal for the *RNU2* array repeat unit is composed of 1 red horizontal bar. Of note, the probe for the *RNU2* array cross-reacts with *RNU2-4P*. (B) Fourteen fibres displaying different numbers of *RNU2* signals are shown. The first six fibres display the entire bar code from the *BRCA1* GMC to *RNU2-4P*, while the followings miss either the beginning of the *BRCA1* GMC or *RNU2-4P*.
doi:10.1371/journal.pone.0076054.g004

5,340 bp for FP4. PCR products have been cloned within the pCR2.1-TOPO XL vector (Life Technologies) according to the manufacturer's instructions.

The probe used in metaphase chromosome FISH experiments was labelled with fluorescein using the nick translation method. Probes used in combed DNA FISH experiments were labelled by random-priming: 200 ng of each probe were incubated during 10 min at 100°C with 1× random primers (Life Technologies), and then cooled at 4°C during 5 min. Klenow enzyme (40U) and dNTP 1× were added. Depending on the emission color chosen, dNTPs 1mM coupled with biotin (for red emission), digoxygenin (for blue emission), or Alexa-488 (for green emission) were also added. These mixes were incubated overnight at 37°C, and the priming reaction were then stopped with EDTA 2.10^{-2} mM pH 8.

Fluorescent In Situ Hybridization

On metaphase chromosomes. Hybridization was performed as described previously [25] with the probe described above and a 17q subtelomeric probe labelled with rhodamine (Cytocell, Cambridge, UK). After denaturation, overnight hybridization and post-hybridization washes, slides were DAPI counterstained and were read using a fluorescent microscope equipped with a CCD camera.

On combed DNA. One tenth of each random priming mix was precipitated during 1 hour at −80°C with 10 µg of Human Cot1 DNA, 2 µg herring sperm DNA, one tenth of volume of NaAc 3 M pH 5.2 and 2.5 volumes of Ethanol 100%. After centrifugation during 30 min at 4°C and at 13,500 rpm, the supernatant was discarded and the pellet dried at 37°C and dissolved with hybridization buffer (deionized formamide, SSC (salt sodium citrate) 2X, Sarcosyl 0.5%, NaCl 10 mM, SDS 0.5%, Blocking Aid). 20 µL of the mixes were laid on a coverslip with combed DNA, denatured at 95°C during 5 min, and incubation was then performed overnight at 37°C in a hybridizer (Dako, Les Ulis, France). For probe detection, hybridized coverslips were washed three times (3 min each) with formamide-SSC 2X, and three times with SSC 2X. Coverslips were then incubated 20 min at 37°C in a wet room with the first reagents: Streptavidine-A594 for Biotin-dNTP (1), Rabbit anti-A488 antibody for Alexa-A488-dNTP (2),

and Mouse anti-Dig AMCA antibody for Digoxygenin-dNTP (3). Coverslips were washed with three successive baths of SSC 2X-Tween20 1%. Similarly, coverslips were incubated with the second reagents: Goat anti-streptavidine biotinylated antibody (1), Goat anti-rabbit A488 antibody (2) and Rat anti-mouse AMCA antibody (3). Coverslips were washed and incubated with the third reagents: Streptavidine A594 (1), and goat anti-rat A350 antibody (3). Coverslips were dehydrated with three successive baths of ethanol (70-90-100%). Image acquisition was performed with a customized automated fluorescence microscope (Image Xpress Micro, Molecular Devices, Sunnyvale, CA, USA) at 40× magnification, and image analysis and signal measurement were performed with ImageJ (available from NIH) and GVlab (Genomic Vision) softwares. Allelic number of copies was determined by counting the number of signals corresponding to a repeat unit only on fibres for which intact flanking probes could be observed. In all cases, the number of copies has been determined by at least two individuals, resulting in differences of one copy at the most. For the nicest 72 fibres obtained from 21 individuals, we determined the individual exact stretching factor by measuring the length of a motif covering 128 kb within the *BRCA1* bar code, which in turn allowed us to determine the physical distance separating *BRCA1* and the *RNU2* locus.

Results

Precise Localisation of the *RNU2* Array

The organization of the *RNU2-BRCA1* region as published in the literature is presented in Figure 1A: the genes described within this interval are *NBR1*, *BRCA1P1* (a *BRCA1* pseudogene) and *NBR2* [16,19]. The distance between the *RNU2* locus and D17S1322, a microsatellite located within *BRCA1* intron 19, is reported to be ~175 kb based on physical maps. This would locate *BRCA1* ~113 kb away from the *RNU2* locus. In contradiction with the literature, a single *RNU2* gene described as a pseudogene, *RNU2-4P* (289 bp long), also known as *RNU2P2*, is found on the chromosome 17 reference assembly Build 37 in the first intron of an uncharacterised gene named *LOC100130581*, ~187 kb away from *BRCA1* (Figure 1B and Table S1 in Additional file). Along with *NBR1*, *BRCA1P1* and *NBR2*, one more gene, *TMEM106A*, has been identified by sequence analysis within this region.

As shown previously [11,26–27], FISH on mitotic metaphase chromosomes using a probe obtained by labelling a 6.1 kb PCR fragment amplified with primers flanking the *RNU2* repeat unit gave a unique signal over band 17q21 (Figure 2), which indicated that the repeat unit is located at the same cytogenetic band as the *BRCA1* gene. Furthermore, the high intensity of the signal was consistent with the repeat unit being present in multiple copies.

Seven *RNU2* genes could be found in Entrez Gene (NCBI's repository for gene-specific information), among which five are considered to be pseudogenes. *RNU2-1* (GenBank accession number NR_002716.3), assigned to chromosome band 17q12-q21, is identical to the gene found in the *RNU2* repeat unit, but this locus, which in Build 36 was annotated on an unplaced contig based on a single unfinished BAC (Bacterial Artificial Chromosome) sequence, is no longer present in Build 37 as the BAC was removed from the assembly. A portion of the *RNU2* repeat unit (corresponding to positions 1440-3036 of U57614.1) is nevertheless present at position 41,399,577-41,401,198 (Figure 3A). The right junction of the *RNU2* array sequenced in 1995 [18] (416 bp: 36 bp of the repeat unit+380 bp of flanking sequence) and located telomeric to the *RNU2* locus [19] could be found as well at position 41,401,163-41,401,579, while the left junction (92 bp: 47 bp of the repeat unit+45 bp of flanking sequence) is missing from the human genome assembly, probably due to sequence assembly errors. In light of these data, we hypothesised that the *RNU2* array was located between positions 41,399,577, and 41,401,198, where the right junction and part of the *RNU2* array repeat unit can be found.

Table 1. Description of the *RNU2* array alleles identified in 46 unrelated chromosomes.

Alleles (N = 28)	Number of repeat units	Number of occurrence	Frequency of each allele
1	6	1	0.02
2	8	1	0.02
3	9	1	0.02
4	11	2	0.04
5	12	1	0.02
6	13	1	0.02
7	14	2	0.04
8	15	1	0.02
9	16	1	0.02
10	17	1	0.02
11	18	3	0.07
12	19	5	0.11
13	20	1	0.02
14	21	2	0.04
15	22	2	0.04
16	23	1	0.02
17	25	1	0.02
18	27	2	0.04
19	28	2	0.04
20	29	2	0.04
21	30	1	0.02
22	32	2	0.04
23	34	2	0.04
24	35	2	0.04
25	36	1	0.02
26	37	2	0.04
27	47	2	0.04
28	82	1	0.02

doi:10.1371/journal.pone.0076054.t001

In order to sustain this hypothesis, we extracted from the databases the sequences covering this region and analyzed them. The complete sequence of a 41 kb fosmid (ABC10-44487500M2) reported in AC231386.2 confirmed the localisation of the *RNU2* macrosatellite, as it displayed 5 complete repeat units followed by sequences matching Build 37 from position 41,399,577 to 41,413,658 (Figure 3B). We also analysed the unfinished sequence of the RP11-570A16 BAC clone (AC087365.3), namely 16 unordered contigs covering 104,495 bp. Part or the entire sequence of the *RNU2* array repeat unit is found in all but contigs 1, 5 and 6. Contig 1, which contains *TMEM106A* and the end of *NBR1*, and contigs 6 and 5 match adjacent sequences on chromosome 17 (Figure 3). The main parts of contigs 15 and 16 also match adjacent sequences, with an overlap of 1.3 kb between contigs 15 and 16 corresponding to a portion of the *RNU2* array

repeat unit, and of ~500 bp between contigs 5 and 15. This ~500 bp overlap precedes a portion of the *RNU2* array repeat unit sequence in contig 15, while it is at the end of contig 5, which suggests that contig 15 has been incorrectly assembled and that all the sequences matching the *RNU2* array repeat unit should be placed at the other end of the contig. Indeed, the assembly of these contigs is comforted not only by the chromosome 17 reference sequence assemblies but also by the complete sequence of a fosmid (AC160862.2) that covers this region. In conclusion, these data are all in agreement with a localisation of the *RNU2* macrosatellite between positions chr17:41,399,577, and chr17:41,401,198, which puts it ~124 kb telomeric to the *BRCA1* gene and ~63 kb centromeric to the *RNU2-4P* gene.

Variation of the Number of *RNU2* Array Repeat Unit in the Human Population

We next undertook to directly visualize the proximity of the *RNU2* array with the *BRCA1* gene by using the molecular combing technology. We completed the existing bar code that allows to get a panoramic view of *BRCA1* and its flanking genes, namely *TMEM106A, NBR1*, *BRCA1P1*, and *NBR2* [23], with a probe obtained by labeling two PCR fragments amplified with primers flanking close regions devoid of repeat sequences within the *RNU2* array repeat unit (1.96 and 0.46 kb). We also generated four probes expected to hybridize regions flanking the *RNU2* macrosatellite based on our assumption of its location, respectively 7.3 kb downstream in the case of probes FP1 and FP2, on the centromeric side, and 2 kb upstream in the case of probes FP3 and FP4, on the telomeric side (Figure 4). Hybridisation of these probes with combed DNA of very good quality generated a consistent pattern of signals covering a genomic region >350 kb. This pattern shows the juxtaposition, from chromosome 17 centromere to telomere, of the *BRCA1* bar code, FP1-2 probes, *RNU2*, FP3-4 probes and *RNU2-4P* (the *RNU2* probes cross-react with the *RNU2* pseudogene), thus validating our tentative map (Figure 4). The average size of the interval between the end of the *BRCA1* gene and the *RNU2* array boundary was 126 kb±3 when measuring 72 fibres from 21 individuals (expected size based on the chromosome 17 reference assembly: 123.7 kb). The distance between the array boundary and *RNU2-4P* seemed consistent with that expected from our tentative map (63.4 kb), but the paucity of the number of fibres displaying both *BRCA1* and *RNU2-4P* precluded us from doing precise measures. Measurement of the *RNU2* signals gave an average size of 2.15 kb±0.63, while the average size for the gap between two *RNU2* signals was 4.30 kb±2.21, as expected on the basis of the sequence of the *RNU2* array repeat unit.

In total, we analysed 41 individuals with this technique. All but one of them displayed two populations of fibres containing different numbers of *RNU2* signals, confirming the high level of heterozygosity of the *RNU2* macrosatellite, which reached 0.98 in our small sample. Examples of fibres displaying different numbers of repeats are shown in Figure 4. The 28 different alleles that we identified among the 46 unrelated chromosomes analysed (five of which carrying a *BRCA1* mutation) are presented in Table 1: 14 of them (50%) were found only once while twelve were found twice (43%), one three times (3.5%) and one six times (3.5%). The number of *RNU2* array repeat units was found to range from 6 to 82 copies, and most of the alleles differed from their closest allele by one copy.

Discussion

The gaps in the finished human genome-assemblies are likely to host undiscovered CNVs. Some long-published and well documented structural variations are also missing from human genome-assemblies due to the difficulty to assemble repeated regions [28–29]. Indeed repeats confuse the assembly process, often resulting in contig mis-assembly [30]. The determination of which segments of the genome are affected by CNVs and the mapping of each CNV to a human genomic region is, however, an important step to assess the phenotypic and pathologic potency of these structural variations. To date, less than a dozen macrosatellites have been characterized although this type of CNVs consisting typically of dozens of repetitive units of several kilobases are among the most polymorphic structural variations and the most likely to impact chromatin organisation and human health [31].

Here, we have determined the exact localisation of the human *RNU2* macrosatellite within chromosome 17 genome-assembly (Build 37), between positions chr17:41,399,577, and chr17:41, 401,198, ~124 kb telomeric to the *BRCA1* gene and ~63 kb centromeric to one of the numerous *RNU2* pseudogenes, *RNU2-4P*, the only one present on chromosome 17. We validated this location by a FISH analysis of combed DNA ("molecular combing") using a *BRCA1* Genomic Morse Code [23] completed by probes complementary to the *RNU2* array repeat unit and to flanking regions. This approach allowed us to determine the exact number of repeats carried by 46 independent chromosomes (41 individuals analysed in total), revealing 28 different alleles that display from 6 to 82 monomers. Up to now, two studies on the *RNU2* macrosatellite alleles have been published in which the *RNU2* array sizes were estimated from FIGE- or PFGE-separated EcoRI (a null cutter) genomic fragments visualised with a *RNU2*-specific probe [20–21]. Interestingly, the minimal number of repeats is the same in the three studies (i.e. 6). Liao et al. (1999) identified 15 different alleles in 28 chromosomes, but FIGE could not resolve alleles with array length >200 kb (33 copies) [20]. PFGE resolution appeared better as Schaap et al. (2013) were able to identify 58 different alleles (6–63 copies) differing from their closest allele by one copy by analysing 210 human DNA samples from four populations [21]. However, the electrophoresis-based methods may lack precision in determining the exact number of repeats for large arrays, especially for those exceeding 500 kb, while repeat number counting following molecular combing is not sensitive to array length as long as probes complementary to the *RNU2* locus flanking regions are used to assess fibre integrity. Moreover, the molecular combing technique allows the identification of possible complex repeat patterns resulting from large insertions of foreign DNA into the array and/or repeat inversions. However, electrophoresis-based methods are better suited to detect mosaicism, which is quite common in the case of macrosatellites [31]. These two techniques are therefore complementary for the study of macrosatellite repeats.

U2 snRNAs play an essential role in formation of the catalytically active spliceosome by base pairing with both the intron branch point and the U6 snRNA [32]. A five nucleotide deletion in one of the five murine U2 snRNA genes causes ataxia and neurodegeneration, neuron loss being strongly dependent on the dosage of wild-type and mutant U2 snRNAs [33]. This finding suggests that *RNU2* might be associated with disease in humans as well. Growing evidence links splicing factor dysfunction with disease, particularly cancer [34]. Furthermore, the proximity of this macrosatellite to the *BRCA1* gene combined with its high degree of polymorphism raise the interesting possibility that it could be involved in breast cancer susceptibility. Indeed, investigations in mice have suggested that the effect of CNVs on the expression of flanking genes could extend up to 450 kb away from their location, all the more in the case of long CNVs (> 50kb) [35]. In humans, the stronger evidence of such an effect so far

came from the study of the Williams-Beuren syndrome, where not only hemizygous genes that map within the microdeletion responsible for the disease but also normal copy neighboring genes show decreased relative levels of expression [36]. Our study, which gives a precise localization and better characterizes the *RNU2* locus, provides the foundation for testing the association between copy number at this locus and breast cancer or other diseases risk.

Supporting Information

Table S1 Genomic coordinates of 17q21 genes and sequences (Build 37.p10).
(DOCX)

References

1. Eichler EE, Nickerson DA, Altshuler D, Bowcock AM, Brooks LD, et al. (2007) Completing the map of human genetic variation. Nature 447: 161–165.
2. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85–97.
3. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, et al. (2006) Copy number variation: new insights in genome diversity. Genome Res 16: 949–961.
4. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, et al. (2006) Global variation in copy number in the human genome. Nature 444: 444–454.
5. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166–1174.
6. Warburton PE, Hasson D, Guillem F, Lescale C, Jin X, et al. (2008) Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics 9: 533.
7. Abyzov A, Urban AE, Snyder M, Gerstein M (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21: 974–984.
8. Van Arsdell SW, Weiner AM (1984) Human genes for U2 small nuclear RNA are tandemly repeated. Mol Cell Biol 4: 492–499.
9. Westin G, Zabielski J, Hammarstrom K, Monstein HJ, Bark C, et al. (1984) Clustered genes for human U2 RNA. Proc Natl Acad Sci U S A 81: 3811–3815.
10. Hammarstrom K, Santesson B, Westin G, Pettersson U (1985) The gene cluster for human U2 RNA is located on chromosome 17q21. Exp Cell Res 159: 473–478.
11. Lindgren V, Ares M Jr, Weiner AM, Francke U (1985) Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. Nature 314: 115–116.
12. Abel KJ, Boehnke M, Prahalad M, Ho P, Flejter WL, et al. (1993) A radiation hybrid map of the BRCA1 region of chromosome 17q12-q21. Genomics 17: 632–641.
13. Albertsen HM, Smith SA, Mazoyer S, Fujimoto E, Stevens J, et al. (1994) A physical map and candidate genes in the BRCA1 region on chromosome 17q12-21. Nat Genet 7: 472–479.
14. Black DM, Nicolai H, Borrow J, Solomon E (1993) A somatic cell hybrid map of the long arm of human chromosome 17, containing the familial breast cancer locus (BRCA1). Am J Hum Genet 52: 702–710.
15. Flejter WL, Barcroft CL, Guo SW, Lynch ED, Boehnke M, et al. (1993) Multicolor FISH mapping with Alu-PCR-amplified YAC clone DNA determines the order of markers in the BRCA1 region on chromosome 17q12-q21. Genomics 17: 624–631.
16. Liu X, Barker DF (1999) Evidence for effective suppression of recombination in the chromosome 17q21 segment spanning RNU2-BRCA1. Am J Hum Genet 64: 1427–1439.
17. Neuhausen SL, Swensen J, Miki Y, Liu Q, Tavtigian S, et al. (1994) A P1-based physical map of the region from D17S776 to D17S78 containing the breast cancer susceptibility gene BRCA1. Hum Mol Genet 3: 1919–1926.
18. Pavelitz T, Rusche L, Matera AG, Scharf JM, Weiner AM (1995) Concerted evolution of the tandem array encoding primate U2 snRNA occurs in situ, without changing the cytological context of the RNU2 locus. EMBO J 14: 169–177.
19. Pavelitz T, Liao D, Weiner AM (1999) Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. EMBO J 18: 3783–3792.
20. Liao D, Pavelitz T, Kidd JR, Kidd KK, Weiner AM (1997) Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. EMBO J 16: 588–598.
21. Schaap M, Lemmers RJ, Maassen R, van der Vliet PJ, Hoogerheide LF, et al. (2013) Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. BMC Genomics 14: 143.
22. Sinilnikova OM, Mazoyer S, Bonnardel C, Lynch HT, Narod SA, et al. (2006) BRCA1 and BRCA2 mutations in breast and ovarian cancer syndrome: reflection on the Creighton University historical series of high risk families. Fam Cancer 5: 15–20.
23. Cheeseman K, Rouleau E, Vannier A, Thomas A, Briaux A, et al. (2012) A diagnostic genetic test for the physical mapping of germline rearrangements in the susceptibility breast cancer genes BRCA1 and BRCA2. Hum Mutat 33: 998–1009.
24. Michalet X, Ekong R, Fougerousse F, Rousseaux S, Schurra C, et al. (1997) Dynamic molecular combing: stretching the whole human genome for high-resolution studies. Science 277: 1518–1523.
25. Schluth-Bolard C, Delobel B, Sanlaville D, Boute O, Cuisset JM, et al. (2009) Cryptic genomic imbalances in de novo and inherited apparently balanced chromosomal rearrangements: array CGH study of 47 unrelated cases. Eur J Med Genet 52: 291–296.
26. Bailey AD, Li Z, Pavelitz T, Weiner AM (1995) Adenovirus type 12-induced fragility of the human RNU2 locus requires U2 small nuclear RNA transcriptional regulatory elements. Mol Cell Biol 15: 6246–6255.
27. Yu A, Fan HY, Liao D, Bailey AD, Weiner AM (2000) Activation of p53 or loss of the Cockayne syndrome group B repair protein causes metaphase fragility of human U1, U2, and 5S genes. Mol Cell 5: 801–810.
28. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.
29. Eichler EE, Clark RA, She X (2004) An assessment of the sequence gaps: unfinished business in a finished human genome. Nat Rev Genet 5: 345–354.
30. Phillippy AM, Schatz MC, Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. Genome Biol 9: R55.
31. Tremblay DC, Alexander G Jr, Moseley S, Chadwick BP (2010) Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. BMC Genomics 11: 632.
32. Wahl MC, Will CL, Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. Cell 136: 701–718.
33. Jia Y, Mu JC, Ackerman SL (2012) Mutation of a U2 snRNA gene causes global disruption of alternative splicing and neurodegeneration. Cell 148: 296–308.
34. Padgett RA (2012) New connections between splicing and human disease. Trends Genet 28: 147–154.
35. Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, et al. (2009) Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41: 424–429.
36. Merla G, Howald C, Henrichsen CN, Lyle R, Wyss C, et al. (2006) Submicroscopic deletion in patients with Williams-Beuren syndrome influences expression levels of the nonhemizygous flanking genes. Am J Hum Genet 79: 332–341.

Acknowledgments

We thank Richard Redon for helpful discussions, for reviewing the manuscript and offering helpful comments and suggestions, as well as Francesca Damiola for her critical reading of the manuscript. We thank Gaël Yvert, Rémy Bonnavion and Ivan Mikaélian for helpful discussions, Anne Vannier for expert advice on molecular combing and Jennifer Abscheidt, Mélanie Léoné and José Garcia for skilled technical assistance.

Author Contributions

Conceived and designed the experiments: CT SM. Performed the experiments: CT MB NM MI. Analyzed the data: CT MB NM MI SM. Contributed reagents/materials/analysis tools: LB CSB DS EC MC OMS. Wrote the paper: CT SM.

Table S1. Genomic coordinates of 17q21 genes and sequences (Build 37.p10).

Gene or sequence name	Genomic coordinates
BRCA1	Chr17:41,196,312 – 41,277,500
NBR2	Chr17: 41,277,600 – 41,292,342
NBR1	Chr17: 41,323,246 – 41,363,707
TMEM106A	Chr17: 41,363,894 – 41,371,589
FP1	Chr17:41,383,743 – 41,388,135
FP2	Chr17:41,387,378 – 41,392,237
FP3	Chr17:41,403,257 – 41,410,265
FP4	Chr17: 41,410,222 – 41,415,561
LOC10030581	Chr17: 41,447,213 – 41,466,266
RNU2-4P	Chr17:41,464,594 – 41,464,785
ARL4D	Chr17: 41,476,353 – 41,478,504

1.3 Discussion

A l'heure où les nouvelles technologies de séquençage foisonnantes et en constante innovation nous font miroiter le déchiffrage de notre propre génome en un clin d'œil, il est légitime de s'interroger sur l'intérêt et les espoirs que suscitent ces avancées. Séquencer tout ? Rapidement ? A bas prix ? Oui, mais pourquoi et comment ? Il est encore utile de le rappeler, le génome humain de référence est incomplet. Combien de personnes analysant des génomes à longueur de journées ont conscience qu'un gène aussi fondamental que le gène *RNU2* est absent du génome de référence ? Combien d'autres gènes tout aussi cruciaux sont-ils absents ? Quel est donc l'intérêt de séquencer tout s'il manque des informations qui pourraient s'avérer essentielles ?

Il apparaît évident que le critère qui devrait prendre le pas sur le coût et la rapidité est la qualité de l'assemblage du génome ainsi séquencé. Pourtant, juger de la qualité d'un assemblage n'est pas chose facile. Actuellement, différentes mesures le permettent, et sont principalement utilisées dans le cas d'un séquençage de nouvelles espèces. Un des indices le plus couramment utilisé est le N50. Le N50 est la taille du plus petit contig tel que 50 % de la longueur cumulée de l'ensemble des contigs obtenues après assemblage soit contenue dans des contigs de taille égale ou supérieure, que l'on pourrait également définir comme la taille médiane des contigs. Cependant, Parra *et al.* ont montré qu'un assemblage du génome de *Ciona intestinalis* présentant un N50 supérieur à l'assemblage antérieur ne contenait pas plusieurs gènes conservés, vraisemblablement parce que les algorithmes utilisés n'ont pas tenu compte des séquences répétées (Parra et al., 2009). Il est donc délicat de définir un (ou des) bon indicateur(s) de la qualité d'un génome, étant donné que les techniques de séquençage évoluent très rapidement.

Plutôt que de juger de la qualité de l'assemblage complet, il serait certainement plus informatif de passer au crible chaque région de l'assemblage. En dépit de la difficulté pour s'accorder sur la taille des intervalles à considérer, cela permettrait de mettre le doigt sur des régions peu ou mal assemblées. D'une certaine façon, cette

information peut être trouvée par le biais des trous annotés dans la séquence de référence. Ces trous sont petit à petit comblés, parfois correctement, mais comme l'illustre l'exemple du locus *RNU2* parfois de façon incorrecte, seulement pour augmenter la qualité apparente de l'assemblage. Ainsi qu'il est indiqué sur le site internet du Projet du Génome Humain (HGP), l'assemblage de ce gigantesque puzzle nécessite encore des vérifications humaines voire expérimentales avec des *vieilles* techniques, qui sont coûteuses en temps et en argent. Cela semble être un paradoxe à l'heure où l'on veut tout séquencer rapidement et à bas prix !

Pourtant, cet effort, loin d'être simplement ingrat, peut se révéler passionnant. Le locus *RNU2* avait été localisé en 17q21, par des techniques que certains jugent maintenant désuètes, en 1985, c'est-à-dire 16 ans avant la mise à disposition du premier génome de référence. On peut regretter ici que les algorithmes ainsi que les *relecteurs* de l'assemblage n'aient pas tenu compte des données de la bibliographie avant d'enlever ce locus des dernières versions.

Certes, les techniques de séquençage et d'assemblage ne nous permettent pas de déchiffrer avec certitudes les 3 milliards de nucléotides de notre ADN. Cependant, les données ainsi générées sont d'une aide précieuse pour conduire des approches plus expérimentales. Grâce aux séquences de contigs non assemblés, j'ai pu reconstruire *in silico* l'environnement nucléotidique de *BRCA1*. C'est grâce à son absence du génome de référence que nous avons été amenés à conduire une analyse plus fine de sa localisation en utilisant le peignage moléculaire. Grâce à cette technique, totalement fascinante car elle nous offre la possibilité de *voir* des gènes et des régions précises d'ADN, j'ai pu étudier d'autres facettes de ce locus ainsi qu'il sera exposé dans le chapitre suivant, ce que nous n'aurions vraisemblablement pas entrepris si nous avions connu précisément sa séquence et sa localisation par rapport à *BRCA1*. L'analyse des séquences de contigs non assemblés m'a permis de développer un code-barres spécifique de la région, le code-barres *RNU2*, qui s'est révélée être un atout majeur pour caractériser le polymorphisme de ce locus. En effet, l'étude par peignage moléculaire est actuellement la seule permettant de déterminer avec certitude le nombre allélique de copies. Notre étude, bien que portant sur un faible nombre d'individus (42) a mis en évidence des allèles comportant un nombre de copies plus

élevé que celui retrouvé dans des études antérieures : 82 copies, contre tout récemment 63 pour Schaap *et al.* sur 210 individus (Schaap et al., 2013), alors que le flou avait longtemps prévalu (> 30 sur 40 individus) (Liao et al., 1997). J'ai par la suite précisé ce niveau de polymorphisme en analysant un plus grand nombre d'individus, dont les résultats vous sont présentés dans la section suivante.

Les données de séquençage, combinées avec le peignage moléculaire, ont également apporté des informations nouvelles sur cette région, telles que la présence d'un pseudogène *RNU2-4P* à 60 kilobases en aval du locus *RNU2*. J'ai pu constater que ce pseudogène était séparé des autres répétitions par une insertion massive de LTRs et présente au sein d'une séquence de 17,8 kb (identifiée comme LOC100130581 lorsque j'ai débuté ma thèse) présentant plusieurs régions d'homologie avec les 6,1 kb de l'unité répétée du CNV. Il serait très intéressant d'étudier le mécanisme à l'origine de cette divergence, ainsi que l'âge de la divergence.

En conclusion, la connaissance précise de la localisation du CNV *RNU2*, et ainsi de son environnement proche, a représenté une étape nécessaire pour l'étude des effets d'un changement du nombre de copies. Cette première étape de mon projet de thèse m'a permis de confirmer que le macrosatellite *RNU2* se trouvait à une distance faible du promoteur du gène *BRCA1*, c'est-à-dire à une distance susceptible d'influencer son niveau d'expression (ou celui d'autres gènes localisés dans son voisinage), ce que nous avons testé par la suite.

2 Estimation du taux de mutation du CNV *RNU2* et de son niveau de polymorphisme dans la population

2.1 Introduction

Pour les nombreuses raisons exposées précédemment, le lien entre le phénotype et le génotype en ce qui concerne les CNVs multialléliques est plus difficile à mettre en évidence que dans le cas des CNVs bialléliques. Cet effet peut dépendre du nombre d'allèles présents dans la population générale. Pour caractériser finement un macrosatellite, il est donc primordial de définir le nombre d'allèles présents dans la population générale, mais également le nombre minimum, maximum et moyen de copies.

La technique de peignage moléculaire mise au point au début de ma thèse, bien que très précise, ne permet pas à l'heure actuelle de caractériser un grand nombre d'individus dans un laps de temps court. J'ai alors entrepris d'utiliser les données de séquençage du projet 1000 Génomes obtenues pour plus de 1000 individus, en collaboration avec l'équipe de Laurent Duret (LBBE, Villeurbanne), afin de préciser le niveau de polymorphisme du locus *RNU2*. En tirant profit du code-barres *RNU2*, nous avons également pu déterminer la fréquence de mutation de ce locus. Ces résultats vous sont présentés dans l'Article 2, accepté pour publication dans le journal Nucleic Acid Research et bientôt disponible.

2.2 Article 2 : Estimation of the *RNU2* macrosatellite mutation rate by *BRCA1* mutations tracing

Estimation of the *RNU2* macrosatellite mutation rate by *BRCA1* mutation tracing

Chloé Tessereau[1,2], Yann Lesecque[3], Nastasia Monnet[1], Monique Buisson[1], Laure Barjhoux[1], Mélanie Léoné[4], Bingjian Feng[5], David E. Goldgar[5], Olga M. Sinilnikova[1,4], Sylvain Mousset[3], Laurent Duret[3] , Sylvie Mazoyer[1,*]

[1]«Genetics of Breast Cancer» team, Cancer Research Centre of Lyon, CNRS UMR5286, Inserm U1052, Université Lyon 1, Centre Léon Bérard, Lyon, France, [2]Genomic Vision, Bagneux, Paris, France, [3]Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR5558, Université Lyon 1, France, [4]Unité Mixte de Génétique Constitutionnelle des Cancers Fréquents, Hospices Civils de Lyon / Centre Léon Bérard, Lyon, France, [5]Department of Dermatology and Huntsman Cancer Institute University of Utah School of Medicine, Salt Lake City, Utah, USA

*To whom correspondence should be addressed. Tel: +33 4 69 16 66 79; Fax: +33 4 78 78 27 20 ; E-mail: sylvie.mazoyer@lyon.unicancer.fr

KEYWORDS: Macrosatellite; *RNU2*; *BRCA1*; fiber FISH.

ABSTRACT

Large tandem repeat sequences have been poorly investigated as severe technical limitations and their frequent absence from the genome reference hinder their analysis. Extensive allelotyping of this class of variation has not been possible until now and their mutational dynamics are still poorly known. In order to estimate the mutation rate of a macrosatellite, we analysed in detail the *RNU2* locus, which displays at least 60 different alleles containing 6-82 copies of a 6.1 kb repeat unit. Mining data from the 1000 Genomes Project allowed us to precisely estimate copy numbers of the *RNU2* repeat unit using read depth of coverage. This further revealed significantly different mean values in various recent modern human populations, favouring a scenario of fast evolution of this locus. Its proximity to a disease gene with numerous founder mutations, *BRCA1*, within the same LD block, offered the unique opportunity to trace *RNU2* arrays over a large timescale. Analysis of the transmission of *RNU2* arrays associated with one "private" mutation in an extended kindred and four founder mutations in multiple kindreds gave an estimation by maximum likelihood of $5x10^{-3}$ mutations per generation, which is close to that of microsatellites.

INTRODUCTION

Tandem repeat DNA sequences, also called satellite DNA, represent a high proportion of the human genome. Due to their unique structural features, they are likely to contribute to genetic diversity and thereby to the variability of human traits, including diseases. The most extensively studied tandem repeat DNA sequences so far have been microsatellites (composed of unit from 1 to 10 bp), in particular those involved in trinucleotide repeat disorders, and minisatellites (unit from 10 to a few hundreds bp long). Comparatively, little is known about tandemly repeated sequences longer than 1 kb, also known as multi-allelic tandem Copy Number Variants (CNVs) or macrosatellites. Although macrosatellites constitute a sizeable fraction of large CNVs (1), can encompass large genomic intervals and are highly enriched with gene content, their impact on human genome plasticity has been poorly investigated despite the association of some arrays with disease susceptibility (2-4). This can be explained by the fact that they are very difficult to genotype directly using genome-wide platforms and are poorly tagged by single-nucleotide polymorphisms (SNPs) (5-7). Moreover, such regions are difficult to sequence and assemble and, as a result, they tend to be omitted from the human reference assembly (1), which excludes them from genome-wide investigations.

Nowadays, less than a dozen macrosatellites have been characterised although several attempts to identify new tandem repeats in the sequenced human genome have been made (1,8-9). The level of polymorphism of most of them has been evaluated on a low number of individuals using relatively low resolution techniques such as pulse field- or field inversion-gel electrophoresis (PFGE or FIGE), yet these were sufficient to reveal one of the highest degrees of allelic diversity in the human genome, attesting their rapid evolution. Likewise, the mutational rate of macrosatellites has been studied only in a limited number of meioses. A high meiotic mutation rate ($8.3x10^{-2}$ mutations per generation) was found for two macrosatellites, DXZ4 and RS447 (9-10), when analysing 24 and 60 parents-to-offspring

transmissions respectively, but was not observed anymore when evaluated on a larger number of meioses (120) (11).

The *RNU2* macrosatellite has been found to be hypervariable, with a 6.1 kb-long unit repeated from 5 to 82 times in a total of ~ 300 individuals of Caucasian, Asian and African origin (11-12). In a previous study that we conducted on 41 Caucasian individuals, the heterozygosity level of this microsatellite reached 98%. The *RNU2* locus has been localised by FISH at 17q21 close to the breast cancer predisposing gene *BRCA1* (13-14), precisely at a distance of 124 kb (12), within the same strong disequilibrium block. Despite this striking proximity, it has never been studied by whole-genome analysis perhaps because it is currently absent from the last human genome reference assembly. Notwithstanding the impact that the *RNU2* locus might have on *BRCA1*, the nearly total absence of recombination between these two loci can be used to follow *RNU2* alleles through many generations in *BRCA1* mutation carriers, and thus to estimate the timescale of evolutionary events.

We decided to explore more extensively the polymorphic state of macrosatellites by using, for the first time, the Depth of Coverage (DOC) of the *RNU2* repeat unit. To do so, we used data from the 1,000 Genomes Project. We validated this approach by obtaining a very good correlation between the numbers of *RNU2* repeats estimated by DOC and by fibre-FISH in 8 individuals. This method, that we recently developed in order to visualise the *RNU2* locus, is presently the most precise to directly count the allelic number of repeats (12). The wide variability of copy number we observed for 1,106 individuals confirmed the extremely high level of polymorphism of this locus. We found a statistically different mean copy number in various recent modern human populations, favouring a scenario of fast evolution of this locus. We then studied a large number of meiosis to determine more accurately the mutation rate of this macrosatellite by *BRCA1* mutation tracing rather than by investigating numerous trios. Contrasting with previous reports on macrosatellite extreme meiotic instability, we found that

RNU2 mutation rate in human modern population ($5x10^{-3}$ per generation) is close to that of microsatellites (estimated between 10^{-5} and 10^{-2} per locus per generation (15)).

MATERIALS AND METHODS

Cell lines

Human lymphoblastoid cell lines (LCLs) from eight CEPH individuals sequenced in the 1,000 Genomes Project, chosen on the basis of their *RNU2* copy number, were purchased from the Coriell Institute for Medical Research (Camden, NJ, USA). Their reference number (and estimated copy number) are the following: NA12272 (10), NA12275 (12), NA07048 (20), NA12006 (32), NA11840 (40), NA06989 (52), NA07051 (72), NA12718 (94). The other studied LCLs were from subjects that belonged to *BRCA1* families and either carried the *BRCA1* mutation present in the family or are non-carriers. All LCLs were maintained in RPMI 1640 medium (Life Technologies, Saint Aubin, France) supplemented with 10% fetal calf serum (VWR, Fontenay sous Bois, France) and 1% penicillin– streptomycin (Life Technologies).

Plug preparation, molecular combing of DNA and FISH

The procedure has been described in detail elsewhere (12). Briefly, EBV-immortalized lymphoblastoid cells were embedded in agarose blocks, DNA was purified, recovered from agarose and stretched on a silanized coverslip at constant speed (constant DNA stretching factor of 2 kb/µm). The quality of combing (linearity and density of DNA molecules) was estimated before performing hybridisation with a bar code consisting of probes hybridizing to regions flanking the *RNU2* array, extending on the centromeric side up to *NBR1*, as well as probes hybridizing to parts of the *RNU2* repeat unit. Following the probe detection step, image acquisition was performed with a customized automated fluorescence microscope. Allelic number of copies was determined by measuring fibre length and by counting the number of signals corresponding to a repeat unit only on fibres for which intact flanking

probes could be observed. Occasionally, a gap was observed in a random way within the *RNU2* array, most probably due to the absence of hybridisation of one *RNU2* probe.

BRCA1 mutation age estimation

In order to estimate the age of the mutation (or more precisely, the number of generations since the most recent common ancestor, MRCA) of the carriers of the two mutations analyzed in this study, we used the method that was first used to estimate the age of several *BRCA1* mutations (16). It was then extended and applied to *BRCA2* mutations (17) and used in several other similar studies, most recently in an analysis of the c.5266dupC *BRCA1* mutation (18). This method uses maximum likelihood and allows for both recombination and mutational events at the marker loci as means of altering a presumed ancestral haplotype. Phased haplotypes were used if these could be inferred from available family data; otherwise, all possible haplotypes were constructed from multi-locus genotype data and weighted according to their probability. For each value of G (the number of Generations since the MRCA), the relative likelihood that each haplotype is descended from the ancestral haplotype via mutation and recombination is calculated compared to the likelihood that it is a totally independent haplotype (i.e., an independent recurrent mutation on a different haplotype background). The value of G which maximizes this likelihood is obtained through iterative search. 95% support intervals were constructed by identifying those points GL and GU where the likelihood differed from the maximum by 0.86 (corresponding to a chi-squared likelihood ratio statistic of 3.84, e.g., $p = 0.05$).

From the set of 323 SNPS in the *BRCA1* +/- 2MB region we selected 17 SNPs on the basis of LD patterns and allele frequency to cover the region. These 17 SNPs spanned a region of 3.2Mb. To obtain genetic positions of each marker analyzed, we estimated the genetic

position from the proportion of physical distance between the known markers and then translated this to the genetic scale, assuming 1cM = 1Mb. As our method uses marker allele frequencies in the calculation of the likelihood, we estimated these frequencies from a large sample of >10000 *BRCA2* carriers genotyped for these SNPs as part of the iCOGS/CIMBA data.

Determination of the number of *RNU2* repeat unit by using the Depth of Coverage value

Depth of coverage (DOC) were calculated using mpileup (SAMtools) at position hs37d5:7,361,154-7,366,066 for the *RNU2* CNV repeat unit (avoiding interspersed sequences identified by RepeatMasker) and 17:41,400,862-41,401,838 for the right junction. The whole genome coverage has been estimated using the number of mapped bases according to data provided for each individual by the 1,000 Genomes Consortium. The *RNU2* copy number per genome has been estimated by dividing the *RNU2* CNV repeat unit DOC by the whole-genome DOC and by doubling this value, giving the DOC-estimated *RNU2* copy number (DCN).

Estimation of the *RNU2* mutation rate by maximum likelihood

As no information on the relationship between families was available, we assumed that those carrying the same *BRCA1* mutation radiated simultaneously from a single ancestor following a star phylogeny (Supplementary Figure 4).

The number of mutations along a lineage with t generations is denoted X_t and follows a Poisson distribution with parameter μt, where μ is the mutation rate per generations. Because only a few generations were observed with low mutation rates, we chose to neglect

reversions. We constructed the likelihood function $L(\mu)$ of the data as shown in formula (1): as we can count 639 generations without any mutation and independently 61, 73, 72 and 72 generations with at least one mutation occurring (Supplementary Figure 4).

$$L(\mu) = p(data|\mu) = p(X_{639} = 0) \times p(X_{61} \geq 1) \times p(X_{73} \geq 1) \times p(X_{72} \geq 1) \times p(X_{72} \geq 1) \quad (1)$$

Because $p(X_t \geq 1) = 1 - p(X_t = 0)$, the log-likelihood function $log(L(\mu))$ is:

$$\log\big(L(\mu)\big) = -639\mu + \log(1 - e^{-61\mu}) + \log(1 - e^{-73\mu}) + \log(1 - e^{-72\mu}) + \log(1 - e^{-72\mu}) \quad (2)$$

We maximized this log-likelihood using the *optimize()* function implemented in the R software with default parameters (19). We calculated the 95% confidence interval by inverting the acceptance region of the Wald test as described e.g. in (20). This method relies on the asymptotic normality of the estimator. Visual inspection of the likelihood function suggested that the original parameterization did not lend itself well to this assumption. We therefore reparameterized our model using the log function to construct the interval and converted the interval back using the exponential function.

Estimation of *RNU2* scaled mutation parameter $(\theta = 4N_e\mu)$

We used two estimators $\hat{\theta}_{\bar{x}}$ and $\hat{\theta}_{NA}$ of the scaled mutation parameter $\theta = 4N_e\mu$, where N_e is the effective population size and μ the mutation rate per generation developed previously for microsatellite loci (21). These estimators rely on a stepwise mutation model and were shown to accurately estimate θ from the total number of alleles n_A or the mean allele frequency $\overline{x}$ in a sample. We used the allele distribution data described in two independent studies (11-12),

leading to a total sample size of n = 504 haploid genotypes. These two studies identified n_A=53 different alleles with a mean allele frequency $\overline{x} = 0.018$. Estimator $\hat{\theta}_{\overline{x}}$ is directly related to :

$$\hat{\theta}_{\overline{x}} = \frac{1}{8\overline{x}^2} - \frac{1}{2}.$$

$\hat{\theta}_{NA}$ is obtained from n_A by solving for θ in the equation (7) of (21): $n_A = (c_0 + c_1 \ln(\theta) + c_2 \ln(\theta)^2)\sqrt{1+2\theta}$,

where c_0 to c_2 are coefficients estimated through simulations that depend on sample size. We used the coefficient values provided for a sample size of n = 500: c_0 = 2.0357, c_1 = -0.07910 and c_2 = -0.00007 (21). The equation was solved using the uniroot function of the R software (19).

RESULTS

Large diversity of *RNU2* copy number in 1,000 Genomes Data

To estimate the degree of variability of the *RNU2* repeat unit copy number in a large set of individuals, we analysed the sequence data generated by the 1,000 Genomes Project (1,000 Genomes Phase 2). We mapped sequence reads obtained in a set of 1,106 individuals using not only the reference genome assembly (Build 37/hg19) but also all known but unlocalized human genomic contigs (reference sequence set hs37d5) (22) that included at least one copy of the *RNU2* repeat unit. Sequence depth of coverage (DOC) of the *RNU2* repeat unit is systematically higher than whole genome DOC; by contrast, DOC of the macrosatellite right junction, which is not expected to be repeated, is not different than that of the whole genome (Figure 1A). By dividing the *RNU2* depth of coverage (DOC) by the whole genome DOC, we estimated the (diploid) *RNU2* copy number (DOC-estimated copy number, DCN). This DCN is on average 40.6, and is highly variable among individuals [2.5-160] (Figure 1B), which confirms that this locus is highly repetitive and polymorphic. To validate this macrosatellite genotyping approach, we first showed that the whole genome DOC has a negligible effect on the DOC-estimated copy number ($r^2=5x10^{-3}$, p-value=0.0126), suggesting that the *RNU2* macrosatellite polymorphic status hence estimated is only weakly affected by sequence coverage. It is known that the sequencing coverage is not uniform across the genome and that it notably decreases in regions of extreme GC-content (23-24). Given that the *RNU2* repeat unit is GC-rich (GC content: ~65%), the number of copy of this unit could be underestimated by the DOC method. To test this, we used a fibre-FISH approach that allows the number of *RNU2* repeat units to be precisely counted (12) in 8 individuals with various *RNU2* DCN (10-94). We found a strong agreement between these two techniques (Figure 2, $r^2= 0.93$, p-value < 0.0001), suggesting that the *RNU2* DCN can be used as a surrogate for the true underlying

RNU2 copy number variant although we observed a slight underestimation of small arrays and overestimation of large arrays by DOC.

As the goal of the 1,000 Genomes project is to to generate a comprehensive resource on human genetic variation in multiple human populations, we further explored variations in *RNU2* DOC in each of the five major population groups sequenced. Statistically significant differences in mean DOC-estimated *RNU2* copy numbers were seen between populations (Figure 3, Krustal-Wallis test p-value < 0.0001). Moreover, we observed a high diversity of distribution among super populations, such as for example in Europe between the Tuscans from Italy (TSI) and the British from England and Scotland (GBR). Strikingly, the populations that seem to be among the lowest in copy numbers are two Chinese populations, while the highest values come from another Chinese population. We also observed a high heterogeneity between individuals within populations, especially for the Peruvians from Lima (PEL) and the Tuscans from Italia (TSI). Such diversity reflects the rapidity with which these sequences evolve.

Transmission study of *RNU2* arrays in *BRCA1* families

As the mutation rate of macrosatellites is poorly documented, we next undertook to study the stability of the *RNU2* locus. This could be achieved by analysing many parent-to-offspring transmissions, a time-consuming task in the case of macrosatellites given that all the techniques currently available to genotype this category of polymorphisms (PFGE, FIGE or FISH on combed DNA) are low-throughput. Alternatively, one could follow alleles through a large time scale by genotyping distantly related individuals, the difficulty in this approach being the identification of such individuals. Here, we took advantage of the fact that the *RNU2* array and the *BRCA1* gene are located in the same haplotype block showing nearly

complete linkage disequilibrium (LD) (25), and of the identification of many *BRCA1* founder mutations indicating unsuspected familial relationships between apparently unrelated individuals. Indeed, the *BRCA1* LD block, first described in 1999 (25), has been delineated by the international HapMap project (August 2010 release) in Caucasians as a ~290 kb-long interval comprising 23 kb centromeric and 185 kb telomeric to the *BRCA1* gene based on Build 37 assembly and dbSNP b126, thus comprising the *RNU2* locus. Consequently, recombinations between *BRCA1* and *RNU2* are expected to be extremely rare, making it possible to use *BRCA1* mutations to trace *RNU2* alleles.

We first examined the stability of the *RNU2* array on a short timescale by measuring the number of repeat units segregating with a *BRCA1* mutation (c.4987-578_5074+342del1008, a 1-kb deletion comprising exon 17) in one of the largest *BRCA1* families reported in the literature, Family 1816. The pedigree records of this family extend over 6 generations, and include 56 *BRCA1* mutation carriers and 100 non carriers. We chose to investigate *RNU2* arrays in 8 family members carefully selected on the basis of them being either as closely or as distantly related as possible. As shown in Figure 4, allele segregation is strictly Mendelian in the nuclear family composed of two parents and three children (individuals 44, 45, 572, 575 and 579). The 5 mutation carriers analysed, separated by 1, 2, 5, 6 or 7 degrees, all share an allele carrying 28 repeat units, which indicates that the *RNU2* array displays meiotic stability on a short timescale and over a relatively small number of generations (at most 12 generations), at least for the array containing 28 repeat units.

We next measured the number of repeat units in the *RNU2* arrays segregating in individuals from different families but carrying the same *BRCA1* founder mutation. Two Ashkenazi Jewish founder mutations were studied: c.5266dupC (also known as 5382insC) and c.68_69delAG (also known as 185delAG), which have been estimated to have arisen respectively about 72 [49–107] and 61 [47-77] generations ago (18,26). We also analysed two

other founder mutations for which no age estimation was available in the literature: c.4186-1787_4357+4122dup (also known as ins6kbEx13), identified in families originating from many different countries (27-31), and c.213-11T>G (also known as 332-11T>G). We therefore first undertook to estimate the number of generations since the most recent common ancestor (MRCA) in 34 carriers of c.4186-1787_4357+4122dup and 86 carriers of c.213-11T>G by using carrier genotypes for SNPs located in the *BRCA1* region, available thanks to the iCOGS study (32). The maximum likelihood estimate of the time to the MRCA for the c.4186-1787_4357+4122dup haplotypes and c.213-11T>G haplotypes were 73 [52-100] generations and 87 [67-111] generations respectively.

The number of *RNU2* repeat units segregating with the *BRCA1* c.213-11T>G mutation, 22, is the same in all the carriers tested in two independent families (4 carriers in F2749 separated by 4 degrees at most; 2 carriers in F3103 separated by 5 degrees; Supplementary Figure 1). Conversely, different numbers of *RNU2* repeat units were shown to segregate with the three other *BRCA1* mutations. For some families, only one individual was available and could be genotyped. In these cases, the *RNU2* allele segregating with the *BRCA1* mutation was inferred, is possible. Concerning c.5266dupC for which carriers belonging to four independent families were tested, three different alleles were identified: 21 (in F1704), 19 (in F1973 and F3715b) and 35 or 13 (in F3574) *RNU2* repeat units (Figure 5). For c.68_69delAG, four independent families were studied and two different alleles were identified: 37 (in F2541, F3079 and F3261) or 47 (in F2979) repeat units (Supplementary Figure 2). For c.4186-1787_4357+4122dup, two different alleles carrying 29 and 14 repeats were identified in two independent families, F3173 and F3653 (Supplementary Figure 3). The results are summarized in Table 1.

To verify that recombinations between *BRCA1* and *RNU2* are indeed extremely rare and do not explain the few occurrences of variation in the number of *RNU2* repeats linked to some

BRCA1 mutations, we analysed SNPs genotypes for some of these individuals across the *BRCA1* LD block available (Supplementary Table 1), available thanks to the iCOGS study (32). Complete LD was systematically observed between flanking polymorphic markers, thus showing that cross-over between nonsister chromatids is extremely rare.

Maximum-likelihood estimation of the mutation rate

To evaluate the mutation rate of the *RNU2* locus, we used a simple Poisson model of the *RNU2* mutation process, with a single mutation rate parameter μ, common to every allele. We used maximum likelihood to estimate this rate. To do so we needed to calculate the number of meioses where *RNU2* alleles were transmitted with unchanged copy numbers and the number of meioses where *RNU2* alleles were transmitted with different copy numbers. To this aim, we used estimates of the maximum number of generations that separate *BRCA1* carriers from their common ancestor, assuming that families are related by a star phylogeny (Supplementary Figure 4). Doing so, we expect to slightly underestimate the mutation rate as those individuals are likely to share a more recent common ancestor than assumed. Secondly, we determined the minimum number of *RNU2* mutations associated with each *BRCA1* founder mutation. As shown in Figures 4-5 and Supplementary Figures 1-4, we observed at least 639 meioses without mutations (17 for F1816, 6 for F2749, 6 for F3103, approximately 174 generations between F2749 and F3103, 5 for F3173, 2 for F3653, 5 for F2979, 4 for F1973, and 20 for F1704, 183 between F3079/F2541/F3261 and their MRCA, approximately 174 generations between F2749/F3103 and their MRCA, 73 between F3173 or F3653 and their MRCA, 144 between F1973/F3715 and their MRCA). Conversely, we identified at least one mutational event between F3173 or F3653 and their MRCA (73 generations), one between F2979 and c.68_69delAG carriers' MRCA (61 generations), one between F3514 and

c.5266dupC carriers' MRCA (72 generations) and one between F1704 and c.5266dupC carriers' MRCA (72 generations). Using those data, we estimated the *RNU2* mutation rate to be $5x10^{-3}$ events per generation (95% confidence interval: [$1.7x10^{-3}$ – $1.6x10^{-2}$]). We calculated again the *RNU2* mutation rate by taking into account the confidence intervals of the estimation of the mutation age and found it to be comprised between $1.3x10^{-3}$ and $2.2x10^{-2}$.

We used a population genetics based approach to confront our estimates with independent estimates of the mutation parameter. We estimated the scaled mutation parameter $\theta=4N_e\mu$ at the *RNU2* locus in human populations from the total number of alleles n_A and the average frequency of alleles $\overline{x}$, and compared it to the value of our mutation rate estimate. Assuming roughly N_e=10,000 for humans (33) leads to an estimate of $4N_e\hat{\mu} = 200$. Because N_e in humans is hard to measure, θ could however be slightly higher (34). Using allelic distribution data taken from two independent studies (11-12) (total sample size n = 504), we obtained the following values: $\hat{\theta}_{\bar{x}}$ = 350.6 and $\hat{\theta}_{NA}$ = 602.2, which are slightly above but within the same range than our $4N_e\hat{\mu}$ estimate, 200. These estimators are based on a stepwise model of satellite mutation with each step leading to an expansion or a reduction of copy number by one unit. This model also assumes no directional bias *i.e.* increases in copy number are as frequent as decreases. As for microsatellite, it is likely that this model is quite far from what really happens in the mutational process of macrosatellites. Indeed, we suspect that *RNU2* mutations lead mostly to multistep expansions (or reductions) in copy number because, as far as we know, some intermediate copy numbers (*e.g.* from 63 to 82) have never been found in any individual analyzed yet, in all *RNU2* studies (11-12). This discrepancy between the model and the suspected mutation behavior at *RNU2* locus is likely to explain why $\hat{\theta}_{\bar{x}}$ and $\hat{\theta}_{NA}$ do not provide values closer to each other. However, authors predicted that $\hat{\theta}_{\bar{x}}$ and $\hat{\theta}_{NA}$ are quite

robust to the violation of the one-step assumption (21) making them the best choices currently available to simply confront our estimate of μ to previous results.

DISCUSSION

The study of macrosatellites is highly challenging. Not only it is very difficult to identify this type of multi-allelic CNV due to their absence from the genome reference assembly, but their genotyping is also problematic. Techniques that can be used presently are either time- and material-consuming and necessitate high skills (PFGE, FIGE or FISH on combed DNA), or are high-throughput but cannot resolve allelic copy number (qPCR). The determination of the spectrum and frequency of their allelic variations in a population is therefore difficult at the present time, especially as they are each likely to have hundreds of different alleles. For the time being, it seems that large scale variability is approachable through global copy number determination only.

While high-resolution sequence data generated by next-generation sequencing has been widely used for CNV detection, it had never been used, to our knowledge, to gain more insight into the variability of macrosatellites. We have shown here for the first time that Depth of Coverage (DOC) gives an accurate estimation of macrosatellite copy number. Indeed, we found a good agreement of the figures obtained by this approach with the numbers of *RNU2* repeat units measured by fibre-FISH for 8 individuals with various values of *RNU2* DOC. Having shown that global copy numbers of the *RNU2* repeat are highly variable in the 1,106 individuals sequenced in the 1,000 Genomes Project, we noticed statistically significant differences in mean *RNU2* copy numbers estimated by DOC between populations. We concluded from this observation that the *RNU2* macrosatellite evolves rapidly and decided to further investigate the mutation rate of this locus. The localisation of the *RNU2* locus within the *BRCA1* LD block gave us the unique opportunity to follow *RNU2* arrays through numerous meiosis spread over a large period of time, as very large *BRCA1* kindreds and several *BRCA1* founding mutations have been identified. We analysed the transmission of *RNU2* arrays associated with five different *BRCA1* mutations, one “private” mutation

identified in a single extended kindred and four founder mutations. Although their age estimate is more or less equivalent, their frequency is highly variable (Table 1). Indeed, c.68_69delAG and c.5266dupC, the two founder Ashkenazi Jewish mutations, were reported each in more than 2000 families while c.213-11T>G and c.4186-1787_4357+4122dup were reported in 56 and 45 families respectively in the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA) database (Lesley McGuffog, personal communication), which contains the world's largest collection of *BRCA1/2* carriers originating from 41 countries on six continents. Neither in the 13 parent-to-child transmissions that we analysed in total nor in *BRCA1* c.213-11T>G or c.4987-578_5074+342del1008 mutation carriers did we identify any copy number alteration of the *RNU2* array associated with the *BRCA1* mutation. However, c.68_69delAG and c.4186-1787_4357+4122dup were associated with two different numbers of repeats and c.5266dupC, the most frequent mutation, with three.

Using an innovative approach taking advantage of the location of the *RNU2* macrosatellite within the *BRCA1* disequilibrium block, we estimated the mutation rate of this locus in human modern population to be about $5x10^{-3}$ per generation, a mutation rate close to that of microsatellites (10^{-5} to 10^{-2} per locus per generation (15)). Our estimation is in agreement with the only published data concerning the *RNU2* locus stability which, although limited to the Mendelian transmission of the number of repeat units in one two-generation and one three-generation families, showed a mutation rate of $8x10^{-3}$ per generation (35-36). It also fits with the allelic diversity reported at this locus by our team and another previous study (11-12). On the other hand, this *RNU2* mutation rate contrasts with both meiotic and mitotic instability of the few macrosatellite sequences studied to date. Indeed, generational transmission studies of a few macrosatellite sequences revealed a high frequency of meiotic instability. For the SST1 macrosatellite that displays 134 different alleles carrying 14 to 154 copies of a 2.4-kb repeat unit, changes in copy number during parent to offspring transmission was evidenced in one of

three three-generation families (with either 6 or 7 children in the last generation) (9). For the TAF11-Like macrosatellite (54 different alleles carrying 10-98 copies of a 3.4-kb repeat unit), one of two three-generation families (with either 6 or 7 children in the last generation) showed copy-number alteration in meiotic transmission. In the case of RS447 for which more than 50 distinct alleles carrying 8-113 copies of a 4.7 kb repeat unit have been described, one study of 60 parent-to-offspring transmission reported an extremely high frequency (~8.3%) of meiotic instability (10) while another one carried on 60 trios found no intergenerational copy number alteration (11). In this latter study, although meiotic mutation rates were low, a high mitotic instability was found in all eight macrosatellites tested (11). In our study, macrosatellite stability was investigated using the molecular combing technology rather than pulse field gel electrophoresis as done previously, allowing determination of unit repeat numbers independently of restriction enzyme digestion that can be skewed by SNP through restriction enzyme site modification. However, we do not believe that this might explain why our results are different. Rather, we favour the hypothesis that the apparent higher stability of the *RNU2* macrosatellite might be due to its location within a strong linkage disequilibrium block, contrary to RS447, SST1 and TAF11-Like.

It is interesting to note that such a relatively stable macrosatellite sequence still has many different alleles and displays a high level of heterozygosity, suggesting that unequal crossing-over is not the principal mechanism generating new macrosatellite alleles. In the case of the *RNU2* locus, the molecular mechanism through which repeat copy number is altered is most likely unequal sister chromatid exchange as we were able to confirm through a haplotype analysis that there is no exchange between flanking markers. It has long been known that interchromosomal genetic exchanges are rare at the *RNU2* locus and that reciprocal nonsister chromatid exchange apparently does not occur, thanks to investigations on the concerted evolution of the *RNU2* repeats (35). Besides, it must be said that the mechanisms that lead to

changes in copy number in this specific class of structural variation have been poorly documented although progress has been made in the understanding of the mutability of microsatellites (37-38) and complex human genomic rearrangements (39). Hopefully, new technological and analytical developments will soon make it possible to study macrosatellites more easily. Indeed, detailed characterisation of macrosatellites and elucidation of their mutational dynamics are an important step towards a better understanding of the genetic instability of the genome and the potential association of structural variations with complex diseases and evolution.

FUNDING

This work was funded by a grant from the “Fondation ARC pour la Recherche sur le Cancer”.

ACKNOWLEDGMENTS

We thank the patients and genetic counselors who contributed to this work. We thank H.T. Lynch, C. Conway, J. Lynch, P. Watson, S. Slominski, C. Snyder, and L. Barjhoux for their contribution. We also thank L. McGuffog, D. Barrowdale, A.C. Antoniou and G. Chenevix-Trench for providing data from the CIMBA database and iCOGS genotypes for some CIMBA mutation carriers. We also thank A. Bensimon, S. Barradeau and E. Conseiller for helpful discussions, and J. Abscheidt and S. Bouchilloux for skilled technical assistance.

REFERENCES

1. Warburton, P.E., Hasson, D., Guillem, F., Lescale, C., Jin, X. and Abrusan, G. (2008) Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics*, **9**, 533.
2. Balog, J., Miller, D., Sanchez-Curtailles, E., Carbo-Marques, J., Block, G., Potman, M., de Knijff, P., Lemmers, R.J., Tapscott, S.J. and van der Maarel, S.M. (2012) Epigenetic regulation of the X-chromosomal macrosatellite repeat encoding for the cancer/testis gene CT47. *Eur J Hum Genet*, **20**, 185-191.
3. Bruce, H.A., Sachs, N., Rudnicki, D.D., Lin, S.G., Willour, V.L., Cowell, J.K., Conroy, J., McQuaid, D.E., Rossi, M., Gaile, D.P. *et al.* (2009) Long tandem repeats as a form of genomic copy number variation: structure and length polymorphism of a chromosome 5p repeat in control and schizophrenia populations. *Psychiatr Genet*, **19**, 64-71.
4. Lemmers, R.J., van der Vliet, P.J., Klooster, R., Sacconi, S., Camano, P., Dauwerse, J.G., Snider, L., Straasheijm, K.R., van Ommen, G.J., Padberg, G.W. *et al.* (2010) A unifying genetic model for facioscapulohumeral muscular dystrophy. *Science*, **329**, 1650-1653.
5. Alkan, C., Cardone, M.F., Catacchio, C.R., Antonacci, F., O'Brien, S.J., Ryder, O.A., Purgato, S., Zoli, M., Della Valle, G., Eichler, E.E. *et al.* (2011) Genome-wide characterization of centromeric satellites from multiple mammalian genomes. *Genome Res*, **21**, 137-145.
6. Campbell, C.D., Sampas, N., Tsalenko, A., Sudmant, P.H., Kidd, J.M., Malig, M., Vu, T.H., Vives, L., Tsang, P., Bruhn, L. *et al.* (2011) Population-genetic properties of differentiated human copy-number polymorphisms. *Am J Hum Genet*, **88**, 317-332.
7. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P. *et al.* (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704-712.
8. Li, W. and Olivier, M. (2013) Current analysis platforms and methods for detecting copy number variation. *Physiol Genomics*, **45**, 1-16.
9. Tremblay, D.C., Alexander, G., Jr., Moseley, S. and Chadwick, B.P. (2010) Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics*, **11**, 632.
10. Okada, T., Gondo, Y., Goto, J., Kanazawa, I., Hadano, S. and Ikeda, J.E. (2002) Unstable transmission of the RS447 human megasatellite tandem repetitive sequence that contains the USP17 deubiquitinating enzyme gene. *Hum Genet*, **110**, 302-313.
11. Schaap, M., Lemmers, R.J., Maassen, R., van der Vliet, P.J., Hoogerheide, L.F., van Dijk, H.K., Basturk, N., de Knijff, P. and van der Maarel, S.M. (2013) Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics*, **14**, 143.
12. Tessereau, C., Buisson, M., Monnet, N., Imbert, M., Barjhoux, L., Schluth-Bolard, C., Sanlaville, D., Conseiller, E., Ceppi, M., Sinilnikova, O.M. *et al.* (2013) Direct visualization of the highly polymorphic RNU2 locus in proximity to the BRCA1 gene. *PLoS One*, **8**, e76054.
13. Hammarstrom, K., Santesson, B., Westin, G. and Pettersson, U. (1985) The gene cluster for human U2 RNA is located on chromosome 17q21. *Exp Cell Res*, **159**, 473-478.
14. Black, D.M., Nicolai, H., Borrow, J. and Solomon, E. (1993) A somatic cell hybrid map of the long arm of human chromosome 17, containing the familial breast cancer locus (BRCA1). *Am J Hum Genet*, **52**, 702-710.
15. Ellegren, H., Primmer, C.R. and Sheldon, B.C. (1995) Microsatellite 'evolution': directionality or bias? *Nat Genet*, **11**, 360-362.
16. Neuhausen, S.L., Mazoyer, S., Friedman, L., Stratton, M., Offit, K., Caligo, A., Tomlinson, G., Cannon-Albright, L., Bishop, T., Kelsell, D. *et al.* (1996) Haplotype and phenotype analysis of six recurrent BRCA1 mutations in 61 families: results of an international study. *Am J Hum Genet*, **58**, 271-280.

17. Neuhausen, S.L., Godwin, A.K., Gershoni-Baruch, R., Schubert, E., Garber, J., Stoppa-Lyonnet, D., Olah, E., Csokay, B., Serova, O., Lalloo, F. *et al.* (1998) Haplotype and phenotype analysis of nine recurrent BRCA2 mutations in 111 families: results of an international study. *Am J Hum Genet*, **62**, 1381-1388.
18. Hamel, N., Feng, B.J., Foretova, L., Stoppa-Lyonnet, D., Narod, S.A., Imyanitov, E., Sinilnikova, O., Tihomirova, L., Lubinski, J., Gronwald, J. *et al.* (2011) On the origin and diffusion of BRCA1 c.5266dupC (5382insC) in European populations. *Eur J Hum Genet*, **19**, 300-306.
19. R Development Core Team. (2012) R: A language and environment for statistical computing. *Vienna, Austria : R Foundation for Statistical Computing* **1**, 1–1833.
20. Shao, J. (2003) Mathematical Statistics 2nd ed. *Springer* **§ 7.3.2 (XVI)**, 1-591.
21. Haasl, R.J. and Payseur, B.A. (2010) The number of alleles at a microsatellite defines the allele frequency spectrum and facilitates fast accurate estimation of theta. *Mol Biol Evol*, **27**, 2702-2715.
22. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.
23. Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*, **21**, 974-984.
24. Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, **19**, 1586-1592.
25. Liu, X. and Barker, D.F. (1999) Evidence for effective suppression of recombination in the chromosome 17q21 segment spanning RNU2-BRCA1. *Am J Hum Genet*, **64**, 1427-1439.
26. Laitman, Y., Feng, B.J., Zamir, I.M., Weitzel, J.N., Duncan, P., Port, D., Thirthagiri, E., Teo, S.H., Evans, G., Latif, A. *et al.* (2013) Haplotype analysis of the 185delAG BRCA1 mutation in ethnically diverse populations. *Eur J Hum Genet*, **21**, 212-216.
27. The BRCA1 Exon 13 Duplication Study Group. (2000) The exon 13 duplication in the BRCA1 gene is a founder mutation present in geographically diverse populations. . *Am J Hum Genet*, **67**, 207-212.
28. Cerutti, R., Sahnane, N., Carnevali, I., Furlan, D., Tibiletti, M.G., Chiaravalli, A.M. and Capella, C. (2010) Identification of the first case of germline duplication of BRCA1 exon 13 in an Italian family. *Fam Cancer*, **9**, 275-282.
29. Hendrickson, B.C., Judkins, T., Ward, B.D., Eliason, K., Deffenbaugh, A.E., Burbidge, L.A., Pyne, K., Leclair, B., Ward, B.E. and Scholl, T. (2005) Prevalence of five previously reported and recurrent BRCA1 genetic rearrangement mutations in 20,000 patients from hereditary breast/ovarian cancer families. *Genes Chromosomes Cancer*, **43**, 309-313.
30. Hofmann, W., Gorgens, H., John, A., Horn, D., Huttner, C., Arnold, N., Scherneck, S. and Schackert, H.K. (2003) Screening for large rearrangements of the BRCA1 gene in German breast or ovarian cancer families using semi-quantitative multiplex PCR method. *Hum Mutat*, **22**, 103-104.
31. Kremeyer, B., Soller, M., Lagerstedt, K., Maguire, P., Mazoyer, S., Nordling, M., Wahlstrom, J. and Lindblom, A. (2005) The BRCA1 exon 13 duplication in the Swedish population. *Fam Cancer*, **4**, 191-194.
32. Couch, F.J., Wang, X., McGuffog, L., Lee, A., Olswold, C., Kuchenbaecker, K.B., Soucy, P., Fredericksen, Z., Barrowdale, D., Dennis, J. *et al.* (2013) Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet*, **9**, e1003212.
33. Takahata, N. (1993) Allelic genealogy and human evolution. *Mol Biol Evol*, **10**, 2-22.
34. Wall, J.D. (2003) Estimating ancestral population sizes and divergence times. *Genetics*, **163**, 395-404.
35. Liao, D., Pavelitz, T., Kidd, J.R., Kidd, K.K. and Weiner, A.M. (1997) Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid

intrachromosomal homogenization and rare interchromosomal gene conversion. *EMBO J*, **16**, 588-598.
36. Liao, D. and Weiner, A.M. (1995) Concerted evolution of the tandemly repeated genes encoding primate U2 small nuclear RNA (the RNU2 locus) does not prevent rapid diversification of the (CT)n.(GA)n microsatellite embedded within the U2 repeat unit. *Genomics*, **30**, 583-593.
37. Ananda, G., Walsh, E., Jacob, K.D., Krasilnikova, M., Eckert, K.A., Chiaromonte, F. and Makova, K.D. (2013) Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. *Genome Biol Evol*, **5**, 606-620.
38. Leclercq, S., Rivals, E. and Jarne, P. (2010) DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biol Evol*, **2**, 325-335.
39. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. and Ira, G. (2009) Mechanisms of change in gene copy number. *Nat Rev Genet*, **10**, 551-564.

FIGURE LEGENDS

Figure 1: Estimation of *RNU2* copy numbers in 1,106 individuals sequenced in the 1,000 Genomes Project using the Depth of coverage (DOC) value. (A) Sequence depth of coverage for the *RNU2* repeat unit or the *RNU2* flanking region *versus* sequence depth of coverage for the whole genome. (B) Distribution of DOC-estimated *RNU2* copy numbers (DCN).

Figure 2: Correlation between DOC-estimated *RNU2* copy number (DCN) and *RNU2* copy number measured by molecular combing for 8 individuals from the 1,000 Genomes Project.

Figure 3: Mean *RNU2* copy numbers estimated by DOC in the different populations of the 1,000 Genomes Project data. Black bars: median, Whiskers: interquartile. ACB: African Caribbean in Barbados (N=55); ASW: African Ancestry in Southwest US (N=50); MXL: Mexican Ancestry in Los Angeles California (N=58); PEL: Peruvian in Lima Peru (N=50); PUR: Puerto Rican in Puerto Rico (N=63); CLM: Colombian in Medellin Colombia (N=55); CEU: Northern Europeans from Utah (N=42); TSI: Toscani in Italy (N=11); GBR: British from England and Scotland (N=59); FIN: Finnish in Finland (N=65); IBS: Iberian populations in Spain (N=76); YRI: Yoruba in Ibadan Nigeria (N=46); LWK: Luhya in Webuye Kenya (N=80); GIH: Gujarati Indian in Houston (N=75); CHB: Han Chinese in Bejing China (N=25); JPT: Japanese in Tokyo Japan (N=65); CHS: Han Chinese South (N=69); CDX: Chinese Dai in Xishuangbann China (N=82); KHV: Kinh in Ho Chi Minh City Vietnam (N=78). As for the 1.000 Genomes Project, these populations have been divided into 5 super populations : African (AFR), Ad Mixed American (AMR), East Asian (ASN), European (EUR) and South Asian (SAN).

Figure 4. Inheritance of the *RNU2* array co-segregating with the c.4987-578_5074+342del1008 *BRCA1* mutation in a large *BRCA1* family. (A) Pedigree of Family 1816. The number of repeat units in *BRCA1* mutation carriers ($^{+}$) or in individuals not carrying the *BRCA1* mutation ($^{-}$) was determined by molecular combing and is shown in red, while the identification number of each individual is shown in black. The number of repeats shared by all the mutation carriers is bolded (28). Only the family members that have been analyzed in the present study are indicated for clarity. (B) Visualization by molecular combing and fibre-FISH of the 17q21 region around the *RNU2* macrosatellite. Probes hybridizing regions flanking the *RNU2* macrosatellite were labeled in green and/or blue, while a probe hybridizing a region within the *RNU2* array repeat unit was labeled in red. For each analyzed individual, two fibers that display the bar code for the *RNU2* macrosatellite and flanking regions are shown, corresponding to the two alleles.

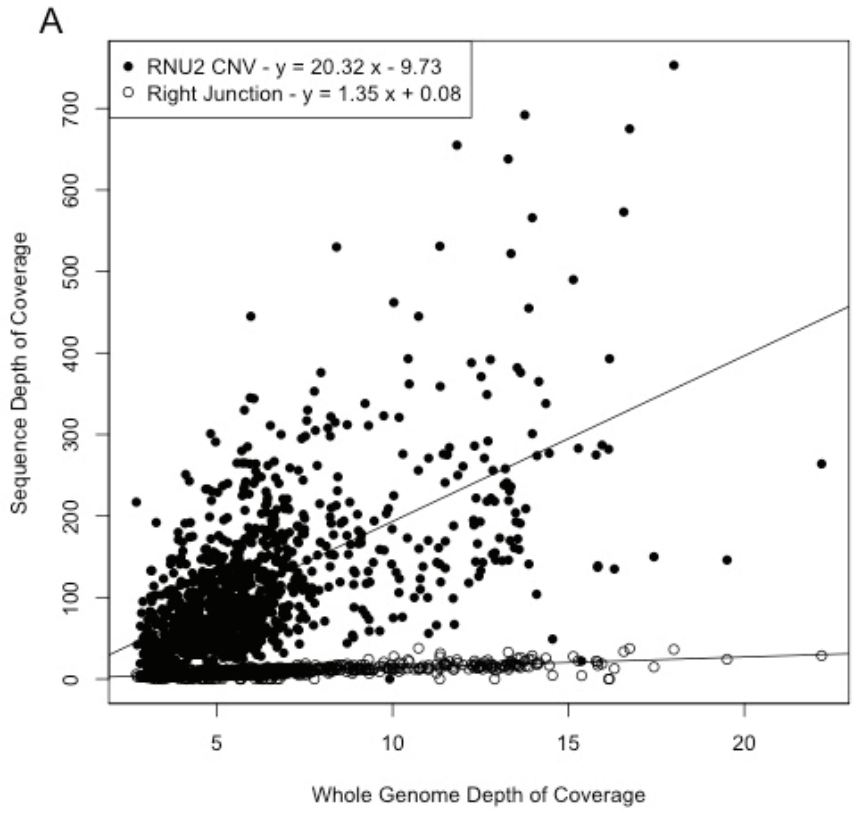
Figure 5. Inheritance of the *RNU2* array co-segregating with the c.5266dupC *BRCA1* mutation in the 1704, 1973, 3715b, and 3574 families. (A) Pedigrees of the families for which more than one individual have been genotyped (F1704 and F1973), (B) Visualization by molecular combing and fibre-FISH of the 17q21 region around the *RNU2* macrosatellite.

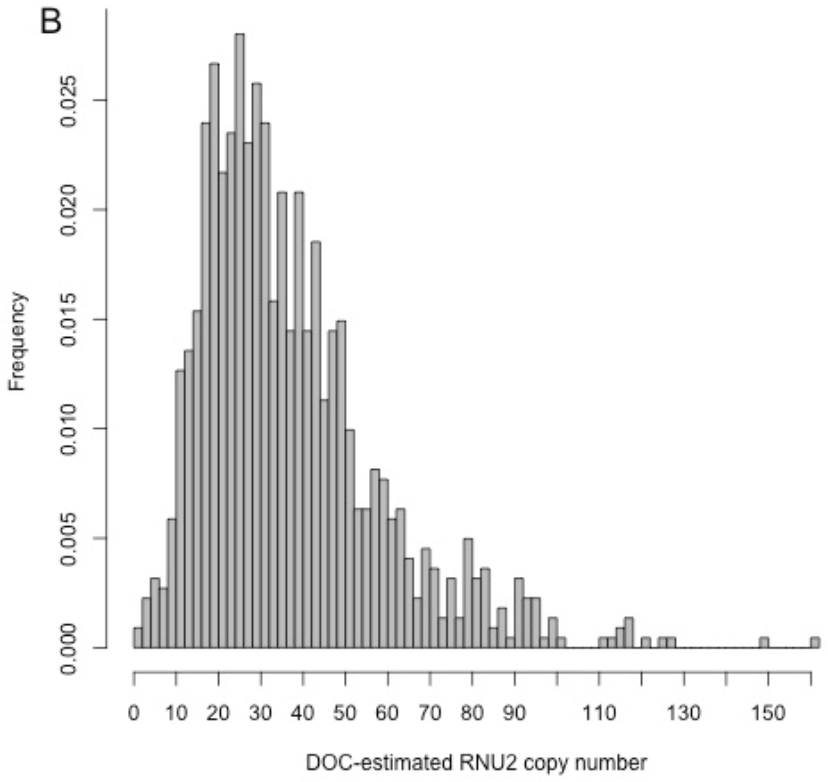
The rest of the legend is as in Figure 4.

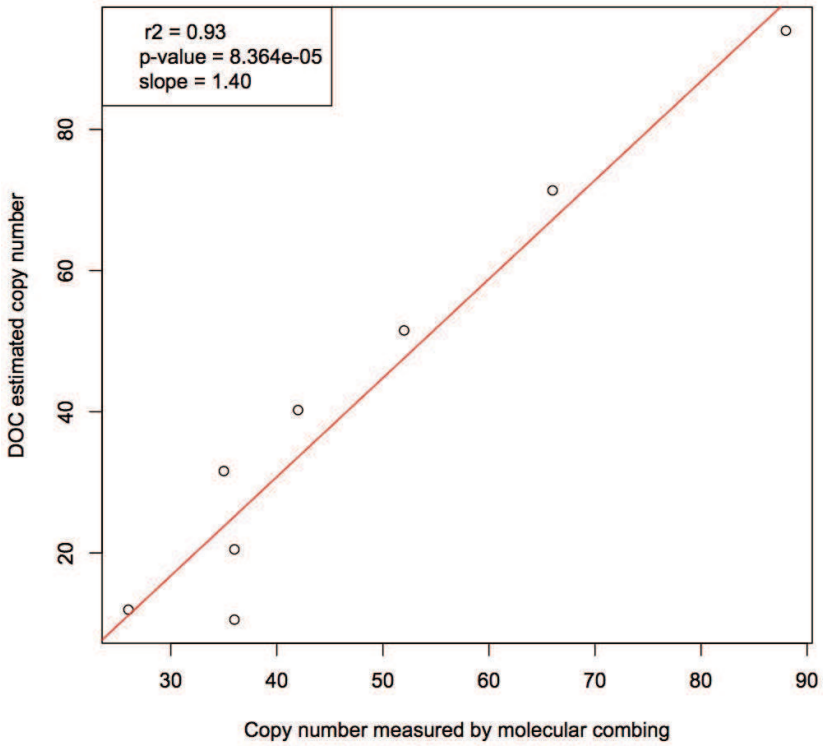
Table 1: Overview of the results of copy numbers of the *RNU2* macrosatellite associated with various *BRCA1* mutations

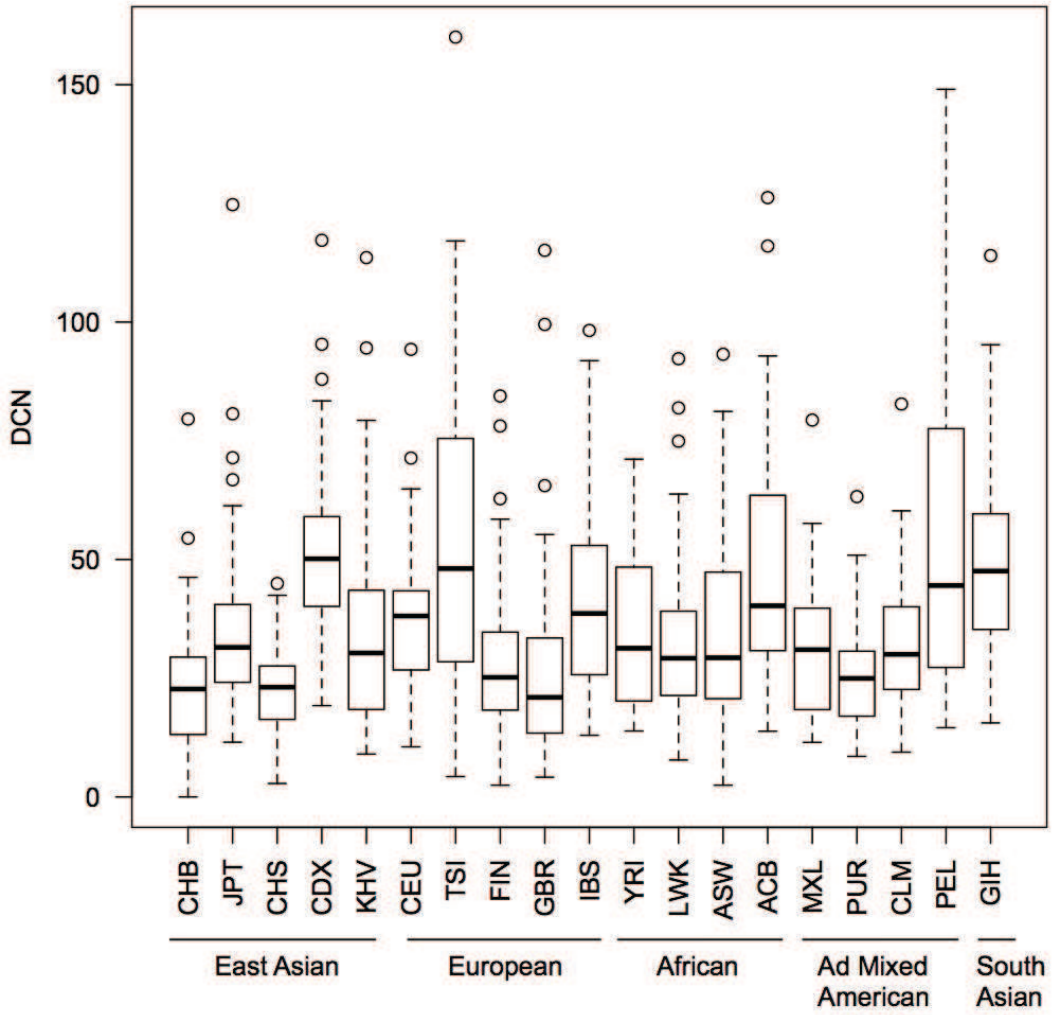
BRCA1 mutation	c.68_69 delAG	c.213-11T>G	c.4186-1787_4357 +4122dup	c.4987-578_5074 +342del1008	c.5266dupC
Age estimate (generations) [95% CI]	61 [47-77][a]	87 [67-111][b]	73 [52-100][b]	-	72 [49–107][c]
Number of independent families in CIMBA database	2097	56	45	1	2703
Number of countries in which the mutation has been identified in CIMBA database	25	6	6	1	24
Number of families analyzed in the present study	4	2	2	1	4
Number of *RNU2* repeats associated with *BRCA1* mutation	37;47	22	14;29	28	13 or 35;19;21

a: estimated in (21); b: estimated in this study; c: estimated in (20).

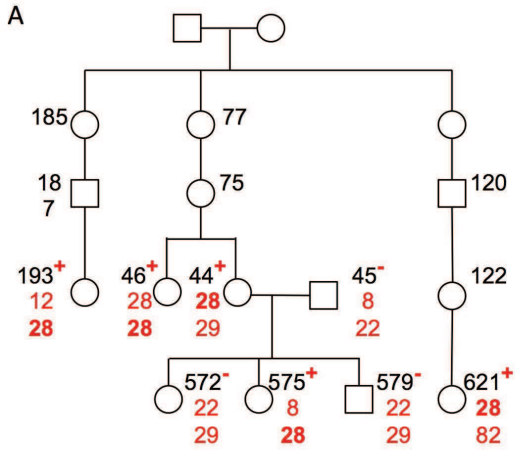

A
RNU2 CNV - y = 20.32 x - 9.73
Right Junction - y = 1.35 x + 0.08
Sequence Depth of Coverage
0
100
200
300
400
500
600
700
5
10
15
20
Whole Genome Depth of Coverage


B
Frequency
0.000
0.005
0.010
0.015
0.020
0.025
0
10
20
30
40
50
60
70
80
90
110
130
150
DOC-estimated RNU2 copy number

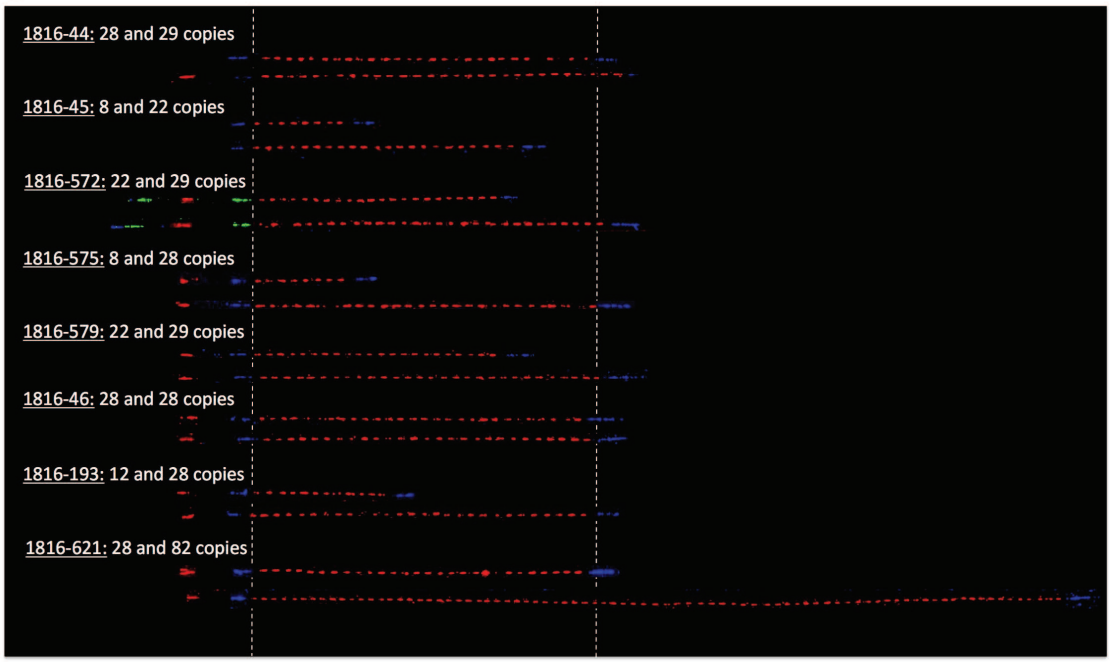

r2 = 0.93
p-value = 8.364e-05
slope = 1.40
DOC estimated copy number
Copy number measured by molecular combing
20
40
60
80
30
40
50
60
70
80
90


DCN
150
100
50
0
CHB
JPT
CHS
CDX
KHV
CEU
TSI
FIN
GBR
IBS
YRI
LWK
ASW
ACB
MXL
PUR
CLM
PEL
GIH
East Asian
European
African
Ad Mixed American
South Asian

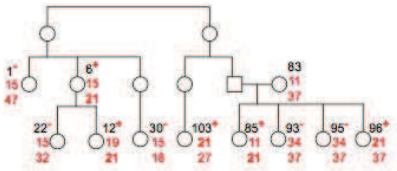
A


185
77
18
7
75
120
193+
12
28
46+
28
28
44+
28
29
45-
8
22
122
572-
22
29
575+
8
28
579-
22
29
621+
28
82

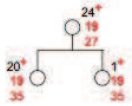
B


1816-44: 28 and 29 copies
1816-45: 8 and 22 copies
1816-572: 22 and 29 copies
1816-575: 8 and 28 copies
1816-579: 22 and 29 copies
1816-46: 28 and 28 copies
1816-193: 12 and 28 copies
1816-621: 28 and 82 copies

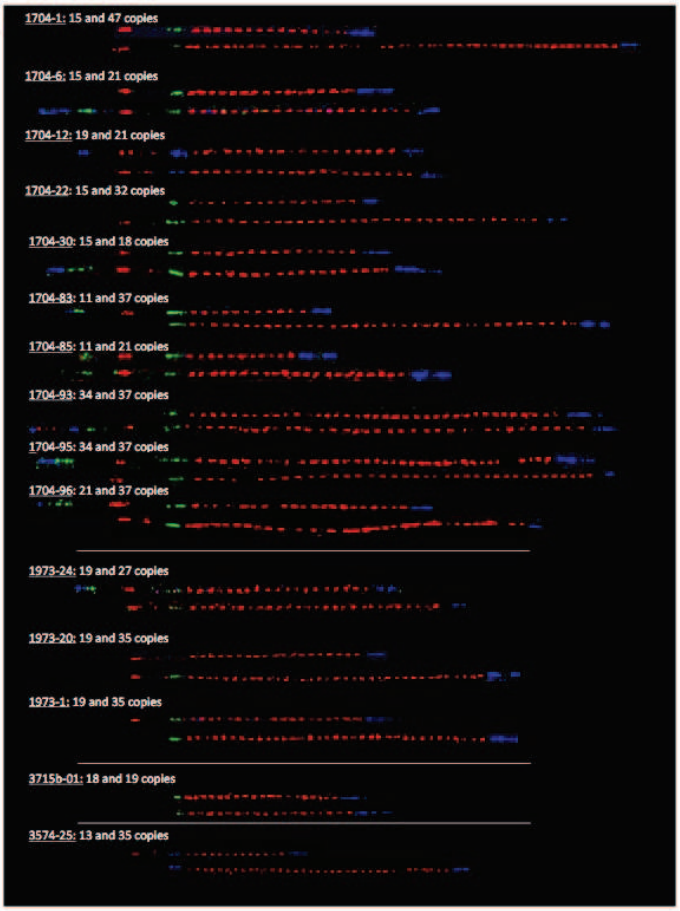
A Family 1704



Family 1973



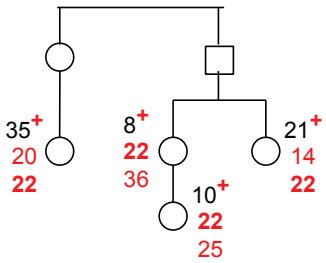
B



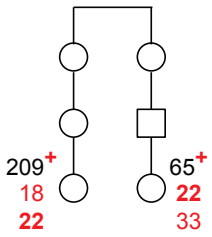
A

Mutation 0332-11 T>G

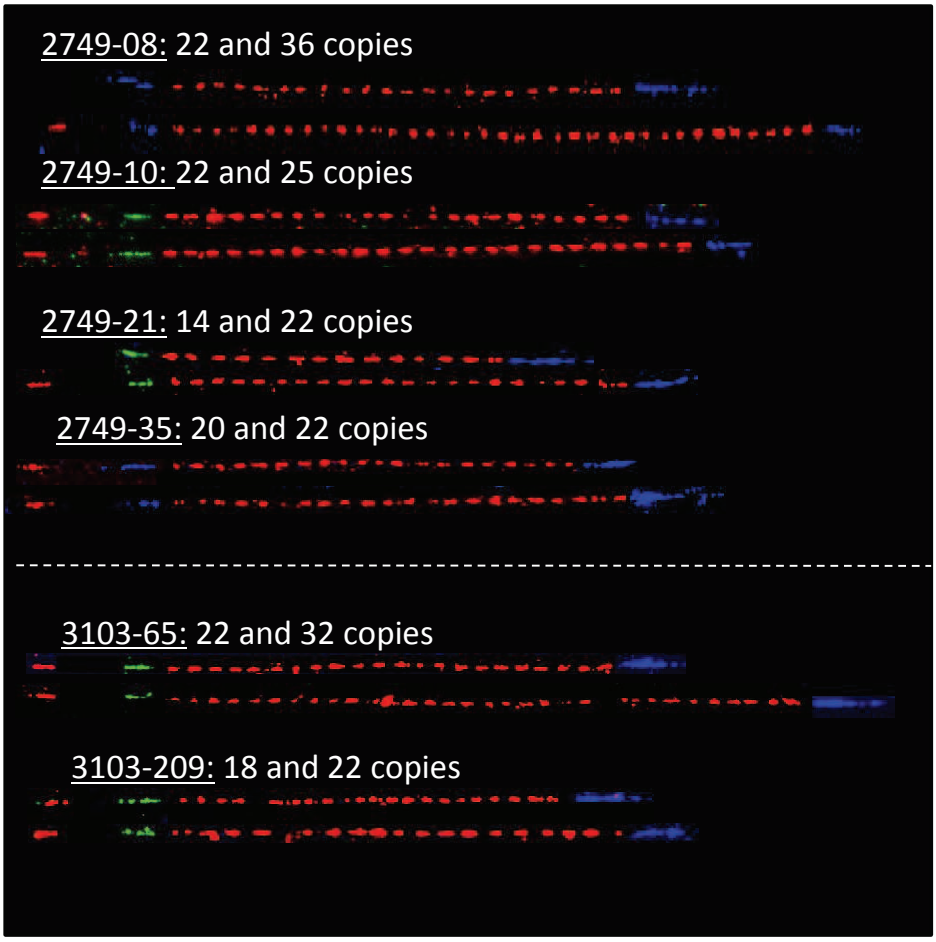
Family 2749



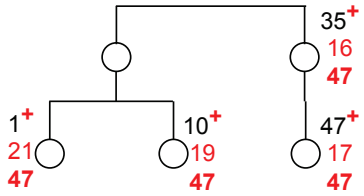
Family 3103



B



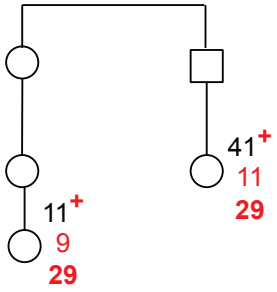
A Mutation 185delAG-ter39
Family 2979



B

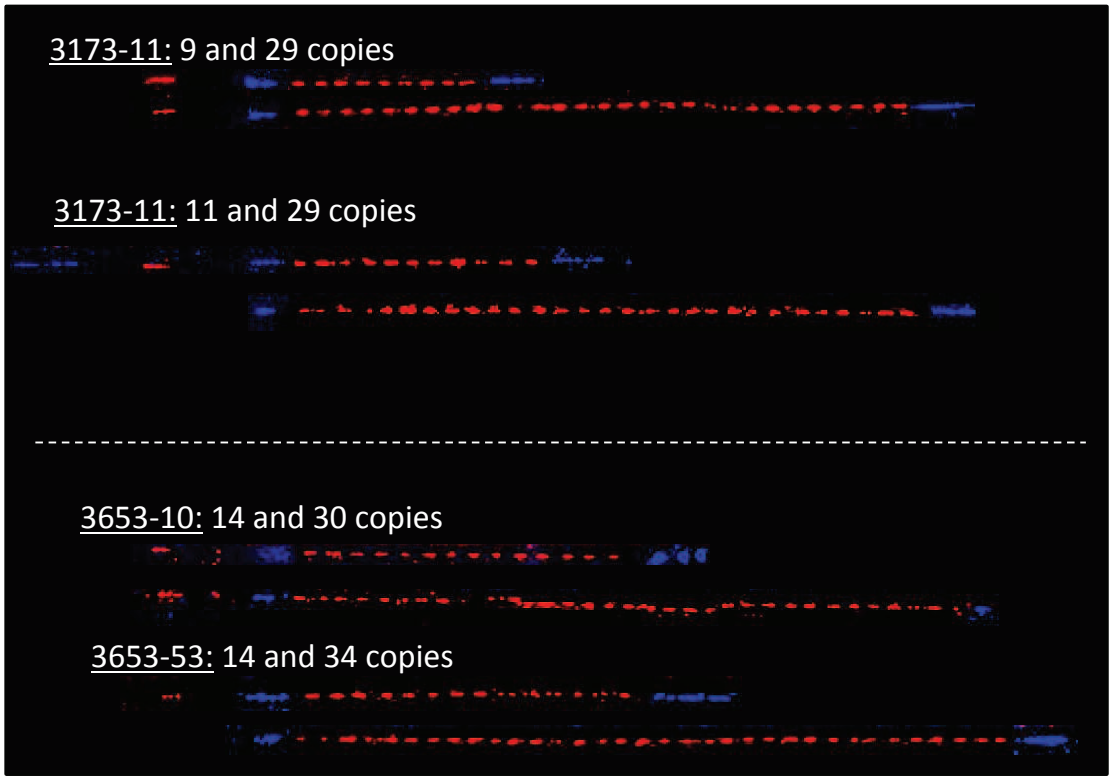
A Mutation 4305-ins6000

Family 3173



Family 3653

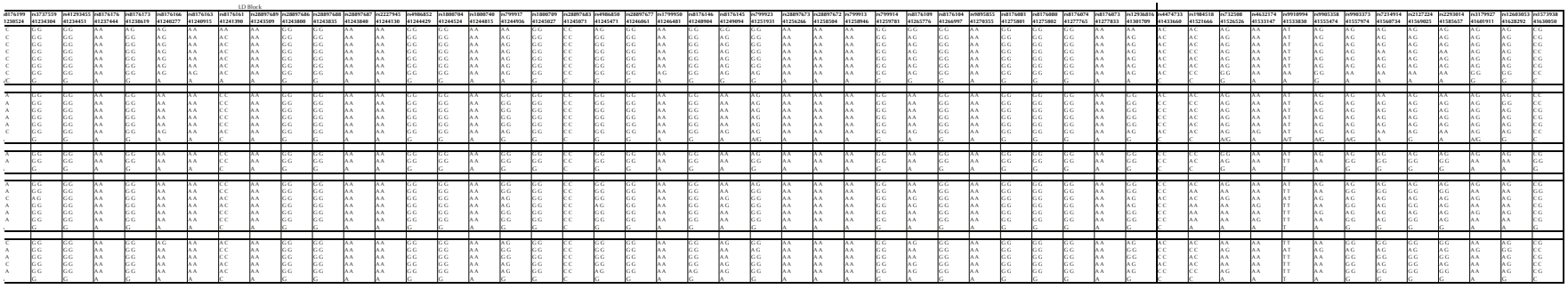
B



Mutation	Individual	SNP	rs4793039	rs11649877	rs2089114	rs12952407	rs11654121	rs2292749	rs2271028	rs1553469	rs1553468	rs4792953	rs752313	rs1078523	rs873084	rs1011157	rs11552192
		Position chr17	40760234	40772397	40777251	40791178	40797461	40818584	40841397	40842762	40849842	40873303	40901824	40912760	40930095	40960253	40962509
c68_69delAG	**2541-04**		A A	A G	A G	G G	A A	A G	A A	C C	G G	G G	A G	A G	A A	A G	A T
	2979-10		A A	A A	G G	G G	A A	G G	A G	C C	A G	A G	A A	G G	A A	G G	A A
	2979-01		A A	A A	G G	G G	A A	G G	A G	C C	A G	A G	A A	G G	A A	G G	A A
	2979-35		A A	A G	A G	G G	A A	A G	A A	C C	G G	G G	A G	A G	A G	G G	A A
	2979-47		A A	A A	G G	G G	A A	G G	A G	C C	A G	A G	A A	G G	A A	G G	A A
	3079-14		A A	A A	G G	G G	A A	G G	A G	C C	A G	A G	A A	G G	A A	G G	A A
	3261-63		A A	A A	G G	G G	A A	G G	A G	C C	A G	A G	A A	G G	A A	G G	A A
	Common haplotype		A	A	G	G	A	G	A	C	G	G	A	G	A	G	A
c.213-11T>G	**2749-08**		A G	G G	A A	A G	A G	A G	A A	C C	G G	G G	G G	A A	A A	G G	A A
	2749-10		A G	G G	A A	A G	A G	A G	A A	C C	G G	G G	G G	A A	A A	G G	A A
	2749-21		A A	G G	A A	G G	A A	A G	A A	C C	G G	G G	G G	A A	A A	G G	A A
	2749-35		A A	A G	A G	G G	A A	A G	A G	C C	A G	A G	A G	A G	A A	G G	A A
	3103-209		A A	A G	A G	G G	A A	A G	A G	C C	A G	A G	A G	A G	A A	G G	A A
	3103-65		A A	A G	A G	G G	A A	A G	A G	C C	A G	A G	A G	A G	A A	G G	A A
	Common haplotype		A	G	A	G	A	A	A	C	G	G	G	A	A	G	A
c.4186-1787_4357+4122dup	**3173-11**		A A	G G	A A	G G	A A	A A	G G	A C	A A	A G	A A	G G	A A	A G	A T
	3653-87		A G	G G	A A	A G	A G	A G	A G	A C	A G	G G	A G	A G	A A	A G	A T
	Common haplotype		A	G	A	G	A	A	G	A/C	A	G	A	G	A	G	A/T
c.5266dupC	**1704-06**		A A	A A	G G	G G	A A	G G	G G	C C	A A	A A	A A	G G	A A	G G	A A
	1704-103		A G	A G	A G	A G	A G	G G	A G	C C	A G	A G	A G	A G	A A	G G	A A
	1704-12		A A	A G	A G	G G	A A	A G	A G	C C	A G	A G	A G	A G	A A	A G	A T
	1973-20		A G	A G	A G	A G	A G	G G	A G	C C	A G	A G	A G	A G	A A	G G	A A
	3574-25		A A	A A	G G	G G	A A	G G	G G	C C	A A	A A	A A	G G	A A	G G	A A
	3715b-01		A A	A A	G G	G G	A A	G G	G G	C C	A A	A A	A A	G G	A A	G G	A A
	Common haplotype		A	A	G	G	A	G	G	C	A	A	A	G	A	G	A
	1816-44		G G	G G	A A	A A	G G	G G	A A	C C	A G	A G	A G	A G	A A	G G	A A
	1816-193		A G	G G	A A	A G	A G	A G	A A	C C	G G	G G	G G	A G	A A	G G	A A
	1816-46		A G	A G	A G	A G	A G	G G	A G	C C	A G	A G	A G	A A	A A	G G	A A
	1816-575		A G	A G	A G	A G	A G	G G	A G	C C	G G	G G	G G	A A	A A	G G	A A
	1816-621		A G	G G	A A	A G	A G	A G	A A	C C	G G	G G	G G	A A	A A	G G	A A
	Common haplotype		G	G	A	A	G	G	A	C	G	G	G	A	A	G	A

Mutation	Individual	rs16968038	rs692198	rs2667933	rs2670854	rs528854	rs6503721	rs455055	rs915945	rs12944458	rs9911630	rs4793189	rs8176320	rs8176306	rs8176305	rs8176302
		40996563	41071396	41077046	41085683	41098108	41127448	41167957	41179864	41180772	41188342	41190715	41196363	41201345	41201364	41201837
c68_69delAG	**2541-04**	A A	A G	A G	G G	A G	A A	C C	A A	A A	G G	A A	G G	C C	A G	A A
	2979-10	A A	G G	G G	A G	A A	A G	A C	A G	A G	A G	A G	G G	C C	A A	A A
	2979-01	A A	G G	G G	A G	A A	A G	A C	A G	A G	A G	A G	G G	C C	A A	A A
	2979-35	A A	G G	G G	A G	A A	A G	A C	A G	A G	A G	A G	G G	C C	A A	A A
	2979-47	A A	G G	G G	A G	A A	A G	A C	A G	A G	A G	A G	G G	C C	A A	A A
	3079-14	A A	G G	G G	A G	A A	A G	A C	A G	A G	A G	A G	G G	C C	A A	A A
	3261-63	A A	G G	G G	G G	A A	A A	A C	A G	A G	A G	A G	G G	C C	A A	A A
	Common haplotype	A	G	G	G	A	A	C	A	A	G	A	G	C	A	A
c.213-11T>G	**2749-08**	A A	A G	A G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	2749-10	A A	G G	G G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	2749-21	A A	A G	A G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	2749-35	A A	G G	G G	A A	A A	G G	A A	G G	0 0	A A	G G	G G	C C	A A	A A
	3103-209	A A	A G	A G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	3103-65	A A	G G	G G	A G	A A	A G	A C	A G	A G	A G	A G	G G	C C	A A	A A
	Common haplotype	A	G	G	A	A	G	A	G	G ?	A	G	G	C	A	A
c.4186-1787_4357+4122dup	**3173-11**	A A	A A	A A	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	3653-87	A A	A G	A G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	Common haplotype	A	A	A	A	A	G	A	G	G	A	G	G	C	A	A
c.5266dupC	**1704-06**	A A	G G	G G	A G	A A	A G	A C	G G	G G	A A	G G	G G	C C	A A	A A
	1704-103	A A	G G	G G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	1704-12	A A	A G	A G	A A	A A	G G	A A	A G	A G	A G	A G	G G	C C	A A	A A
	1973-20	A A	G G	G G	A G	A G	A G	A C	A G	A G	A G	A G	G G	C C	A G	A A
	3574-25	A A	G G	G G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	3715b-01	A A	G G	G G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	Common haplotype	A	G	G	A	A	G	A	G	G	A	G	G	C	A	A
	1816-44	A A	G G	G G	A A	A A	G G	A A	A G	A G	A G	A G	G G	C C	A A	A A
	1816-193	A A	G G	G G	A G	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	1816-46	A A	G G	G G	A A	A A	G G	A A	G G	G G	A A	G G	G G	C C	A A	A A
	1816-575	A A	A G	A G	A G	A A	A G	A C	A G	A G	A G	A G	G G	C C	A A	A A
	1816-621	A A	G G	G G	A G	A G	A G	A C	A G	A G	A G	A G	G G	C C	A G	A A
	Common haplotype	A	G	G	A	A	G	A	G	G	A	G	G	C	A	A

Mutation	Individual	rs8176273	rs8176265	rs8176259	rs28897696	rs41293459	rs8176258	rs8176245	rs8176240	rs8176236	rs8176235	rs8176225	rs8176224	rs1799967	rs4986854	rs1799966	rs28897691	rs8176212
		41211653	41213996	41215806	41215920	41215947	41216021	41217380	41218372	41218805	41219560	41221596	41221674	41222975	41223048	41223094	41226387	41226601
c68_69delAG	**2541-04**	G G	A A	A G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	G G	G G	C C
	2979-10	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	2979-01	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	2979-35	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	2979-47	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	3079-14	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	3261-63	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	Common haplotype	G	A	G	C	G	G	A	A	A	A/G	C	A	G	A	G	G	C
c.213-11T>G	**2749-08**	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	2749-10	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	2749-21	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	2749-35	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	3103-209	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	3103-65	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	Common haplotype	A	G	G	C	G	G	A	A	A	G	C	A	G	A	A	G	G
c.4186-1787_4357+4122dup	**3173-11**	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	3653-87	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	Common haplotype	A	G	G	C	G	G	A	A	A	G	C	A	G	A	A	G	G
c.5266dupC	**1704-06**	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	1704-103	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	1704-12	A G	A G	G G	C C	G G	G G	A A	A A	A A	A G	C C	A A	G G	A A	A G	G G	C G
	1973-20	A G	A G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A G	G G	C G
	3574-25	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	3715b-01	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	Common haplotype	A	G	G	C	G	G	A	A	A	G	C	A	G	A	A	G	G
	1816-44	A G	A G	G G	C C	G G	G G	A A	A A	A A	A A	C C	A A	G G	A A	A G	G G	C G
	1816-193	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	1816-46	A A	G G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A A	G G	G G
	1816-575	A G	A G	G G	C C	G G	G G	A A	A A	A A	A A	C C	A A	G G	A A	A G	G G	C G
	1816-621	A G	A G	G G	C C	G G	G G	A A	A A	A A	G G	C C	A A	G G	A A	A G	G G	C G
	Common haplotype	A	G	G	C	G	G	A	A	A	G	C	A	G	A	A	G	G


LD Block

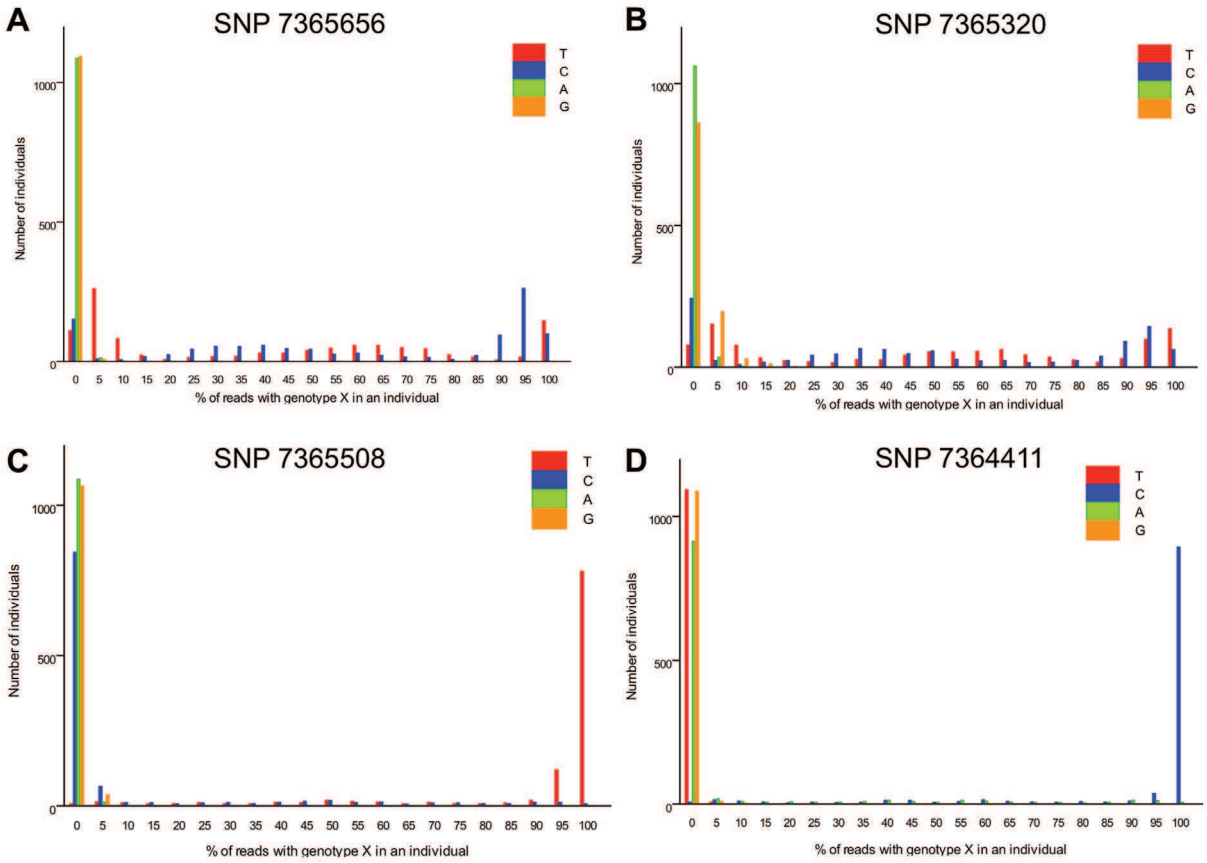


Figure 31 : Génotypage de 4 SNPs présents au sein de l'unité répétée du CNV *RNU2* pour 1106 individus du projet 1000 Génomes.
A. SNP 7365656 (Polymorphisme de restriction Sac I)
B. SNP 7365320
C. SNP 7365508
D. SNP 7364411

2.3 Etude des SNPs présents au sein de l'unité répétée du macrosatellite *RNU2* et confirmation de l'évolution concertée du locus

Nous avons également tiré profit des données de séquençage du Projet 1000 Génomes pour étudier les polymorphismes présents au sein de l'unité répétée du CNV, ainsi que des jonctions droite et gauche. L'équipe de A. Weiner avait préalablement montré dans plusieurs études que le CNV *RNU2* est soumis à l'évolution concertée, comme la plupart des larges répétitions en tandem (Ganley and Kobayashi, 2007; Gonzalez and Sylvester, 2001; Liao, 1999; Liao et al., 1997; Pavelitz et al., 1995). Cela signifie que les répétitions sont quasiment identiques entre elles. En utilisant différentes enzymes de restriction et par séquençage, ils avaient noté seulement 0,4 % de divergence entre les répétitions et identifié seulement 2 polymorphismes au sein de l'unité répétée (l'absence ou la présence d'un site de restriction Sac I, et un polymorphisme de longueur de répétitions (CT)). Deux polymorphismes avaient également été identifiés au sein de chaque jonction. D'après leur étude, seulement 6 haplotypes du locus étaient possibles.

Grâce à l'outil SAMtools SNP caller, nous avons pu identifier 350 polymorphismes potentiels, présents chez 0,09 à 99 % des 1106 individus séquencés. Nous avons décidé de nous intéresser uniquement aux SNPs, dont l'analyse est plus facile que les indels. Dans un premier temps, nous avons décidé d'étudier uniquement les SNPs présents chez 10 à 90 % des individus, afin de s'affranchir des éventuelles erreurs de séquençage qui pourraient être à l'origine des SNPs plus rares. Nous avons ainsi identifié 24 SNPs répondant à ces critères, dont le polymorphisme de restriction Sac I préalablement identifié. Cette analyse souligne que les polymorphismes fréquents sont plus abondants que prévus au sein de l'unité répétée du CNV. Pour chacun de ces 24 SNPs et pour chaque individu, nous avons regardé le nombre de lectures (ou reads) pour la base majoritaire et pour la base minoritaire (Figure 31). Mis à part pour 6 SNPs, nous avons identifié des individus pour lesquels 100 % des lectures comportaient soit

Tableau 7 : Pourcentage de lectures portant l'allèle minoritaire pour les 24 SNPs présents au sein de l'unité répétée du CNV *RNU2* chez 8 individus du projet 1000 Génomes.

Case grisée: proche de la valeur (+/- 5%) déterminée par peignage moléculaire.

SNP	NA12006	NA12272	NA12718	NA11840	NA07048	NA07051	NA12275	NA06989
7365446	0,91	23,08	2,18	41,40	3,23	45,11	0,00	45,26
7364495	0,00	0,00	0,40	47,80	4,11	46,85	0,00	36,23
7365656	5,50	0,00	3,14	42,31	5,77	48,44	3,70	47,58
7361661	0,00	0,00	0,00	49,13	0,00	48,86	0,00	46,67
7365923	7,07	0,00	2,14	45,07	3,57	43,65	0,00	47,19
7361349	0,00	0,00	0,63	44,29	0,00	37,66	0,00	40,28
7361409	1,04	0,00	0,00	44,31	0,00	38,60	0,00	42,20
7366329	1,75	0,00	2,20	45,61	17,24	44,93	8,82	50,00
7365320	8,46	0,00	7,69	46,95	6,06	38,24	12,50	41,07
7362615	11,46	0,00	5,45	46,24	7,46	49,29	3,23	50,00
7361065	0,00	0,00	2,70	0,00	0,00	0,00	0,00	0,00
7365516	28,70	0,00	9,50	48,84	5,00	41,18	25,00	40,88
7362479	47,06	0,00	34,67	29,44	42,42	35,04	18,75	27,86
7361449	0,00	34,78	2,34	2,27	0,00	5,88	3,85	3,23
7364058	27,27	0,00	48,30	29,84	23,08	23,26	23,08	31,82
7362614	0,00	43,59	0,30	8,47	0,00	17,14	0,00	14,39
7364321	27,07	0,00	44,71	3,94	26,53	4,46	26,92	6,73
7364065	48,84	0,00	14,81	9,65	30,00	13,95	14,29	6,25
7361477	0,00	0,00	0,00	6,47	0,00	49,15	0,00	27,27
7363718	0,00	15,79	0,30	21,39	0,00	5,64	0,95	10,76
7362285	13,76	0,00	17,29	16,92	21,28	14,71	6,45	0,00
7365508	0,00	0,00	0,88	0,00	2,33	0,00	0,00	0,00
% Attendu Allèle minoritaire (Peignage Moléculaire)	45,71	22,22	40,91	47,62	38,89	48,48	34,62	36,54

Tableau 8 : SNPs identifiés au sein de la séquence codante du gène *RNU2*, par analyse des données de séquençage de 1106 individus du projet 1000 Génomes.

Position hs37d5	Position U57614	Position snRNA U2	Nombre d'individus porteurs de la base alternative (SNP Caller)	Fréquence	Base majoritaire (1,000G)	Base alternative (1,000G)
7364725	4881	0	46	4,16	G	A
7364707	4899	18	10	0,90	A	G
7364697	4909	28	1	0,09	G	A
7364677	4929	48	1	0,09	T	A
7364663	4943	62	24	2,17	A	G
7364661	4945	64	13	1,18	T	A
7364655	4951	70	4	0,36	G	A
7364652	4954	73	3	0,27	G	A
7364647	4959	78	2	0,18	G	A
7364642	4964	83	1	0,09	T	C
7364640	4966	85	5	0,45	T	C
7364638	4968	87	1	0,09	A	G
7364631	4975	94	1	0,09	T	C
7364629	4977	96	1	0,09	A	G
7364625	4981	100	1	0,09	A	G
7364621	4985	104	29	2,62	A	G
7364611	4995	114	18	1,63	T	C
7364606	5000	119	1	0,09	C	T
7364604	5002	121	1	0,09	T	C
7364602	5004	123	1	0,09	T	C
7364601	5005	124	5	0,45	C	A
7364600	5006	125	1	0,09	C	A
7364597	5009	128	1	0,09	G	A
7364567	5039	158	3	0,27	C	A
7364565	5041	160	1101	99,55	T	C
7364563	5043	162	1	0,09	A	G
7364551	5055	174	14	1,27	T	C
7364537	5069	188	1	0,09	G	A

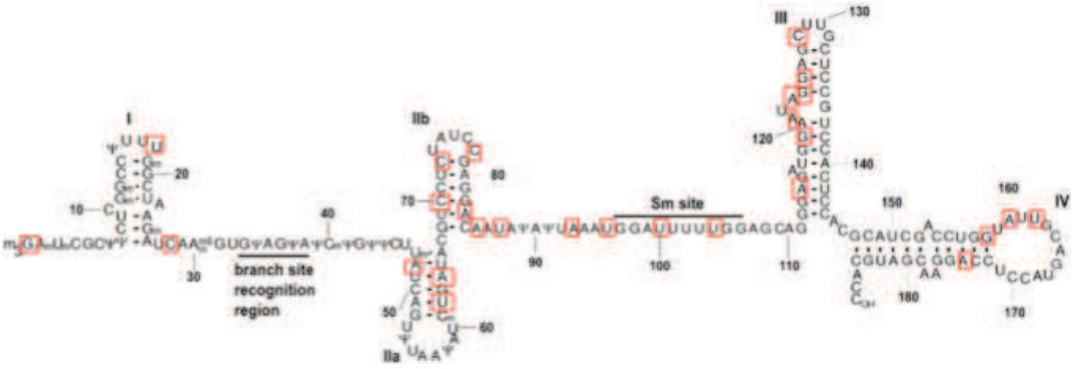


Figure 32 : Structure secondaire de l'ARNsn U2 et position des SNPs identifiés avec les données du projet 1000 Génomes (rouge).
Adapté de Dönmez et al., 2004.

la base minoritaire soit la base majoritaire, et ceci même pour des SNPs peu fréquents. Cela signifie que ces individus sont homozygotes, et que chaque copie de chaque allèle porte la même base, éventuellement rare. Dans ce cas, la probabilité que cette mutation soit apparue sur chaque répétition de manière indépendante est extrêmement faible et une homogénéisation intrachromosomique des répétitions parait beaucoup plus probable pour expliquer de telles observations. Ainsi, ces données sont en accord avec les résultats antérieurs montrant que le locus *RNU2* est soumis à l'évolution concertée. Par ailleurs, pour les individus hétérozygotes (*i.e.* pour lesquels au moins une copie porte la base alternative), nous observons un gradient de fréquence. Dans le cas d'un locus présent en une seule copie par génome haploïde, un individu hétérozygote porte 50 % de lectures avec l'allèle majoritaire, et 50 % de lectures avec l'allèle minoritaire. Pour le CNV *RNU2*, chaque individu portant un nombre variable de copies sur chacun de ses allèles, ce gradient de fréquence était attendu. La fréquence génotypique observée pour les hétérozygotes pourrait, du fait de l'évolution concertée, être le reflet du rapport entre le nombre de copies de chaque allèle (les copies de l'un portant la base majoritaire, et celles de l'autre la base minoritaire). J'ai donc comparé ce rapport avec la valeur théorique pour les 8 individus étudiés par peignage moléculaire (Tableau 7). Pour chacun des individus sauf un, j'ai retrouvé un ou plusieurs SNPs pour lesquels le rapport du nombre de reads est un reflet de la répartition allélique du nombre de copies. Cependant, il n'est pas possible d'identifier un SNP unique qui permettrait de déterminer la répartition allélique chez tous les individus. Ces résultats obtenus sur un faible échantillonnage, bien qu'encourageants, doivent être approfondis par l'étude d'un plus grand nombre d'individus suivis d'une analyse statistique adéquate.

En parallèle, nous avons regardé plus en détail les polymorphismes au sein de la séquence codante de l'ARNsn U2. Nous avons identifié 28 SNPs potentiels, présents chez 0,01 à 99,55 % des individus (Tableau 8), et répartis sur l'ensemble de la séquence de l'ARNsn U2 (Figure 32).

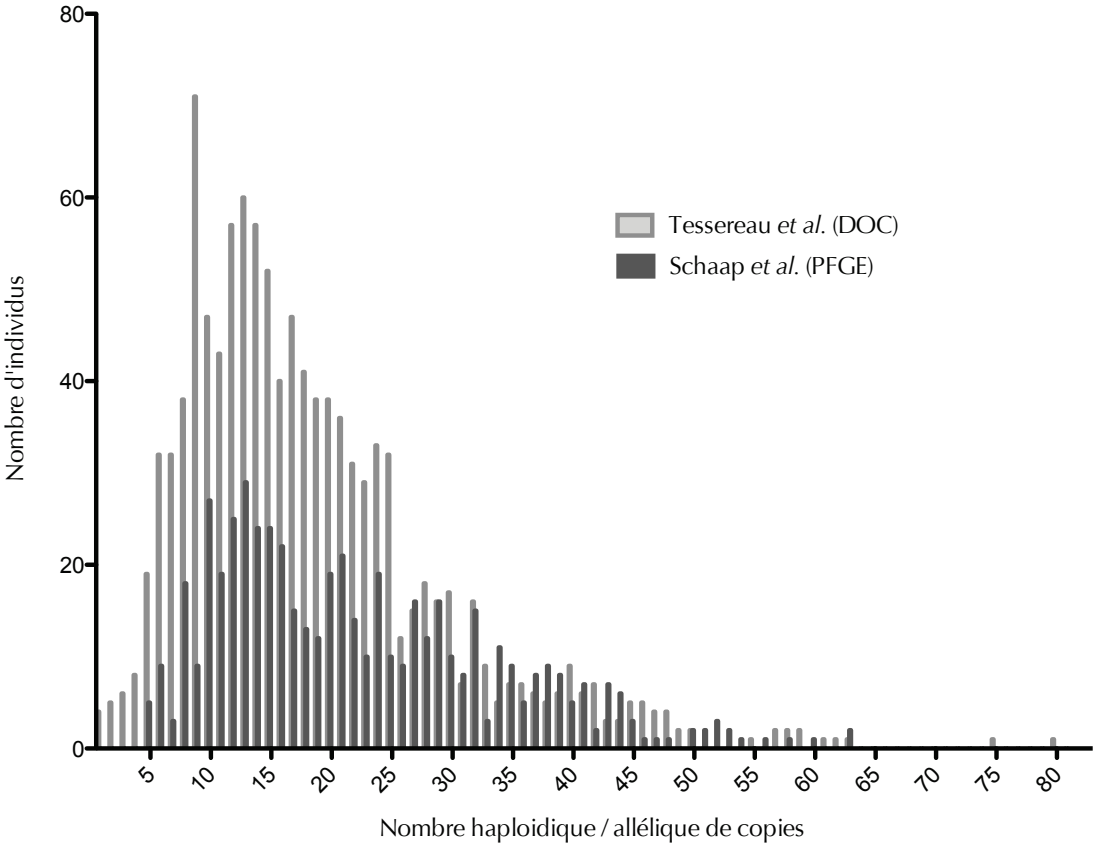


Figure 33 : Distribution des nombres haploïdiques (estimés par profondeur de couverture) ou alléliques (*Schaap et al., 2013,* déterminés par électrophorèse en champs pulsés) de copies du macrosatellite *RNU2*.
DOC: Depth-of-coverage
PFGE: Pulse Field Gel Electrophoresis

2.4 Discussion

L'utilisation des données de séquençage du projet 1000 Génomes a permis pour la première fois d'étudier le niveau de polymorphisme d'un macrosatellite sur un très grand nombre d'individus issus de différentes populations humaines modernes. En comparant nos résultats avec ceux obtenus avec la technique du peignage moléculaire pour 8 individus, nous avons montré que la profondeur de couverture de séquençage d'un locus hautement répété était un très bon indicateur du nombre haploïdique de copies. Il est intéressant de noter que notre étude et celle de Schaap *et al.* (réalisée par électrophorèse en champ pulsé sur 210 individus) (Schaap et al., 2013) donnent des résultats assez similaires quant à la fréquence de chaque allèle du CNV *RNU2* (Figure 33). Le nombre allélique de copies (NAC) le plus fréquemment observé dans l'étude de Schaap *et al.* est de 12, et le nombre allélique moyen de copies est de 22 (Schaap et al., 2013). Le nombre haploïdique de copies le plus fréquent d'après notre étude est de 9. Dans l'étude de Schaap *et al.*, un NAC de plus de 30 n'est observé que chez 25 % des individus. Dans notre étude des données 1000 Génomes, un nombre haploïdique de copies supérieur à 24 est observé chez 25 % des individus. Ainsi, les allèles porteurs d'un grand nombre de copies (> 30) sont extrêmement rares, ce qui augmente la probabilité qu'ils puissent avoir un effet biologique important.

L'utilisation des données de séquençage pour génotyper les macrosatellites permet donc de mieux définir le niveau de polymorphisme de ces séquences et de mieux les caractériser. Il sera nécessaire d'appliquer ce type d'analyse à d'autres macrosatellites connus, présents ou absents du génome de référence. Nous aurions aimé pouvoir utiliser les SNPs fréquents que nous avons identifiés pour quantifier le NAC du locus *RNU2* par séquençage nouvelle génération. En effet, étant soumise à l'évolution concertée, chaque copie d'un allèle devrait porter les mêmes polymorphismes. Pour un individu hétérozygote, il devrait être possible d'estimer le NAC en évaluant le nombre de lectures porteuses d'une base ou de l'autre. La plupart des macrosatellites étant également soumis à l'évolution concertée, cette approche pourrait également leur être appliquée. Ainsi, les données de séquençage pourraient être utilisées non seulement

pour définir le nombre haploïdique (donc global) de copies, mais surtout pour déterminer la répartition entre chaque allèle, ce qui rendrait envisageable le génotypage des macrosatellites lors d'études à large échelle par séquençage nouvelle génération, qui, pour le moment, ne les prennent pas en compte. Cependant, les résultats préliminaires que nous avons obtenus avec 8 individus, pour lesquels nous avons pu déterminer le NAC, n'ont pas permis d'identifier un SNP pour lequel le nombre de lectures de chaque base était systématiquement fonction du NAC. Des analyses plus poussées devront être effectuées.

Par ailleurs, parmi les SNPs que nous avons identifiés, certains sont retrouvés au sein de la séquence codant pour l'ARNsn U2, qui occupe une fonction essentielle au sein du spliceosome. Il serait très intéressant d'étudier les conséquences fonctionnelles de chacun de ces polymorphismes. Grâce aux lignées lymphoblastoïdes établies pour chacun de ces individus, il serait intéressant de regarder si les copies porteuses de ces polymorphismes sont exprimées (puisque toutes les répétitions ne sont pas transcrites). Dans un second temps, il serait intéressant de regarder si cette expression est tissu-dépendante. En effet, une délétion de 5 nucléotides dans l'une des cinq répétitions du gène chez la souris se traduit par une expression différentielle de l'ARN muté dans le cerveau, entraînant de graves défauts d'épissage et une neurodégénération (Jia et al., 2012). En conclusion, le nombre de copies d'un macrosatellite n'est pas le seul paramètre à prendre en compte, les variations génétiques entre chaque copie pouvant également avoir des conséquences phénotypiques.

3 Le CNV *RNU2* et la prédisposition génétique au cancer du sein

3.1 Introduction

Chaque année, en France, environ 1500 familles prédisposées au cancer du sein effectuent un test génétique, *i.e.* une recherche d'altérations sur les gènes *BRCA1/2*. Ce test permet d'identifier l'altération à l'origine de la prédisposition familiale et les individus porteurs de cette altération, et donc à haut risque de développer un cancer du sein. Des recommandations nationales de prise en charge sont préconisées pour les porteuses de mutations, incluant entre autre une surveillance clinique des seins 2 à 3 fois par an dès l'âge de 25 ans, et une mammographie annuelle dès l'âge de 30 ans. Cependant, dans 80 % des familles analysées lors du diagnostic, le test est négatif, aucune anomalie clairement délétère n'ayant pu être identifiée. Cela ne signifie pas qu'il n'existe pas de prédisposition au cancer, mais que le gène ou le mécanisme en cause n'est pas encore connu.

Les études récentes ayant pour but d'identifier de nouveaux allèles de prédisposition au cancer du sein n'ont permis d'identifier que des variants de pénétrance faible à modérée conférant des risques modestes. Nous avons émis l'hypothèse qu'un nombre anormalement élevé de copies du CNV *RNU2* pourrait influencer l'expression de *BRCA1*, et de ce fait être impliqué dans la prédisposition au cancer du sein. Pour tester cette hypothèse, nous avons étudié le nombre de copies dans une large étude de cas de cancer du sein et témoins appariés, ainsi que les répercussions de la variation du nombre de copies sur l'expression globale et allélique de *BRCA1*. Les résultats que nous avons obtenus sur deux études (GENESIS et BCFR) sont présentés dans cette section ainsi que dans l'annexe.

3.2 Matériel et Méthodes

- Sujets

Une partie des sujets analysés provient de l'étude GENESIS (GENE SISter), une étude nationale mise en place en 2007, promue par le réseau UNICANCER et coordonnée par N. Andrieu, D. Stoppa-Lyonnet (Institut Curie) et O. Sinilnikova (notre équipe). Elle a obtenu l'avis favorable du Comité de Protection des Personnes Ile de France III. L'objectif de cette étude est d'identifier de nouveaux allèles de prédisposition par étude cas-témoins. Les personnes ont été recrutées jusqu'en Décembre 2012 par l'intermédiaire de toutes les consultations d'oncogénétique du territoire français. Les cas index sont des femmes ayant développé un cancer du sein, non porteuses de mutations sur *BRCA1/2*, et dont une des sœurs a également été atteinte d'un cancer du sein. Les sœurs, atteintes ou indemnes, ont également été recrutées dans la mesure du possible, ainsi que d'autres membres de leurs familles, atteints ou non. Les témoins appariés de cette étude sont les meilleures amies ou les collègues non atteintes de ces cas index. Les inclusions dans l'étude (information, questionnaires, acheminement des prélèvements sanguins, numérisation des mammographies, ...) sont gérées par le Centre Coordinateur (CC) (Nadine Andrieu, Inserm U900, Paris), les prélèvements biologiques sont centralisés par le Centre de Ressource Biologique (CRB) (Olga Sinilnikova, Hospices Civils /Centre Léon Bérard, Lyon). Actuellement, 5128 individus ont été recrutés : 1663 cas index de cancer du sein sans mutation connue dans les gènes *BRCA1* et *BRCA2*, 719 sœurs atteintes, 1338 apparentés et 1408 témoins appariés. L'ADN de chaque individu a été extrait à partir de sang congelé, grâce au robot Autopure LS (Qiagen), et mis en plaque de 96 puits à la concentration de 25 ng/µl. Pour tous les cas, les lymphocytes B ont également été purifiés et congelés, permettant d'établir éventuellement par la suite des lignées lymphoblastoïdes. Au moment où le génotypage du CNV *RNU2* a été réalisé, 4163 échantillons d'ADN étaient disponibles (1400 cas, 1619 apparentés, 1144 témoins).

La deuxième série de sujets provient de l'étude BCFR (Breast Cancer Families Registry) (John et al., 2004). Les personnes ont été recrutées entre 1995 et 2005 dans 3

centres (Cancer Care Ontario, the Cancer Prevention Institute of California et University of Melbourne). Les cas sélectionnés (n = 1330) ont été diagnostiqués avec un cancer du sein avant 45 ans, et sont d'origine caucasienne, asiatique de l'Est, latino-hispanique ou afro-américaine. Les témoins (n = 1123) de même origine ont été recrutés à un âge inférieur de 10 ans au moins par rapport à l'âge au diagnostic des patients inclus dans le même centre. L'appariement a été fait par *frequency matched* : un témoin est apparié à un cas d'environ le même âge et de même groupe éthnique. D'après les données recueillies, 25,4 % des cas ont au moins un apparenté au premier degré ayant développé un cancer du sein (contre 9,8 % des témoins), 35,7 % au second degré (contre 21,6 % des témoins), 3,4 % des cas ont au moins un apparenté au premier degré ayant développé un cancer de l'ovaire (contre 2,3 % des témoins) et 4,8 % au second degré (contre 4,1 % des témoins) (Le Calvez-Kelm et al., 2012). Actuellement, 2453 ADNs ont été mis en plaque 96 puits et conservés à la concentration soit de 5 ng/µL, soit de 10 ng/µL, soit de 20 ng/µL : 1330 provenant de cas index et 1123 de témoins. Pour notre étude, deux fois 20 ng d'ADN ont été prélevés pour chaque sujet, séchés et mis en plaque 384 puits. Pour quelques cas, des lignées lymphoblastoïdes ont été préalablement établies à partir de lymphocytes B purifiés et congelés. Pour le génotypage du CNV *RNU2*, seulement 1149 cas et 1017 témoins, répartis sur 7 plaques 384 puits, ont été analysés.

- Estimation du nombre global de copies par PCR quantitative

Le nombre global de copies (NGC) du CNV *RNU2* a été estimé en utilisant un protocole de qPCR en duplex en chimie TaqMan. Chaque échantillon a été analysé en duplicat sur des plaques 96 puits pour l'étude GENESIS et en plaques 384 puits pour l'étude BCFR, selon le mélange suivant (Volume final 20 µL): ADN 20 ng, TaqMan Universal PCR Master Mix 1X (Applied Biosystems), TaqMan Copy Number Reference Assay *RNAse P* 1X (Applied Biosystems), sonde *RNU2* (5'-FAM-ACGGAACGCACAGGAGCAGAGTAMRA-TAMRA-3') 50 nM, amorce *RNU2* sens (5'-GAGGTGCAGGTAGTATAAGCCATT-3') 0,1 µM, amorce *RNU2* anti-sens (5'-GAGCCACGATGCTTGGAC-3') 0,1 µM. Les conditions de réaction ont été choisies conformément aux recommandations du fournisseur : 2 minutes à 50°C, 10 minutes à

95°C, puis 40 cycles de 15 secondes à 95°C et 1 minute à 60°C, dans un appareil StepOnePlus Real-Time PCR System (Applied Biosystems) pour l'étude GENESIS et dans un appareil 7900 HT Sequence Detection System (ABI Prism) pour l'étude BCFR.

Un individu calibrateur (GE522), dont j'ai estimé précisément le nombre de copies par peignage moléculaire (Allèle 1 : 13 copies, Allèle 2 : 29 copies), a été analysé sur chaque plaque 96 puits ou 384 puits. Un autre échantillon également caractérisé par peignage moléculaire (Allèle 1 : 14 copies, Allèle 2 : 23 copies), a aussi été analysé sur chaque plaque, servant ainsi de contrôle interne pour tester la reproductibilité de l'essai.

La quantité relative de copies par rapport au calibrateur a été calculée pour chaque échantillon selon les recommandations du logiciel StepOne-Software v 2.1 pour l'étude GENESIS et SDS v 2.1 pour l'étude BCFR d'après l'équation :

1) $RQ = \frac{0{,}98^{(Ct\ CNV_{calib} - Ct\ CNV_{éch})}}{0{,}90^{(Ct\ RNAse\ P_{calib} - Ct\ RNAse\ P_{éch})}}$

Le NGC a ensuite été estimé selon l'équation suivante :

2) $NGC = RQ \times 42$

- Culture cellulaire

Les lignées lymphoblastoides ont été établies après immortalisation par le virus d'Epstein-Barr de lymphocytes B prélevés lors de l'inclusion dans l'étude GENESIS de 20 cas index. Les cellules ont été cultivées dans un milieu RPMI (Roswell Park Memorial Institute) supplémenté avec 20 % de sérum de veau fœtal (SVF) et 1 % de pénistreptomycine, dans une étude à 37°C avec 5 % de CO_2.

- Peignage moléculaire

La caractérisation du NAC par peignage moléculaire a été effectuée en suivant le protocole décrit dans les articles précédents (Article 1 & Article 2). Pour deux témoins (GE1808 et GE3649) et un cas index (GE1966), l'ADN a été extrait à partir de 300 µL d'un prélèvement de sang congelé (et non pas de lignées lymphoblastoïdes), avec le kit Gentra Puregene Blood Kit (Qiagen) en suivant les recommandations du fournisseur et

en réalisant le traitement par la RNAse. La quantité d'ADN ainsi extraite est dosée grâce au kit Qubit dsDNA BR Assay (Invitrogen) en suivant les recommandations du fournisseur.

- Electrophorèse en champs pulsés

Préparation des plugs :

Les plugs pour le protocole de PFGE ont été préparés à partir de 5.10^6 cellules en culture avec le kit CHEF Mammalian Genomic DNA (Biorad) en suivant les recommandations du fournisseur. Brièvement, les cellules sont resuspendues dans un mélange 2:1 Cell Suspension Buffer – Clean Cut Agarose 2 % (préalablement équilibrées à 50°C). Après 30 minutes à 4°C, les plugs sont traitées par la Protéinase K sur la nuit à 50°C, puis lavés 4 fois dans du Wash Buffer 1X ainsi qu'une fois dans du PMSF 1 mM. Les plugs sont ensuite digérés sur la nuit par EcoRI (30 U) à 37°C.

Migration et transfert sur membrane:

La migration est réalisée dans un gel 1,8 % d'agarose (Pulse Field Certified Agarose, Bio Rad) avec du TBE 0,5X pendant 44 h à 180 Volts, avec un temps de pulse variant de 50 à 90 sec. La taille des fragments est estimée en comparaison avec le marqueur de taille Lambda Ladder (Bio Rad). Plusieurs lavages sont ensuite réalisés : 20 min dans une solution 0,25N HCl, 5 min dans l'eau, 30 min dans une solution 0,5M NaOH – 1,5M NaCl, 5 min dans l'eau, 15 min dans du tampon de neutralisation. Le transfert sur une membrane Hybond N+ (Amersham) est réalisé sur la nuit en présence de SSC 10X. La membrane est cross-linkée aux UVs grâce à l'appareil UV Stratalinker 1800 (Stratagene), puis hybridée avec des sondes radioactives.

Préparation de la sonde radioactive :

Le fragment PCR est amplifié à partir de 25 ng d'ADN en utilisant les amorces 5'-ACGACGCAGTTAGGAGGCTA-3' (60°C) et 5'-TTGCGTTAGGAAGGAGGAAG-3' (59°C), puis marqué avec des nucléotides α-32P-

dCTP en suivant les recommandations du kit Prime-It II Random Primer (Stratagene), puis purifié sur colonne MicroSpin G50 (Amersham). Le fragment est ensuite dénaturé à 95°C pendant 5 min.

Hybridation :

La membrane est préhybridée pendant 1 h à 65°C avec du tampon Church, puis hybridée avec la sonde marquée sur la nuit à 65°C, et lavée deux fois pendant 10 min dans du SSC 2X, pendant 10 min avec du SSC 2X SDS 0,1 %, pendant 45 minutes avec du SSC 0,1X SDS 0,1 %, et 2 h dans du SSC 0,1 X SDS 0,5 %. La membrane est ensuite mise en présence avec un film autoradiographique à -80°C pendant 24 à 72 h avant révélation avec l'appareil Curix 60 (AGFA HealthCare).

- Extraction d'ADN à partir de lignées lymphoblastoïdes

L'ADN est extrait à partir des lignées lymphoblastoïdes en culture grâce au kit NucleoSpin Tissue (Macherey-Nagel) en suivant les instructions du fabricant.

- Fluorescent *In Situ* Hybridization

L'hybridation sur chromosomes en métaphase a été réalisée à partir des lignées lymphoblastoides cultivées, en suivant le protocole décrit préalablement (Schluth-Bolard et al., 2009) et également utilisé dans l'Article 1.

- Analyse statistique des résultats de qPCR

Les analyses statistiques des résultats obtenus pour les cas index et les témoins ont été réalisées sous le logiciel R, en utilisant une analyse de la variance ou « Anova » (ANalysis Of Variance) à deux facteurs (probabilité d'être malade et effet de la plaque de qPCR). Afin de savoir si la relation entre la probabilité de développer un cancer du sein et et le NGC est linéaire (proportionnelle), plusieurs modèles linéaires généralisés

ont aussi été testés. La variable réponse a été modélisée par une variable Binomiale et transformée selon un lien logit :

- Modèle nul : le nombre de copies n'a pas d'effet sur la probabilité de développer un cancer du sein.
- Modèle linéaire : la probabilité de développer un cancer du sein augmente de manière proportionnelle avec le nombre de copies.
- Modèle à seuil : la probabilité de développer un cancer du sein augmente de manière proportionnelle avec le nombre de copies à partir d'un certain nombre seuil de copies.
- Modèle quadratique: l'effet du nombre de copies sur la probabilité de développer un cancer du sein est non linéaire et peut être renforcé ou diminué quand le nombre de copies augmente

Le meilleur modèle a été sélectionné par AIC (Critère d'information d'Akaike). L'AIC est un critère d'évaluation des modèles linéaires généralisés, dont les paramètres ont été estimés par maximum de vraisemblance. On compare donc l'AIC obtenu pour chaque modèle incluant notre variable d'entrée (ici le NGC pour les modèles linéaire, à seuil et quadratique) à la valeur de l'AIC obtenue pour un modèle plus simple n'incluant pas cette variable d'entrée (*i.e.* le modèle nul). Le modèle présentant l'AIC le plus faible est le meilleur modèle pour expliquer nos données, *i.e.* celui qui s'ajuste le plus correctement aux données avec le minimum de paramètres (Akaike, 1974).

- Extraction de l'ARN et Rétro-transcription

Les ARN totaux ont été isolés à partir des lignées lymphoblastoïdes avec le kit NucleoSpin miRNA (Macherey-Nagel) selon les recommandations du fournisseur. 50 ng d'ARN ont été rétrotranscrits avec le kit Expand Reverse Transcriptase (Roche Diagnostic) avec 1 µg d'OligodT et de Random Primers (Promega) en suivant les recommandations du fournisseur.

- PCR quantitative

La réaction a été faite sur 2 µl de la réaction de rétrotranscription diluée 50 fois avec le kit GoTaq qPCR Master Mix (Promega), dans un appareil StepOnePlus Real-Time PCR (Life Technologies) dans des plaques 96 puits en suivant les étapes suivantes : 95°C pendant 10 min, 40 cycles à 95°C pendant 15 min et 60°C pendant 1 min. Les amorces utilisées pour amplifier le transcrit *BRCA1* (nucléotides 5052 à 5143 de la séquence U14680) ont été designées avec le logiciel Primer3 v.0.4.0 (http://frodo.wi.mit.edu/primer3/) et synthétisées par Eurofins MWG Operon : 5'-AGGGTCAACAAAAGAATGTCCA-3' et 5'-GTGATGTGGTGTTTTCTGGCAA-3'. L'expression du gène *BRCA1* a été normalisée grâce aux valeurs obtenues pour le gène de ménage *GAPDH* : 5'-CGGAGTCAACGGATTTGGTCGTAT-3' et 5'-AGCCTTCTCCATGGTGGTGAAGAC-3'. Pour quantifier l'expression allélique de *BRCA1*, nous avons utilisé les sondes Taqman ciblant spécifiquement le SNP rs1799966 du kit de génotypage TaqMan SNP Genotyping Assay (Life Technologies) en suivant les recommandations du fournisseur. Chaque échantillon a été analysé en duplicat. L'expression de *BRCA1* a été évaluée avec la méthode comparative du ΔΔCt en suivant les recommandations du fournisseur.

3.3 Résultats

- **Mise au point d'un test haut-débit permettant d'évaluer le NGC du CNV *RNU2***

Afin de mettre en évidence un effet du nombre de copies du CNV *RNU2* sur le risque de développer un cancer du sein familial, il était nécessaire de comparer les valeurs obtenues pour un grand nombre d'individus ayant développé un cancer du sein et pour des témoins. Nous avons montré dans la section précédente (Chapitre 2 – Section 2) que les données de séquençage nouvelle génération peuvent être utilisées

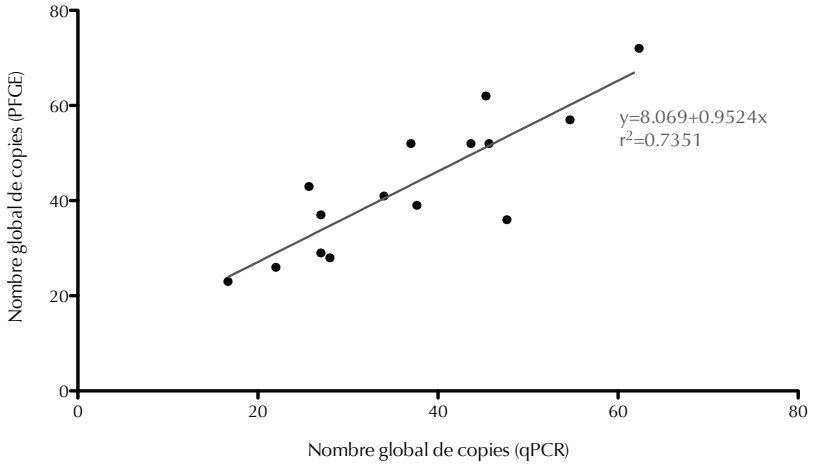


Figure 34 : Comparaison du nombre global de copies du macrosatellite *RNU2* mesuré par qPCR et par électrophorèse en champs pulsés (PFGE) chez 15 individus.

pour conduire ce type d'analyse. Cependant, aucune étude cas-témoins portant sur des cas de cancers du sein familiaux n'a pour le moment été conduite par séquençage nouvelle génération sur le génome complet, rendant l'accès à des données pré-existantes impossible. La technique de peignage moléculaire que nous avons mise au point permet de décrire le NAC très précisément, mais n'est actuellement pas adaptée à la caractérisation d'un très grand nombre d'individus, pas plus que ne l'est l'électrophorèse en champs pulsés étant donné que ces techniques restent très chronophages, et nécessitent d'accéder à des lignées cellulaires pour chaque individu, ce qui limite les échantillons que nous pouvions utiliser.

Nous avons donc mis au point une nouvelle méthode d'analyse du CNV *RNU2* basée sur la PCR quantitative (qPCR) en chimie TaqMan. Ce protocole ne permet pas de quantifier le NAC, mais uniquement le NGC. Bien que ne permettant pas de quantifier à la copie près, ce protocole offre l'avantage de pouvoir analyser un grand nombre d'individus en utilisant l'ADN extrait à partir de prélèvements de sang, et ainsi de pouvoir discriminer les individus présentant un nombre global anormalement élevé ou faible de copies.

Dans un premier temps, nous avons comparé pour 15 individus les résultats obtenus avec notre test qPCR avec ceux obtenus par électrophorèse en champs pulsés, la technique la plus fréquemment utilisée par d'autres équipes (Figure 34) (Liao et al., 1997; Schaap et al., 2013). Nous avons obtenu une bonne corrélation entre les deux techniques (test de corrélation de Pearson, $r^2 = 0,7351$, p-value $< 0,0001$).

Dans un second temps, nous avons étudié l'effet de la qualité de l'ADN ainsi que du type d'extraction sur la mesure par qPCR du NGC. En effet, des études antérieures avaient montré que les résultats obtenus par qPCR étaient dépendants de la qualité et de la concentration de l'ADN (dépendants en partie du type d'extraction), rendant les résultats parfois difficilement interprétables (Armour et al., 2007; Clayton et al., 2005; Fernandez-Jimenez et al., 2011; Guescini et al., 2008). Pour cela, nous avons mesuré par qPCR le NGC sur de l'ADN extrait à partir de lignées lymphoblastoïdes en culture pour 27 individus de l'étude GEMO, qui avaient été également caractérisés par peignage moléculaire. Pour 24 individus pour lesquels le prélèvement était disponible, nous avons comparé ces résultats avec ceux obtenus à partir d'ADN extrait à partir

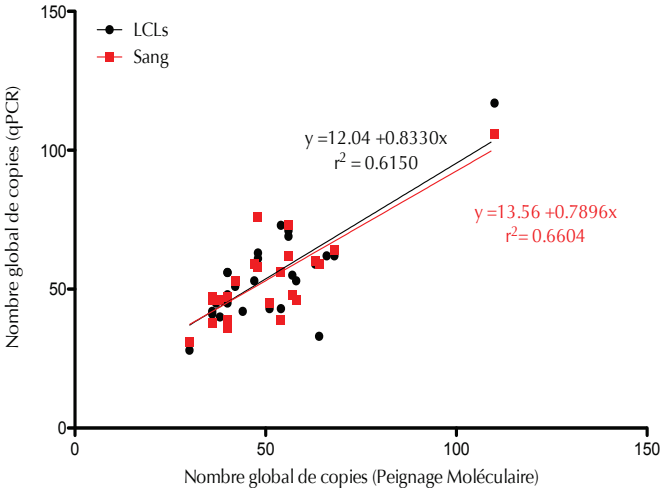


Figure 35 : Comparaison du nombre global de copies du macrosatellite *RNU2* mesuré par peignage moléculaire et par qPCR, sur des échantillons provenant de lignées lymphoblastoïdes (LCLs) ou de sang.

Tableau 9 : Récapitulatif des nombres globaux de copies du macrosatellite *RNU2* déterminés par qPCR dans les échantillons de l'étude GENESIS.

	Cas Index	Sœur atteinte	Sœur indemne	Témoin	Contrôle avec cancer du sein	Autres membres de la famille
Nombre d'échantillons	1400	618	446	1144	90	555
Minimum	14,52	14,09	13,89	11,68	24,54	13,29
25% Percentile	38,98	39,84	38,85	37,76	37,81	38,33
Médiane	50,64	52,98	51,09	49,18	46,30	51,70
75% Percentile	63,68	66,90	64,02	61,32	56,71	64,79
Maximum	243,6	167,1	203,3	157,3	111,3	235,0
Moyenne	52,89	55,33	53,19	51,27	47,84	54,15
Déviation standard	20,55	21,12	20,21	18,65	14,81	22,52
Erreur standard	0,5491	0,8494	0,9569	0,5514	1,561	0,9559
Minimum IC 95% Moyenne	51,81	53,66	51,31	50,18	44,73	52,27
Maximum IC 95% Moyenne	53,96	57,00	55,07	52,35	50,94	56,03
Coefficient de variation	38.85%	38.16%	37.99%	36.38%	30.96%	41.59%
Moyenne Géométrique	49,28	51,52	49,63	48,10	45,81	50,14
Minimum IC 95% Moyenne Géométrique	48,31	49,99	47,91	47,10	43,08	48,53
Maximum IC 95% Moyenne Géométrique	50,27	53,10	51,40	49,11	48,71	51,81

d'un prélèvement de sang (réalisé lors de l'inclusion dans l'étude GEMO et conservé à -20°C pendant un nombre variable d'années). Nous avons obtenu des résultats comparables avec les deux extractions d'ADN (test de corrélation de Pearson, r^2 = 0,7177, p-value < 0,0001, n = 24). Par ailleurs, les résultats obtenus par qPCR, quelles que soient la provenance de l'ADN et sa qualité, sont comparables avec ceux obtenus par peignage moléculaire (test de corrélation de Pearson ; pour l'ADN extrait à partir de LCLs : r^2 = 0,6150, p-value < 0,0001, n = 27 ; pour l'ADN extrait à partir de sang : r^2 = 0,6604, p-value < 0,0001, n = 24) (Figure 35). Ainsi, notre test qPCR donne une estimation fiable du NGC que porte chaque individu analysé, et peut être utilisé pour caractériser un grand nombre d'individus.

- **Etude du lien entre le NGC du CNV *RNU2* et la prédisposition génétique au cancer du sein par une étude cas/témoins**

Nous avons choisi de réaliser une étude cas/témoins en utilisant les individus de l'étude nationale GENESIS, mise en place et coordonnée par D. Stoppa-Lyonnet et N. Andrieu de l'Institut Curie et O. Sinilnikova de notre équipe. Les analyses ont été faites sur des échantillons d'ADN répartis sur 114 plaques de 96 puits. Afin de nous assurer que les résultats obtenus n'étaient pas dépendants de la plaque, le NGC a été calculé par rapport à un individu calibrateur dont le nombre de copies avait été déterminé par peignage moléculaire, analysé sur chacune des plaques (GE522). De plus, un second individu également caractérisé par peignage moléculaire a été analysé sur chacune des plaques.

Le NGC a été mesuré pour 1400 cas index, 1144 témoins, 618 sœurs atteintes, 446 sœurs indemnes et 555 autres membres de la famille des cas index. Une synthèse des résultats obtenus est présentée dans le tableau 9. Le NGC minimum est de 12, en accord avec les études précédentes montrant un nombre allélique minimum de 5 copies. De manière intéressante, on observe que le NGC maximum est différent pour les témoins (157) et les cas (244). Alors que pour les témoins, le NGC maximum est conforme à celui attendu, le NAC maximal identifié préalablement étant de 82, celui

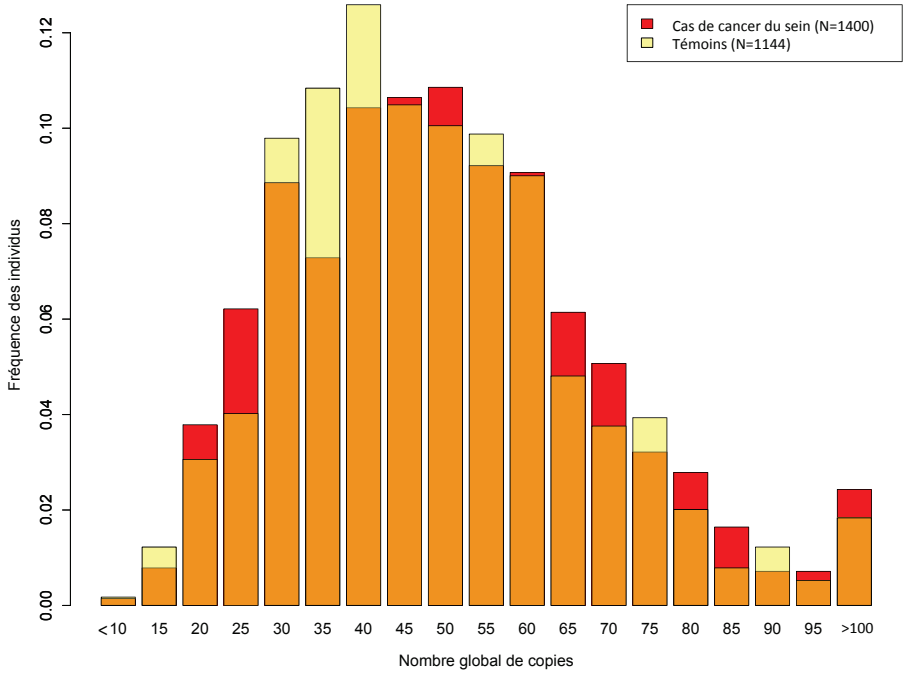


Figure 36 : Distribution du nombre global de copies du macrosatellite *RNU2* estimé par qPCR chez les cas de cancer du sein et les témoins de l'étude GENESIS.

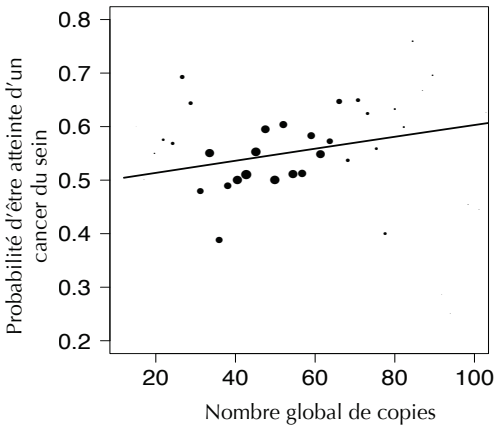


Figure 37 : Probabilité d'être atteinte d'un cancer du sein en fonction du nombre global de copies du macrosatellite *RNU2* d'après le modèle linéaire généralisé.

Tableau 10 : Probabilité d'être atteint par un cancer du sein en fonction du nombre global de copies du macrosatellite *RNU2* d'après le modèle linéaire généralisé.

Nombre global de copies	Probabilité d'être malade
20	0,52
40	0,54
60	0,56
80	0,58
100	0,6
120	0,62
140	0,64
160	0,66
180	0,68
200	0,69
220	0,71
240	0,73

mesuré chez les cas est largement supérieur. J'ai également observé que deux cas index présentent un NGC plus élevé que le témoin ayant le NGC le plus élevé. Par ailleurs, le nombre de cas index présentant un NGC supérieur à 100 est plus élevé que le nombre de témoins (34 vs 21) (Figure 36). La moyenne du NGC est légèrement mais significativement différente pour les cas (52,89) et pour les témoins (51,27) (Anova, p-value = 0,036), et cette différence ne peut pas être attribuée à un effet de plaque.

Pour analyser plus finement ces données et tester l'implication du NGC dans la probabilité de développer un cancer du sein, nous avons testé plusieurs modèles linéaires généralisés : modèle nul (AIC = 3502,9), modèle linéaire (AIC = 3500,7), modèle quadratique (AIC = 3501,9), modèle à seuil (AIC = 3502,3). Le modèle présentant l'AIC le plus faible est le modèle linéaire (voir Matériels et Méthodes). Nous pouvons conclure que la probabilité d'être atteint d'un cancer du sein augmente proportionnellement avec le NGC du CNV *RNU2*. Ainsi, dans notre étude, la probabilité d'être malade augmente de 10 % quand le NGC passe de 60 à 160 (Figure 37 & Tableau 10). Cette analyse statistique, bien que préliminaire, confirme notre hypothèse que le nombre de copies du CNV *RNU2* est associé au risque de cancer du sein.

- **Etude par FISH et par peignage moléculaire de la répartition allélique du NGC estimé par qPCR chez des cas index**

Pour préciser la répartition allélique du nombre de copies, nous avons caractérisé par peignage moléculaire 20 cas index sélectionnés selon les critères suivants : disponibilité dans le centre de ressources biologiques de GENESIS de lymphocytes B congelés, NGC élevé (> 100), NGC faible (< 20), NGC moyen (≈ 50), ou NGC correspondant au 50 % et 150 % de la moyenne (25 ou 75). Afin de nous assurer que les NGC élevés résultaient bien d'une amplification du nombre de copies d'un seul allèle, uniquement en 17q21, et non au sein d'un autre locus du génome humain, nous avons également réalisé des expériences de FISH pour les 10 cas index présentant un NGC supérieur à 100. Les résultats obtenus sont présentés dans le

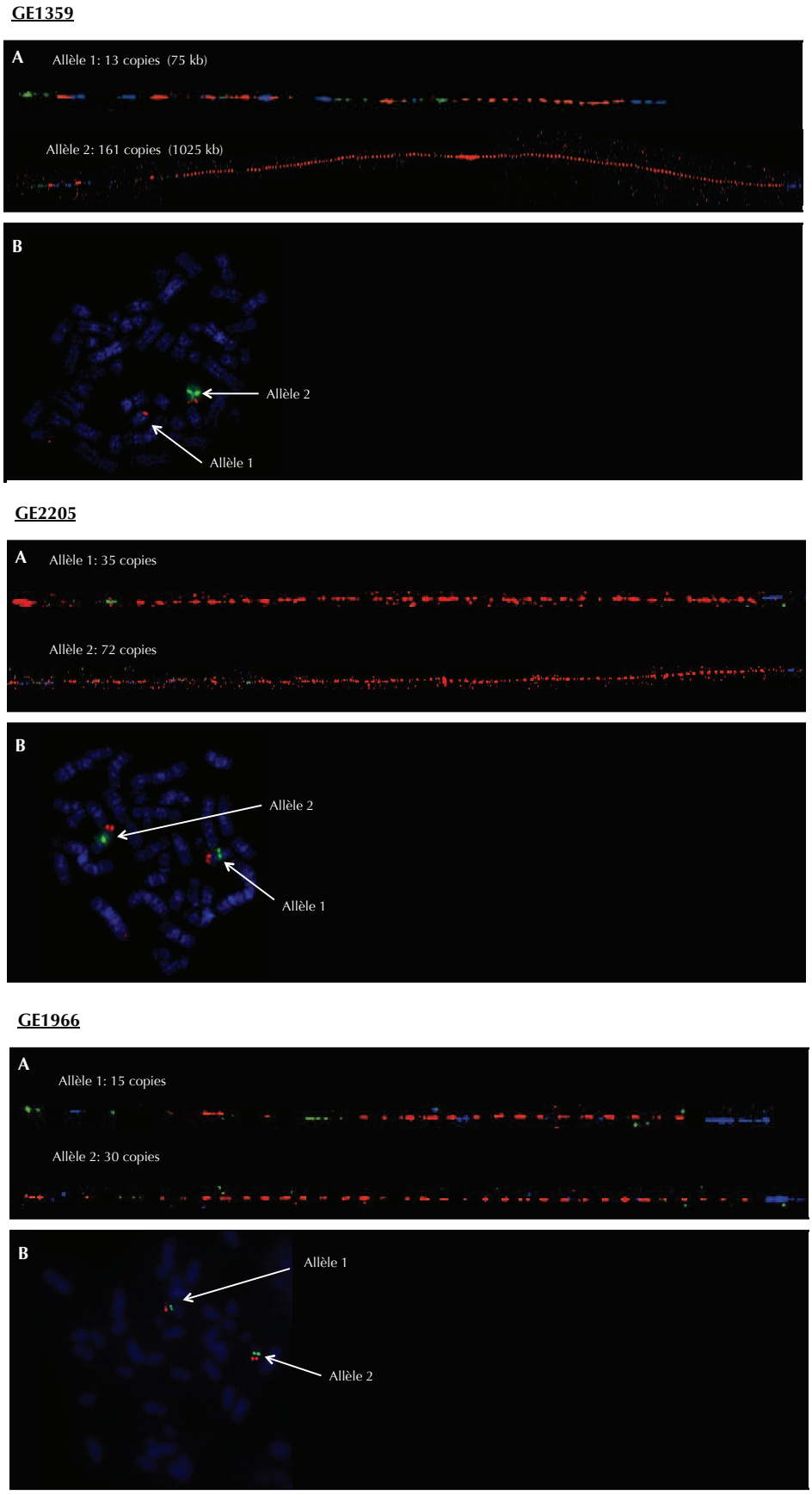
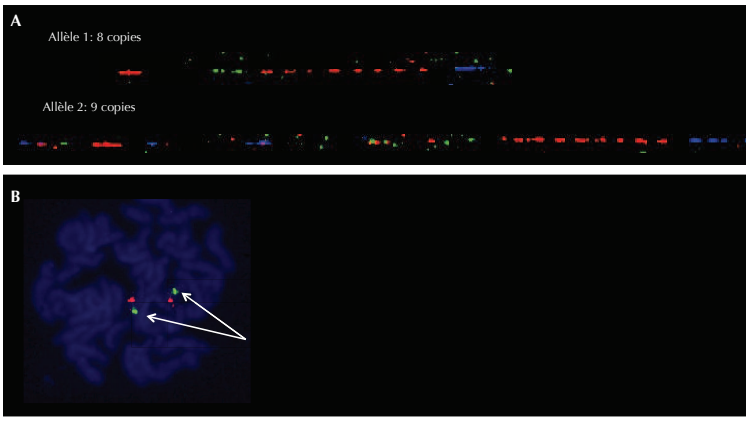
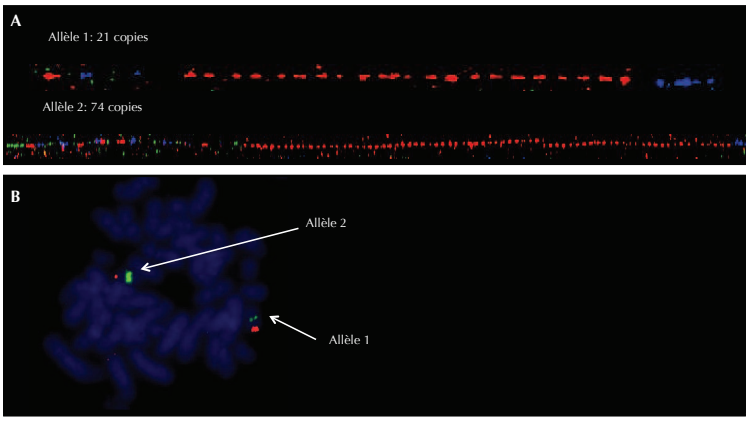


Figure 38 : Analyse par peignage moléculaire (**A.**) et FISH (**B.**) de la répartition allélique du nombre de copies dans 10 échantillons de l'étude GENESIS.

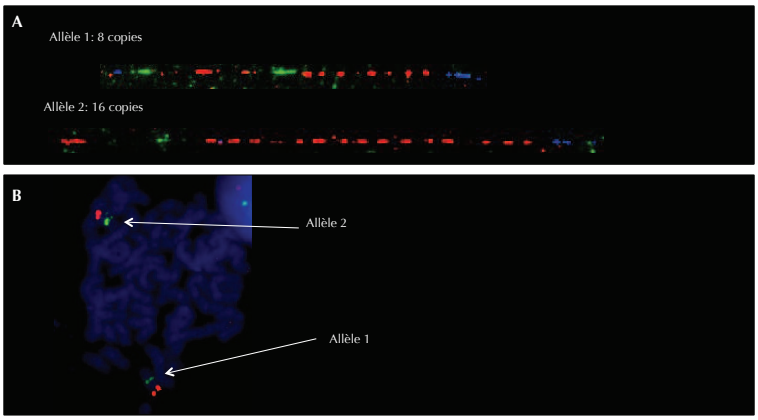
GE2154



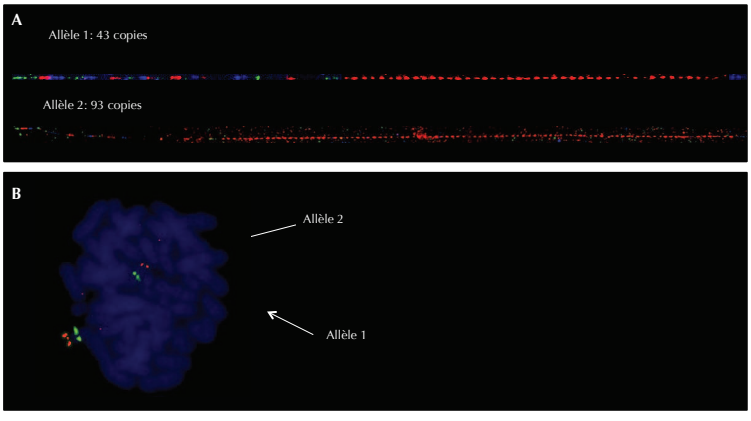
GE2543



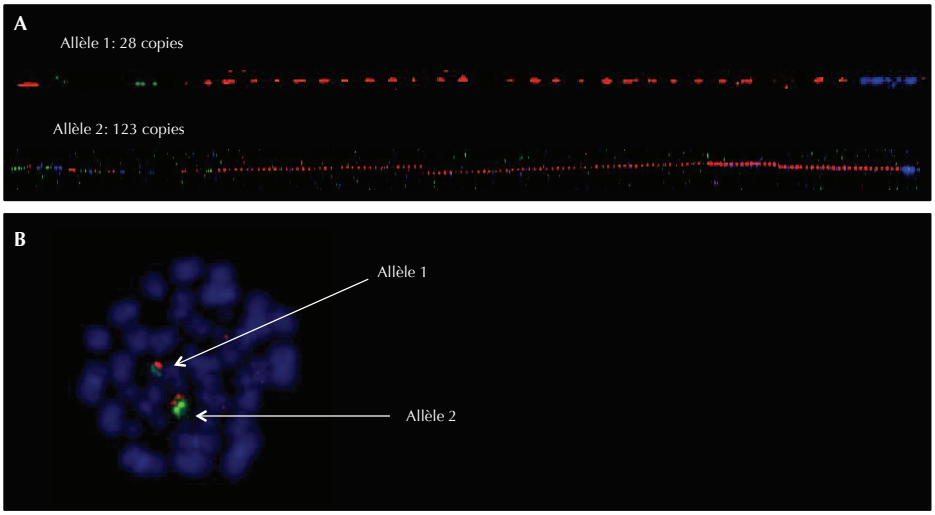
GE311



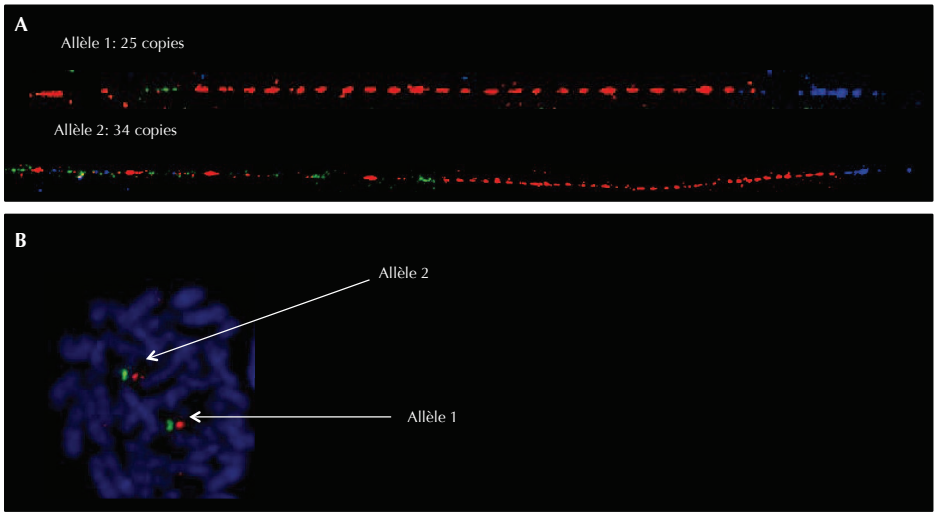
GE862



GE323



GE784



GE837

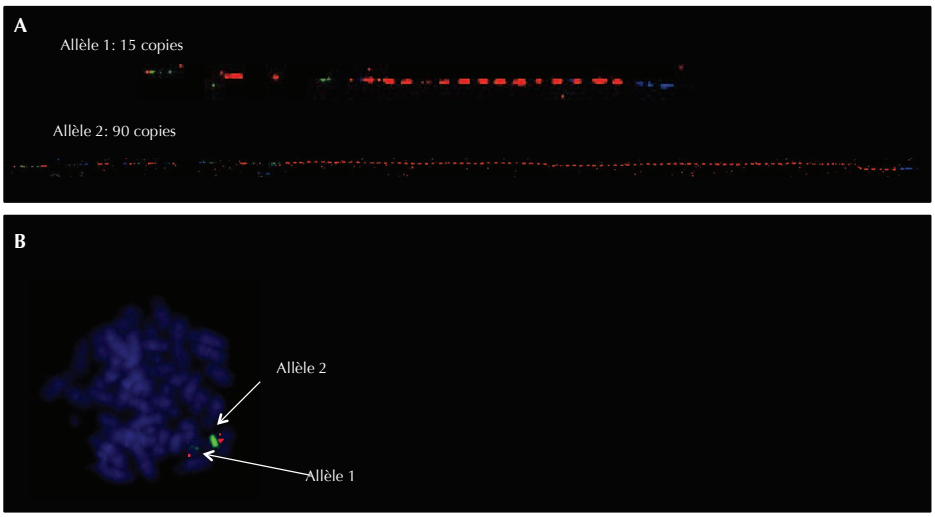


Tableau 11 : Détermination du nombre global de copies (qPCR) et du nombre allélique de copies (peignage moléculaire) pour 20 cas index de l'étude GENESIS.
Voir le texte des résultats pour les critères de sélection. L'allèle 1 a été arbitrairement défini comme étant celui comportant le plus petit nombre de copie.
ND: non déterminé

	Nombre global de copies estimé par qPCR	Nombre global de copies mesuré par peignage moléculaire	Nombre allélique mesuré par peignage moléculaire		Expression relative *BRCA1* (/*GAPDH*)	Déséquilibre d'expression allélique *BRCA1*
			Allèle 1	Allèle 2		
GE9	15	24	12	12	ND	ND
GE35	15	35	13	22	ND	ND
GE1966	16	45	15	30	2,1	0,89
GE2154	17	17	8	9	1,67	0,72
GE311	25	24	8	16	2,1	0,72
GE1510	75	72	24	48	ND	ND
GE784	75	59	25	34	1,69	0,86
GE2249	82	73	32	41	ND	ND
GE1884	90	75	13	62	ND	ND
GE479	100	77	20	57	ND	ND
GE3777	106	92	28	64	ND	ND
GE2543	109	95	21	74	1,74	ND
GE1815	111	94	32	62	ND	ND
GE3572	118	100	7	93	ND	ND
GE1622	122	114	35	80	1,22	0,65
GE837	122	105	15	90	0,67	ND
GE323	130	151	28	123	1,59	0,65
GE2205	134	107	35	72	2,06	ND
GE862	148	136	43	93	3	ND
GE1359	244	174	13	161	1,7	0,89

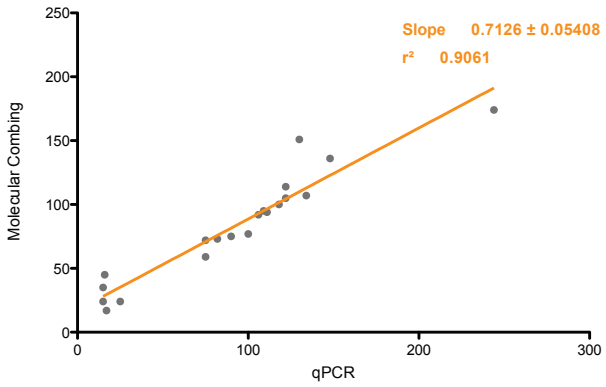


Figure 39 : Corrélation entre le nombre global de copies estimé par qPCR et déterminé après mesure des nombres alléliques de copies par peignage moléculaire pour 20 individus de l'étude GENESIS.

Tableau 11, ainsi que dans la Figure 38. Nous observons une très bonne corrélation entre le NGC estimé par qPCR et le NGC calculé après détermination du NAC par peignage moléculaire (r^2 = 0,9061, p-value < 0,0001) (Figure 39). Par ailleurs, nous avons pu mettre en évidence la présence systématique de seulement 2 signaux par FISH, et pu noter que l'intensité du signal obtenu pour chacun des allèles est un bon indicateur du NAC.

Le NAC minimum pour ces 20 échantillons est de 8. Quatre allèles présentent une taille supérieure aux allèles que nous avions préalablement décrits : ils sont porteurs respectivement de 90, 93, 123 et 161 copies. Ce dernier allèle, qui correspond à plus d'1 Mb de répétitions, a été identifié chez le cas index présentant le NGC le plus élevé en qPCR (GE1359) ; son second allèle contient 13 copies (Figure 38). Comme cela pouvait être attendu, nous pouvons noter que les cas index que nous avons peigné et présentant un NGC supérieur à deux fois la moyenne (104) sont tous porteurs d'un allèle de grande taille (> 60), et non de deux allèles de taille *moyenne*.

- **Etude par peignage moléculaire de la répartition allélique du NGC estimé par qPCR chez des témoins à partir de sang congelé**

Afin de déterminer si les témoins ayant un NGC élevé portent également des allèles de grande taille, nous avons entrepris de caractériser les NAC dans ces échantillons à partir de sang congelé, car c'est le seul matériel à notre disposition (les lymphocytes B n'ont été congelés que pour les cas index dans l'étude GENESIS). Nous avons dans un premier temps essayé d'utiliser l'intensité des signaux obtenus par FISH en tant qu'indicateur du NAC. Les essais réalisés n'ont pas donné de résultats concluants, les cellules étant pour la plupart éclatées et les chromosomes abimés.

Un nouveau protocole d'extraction d'ADN a été utilisé afin de tester le peignage moléculaire (voir Matériels et Méthodes). L'inconvénient de cette extraction est que la plupart des fibres obtenues présente une longueur maximale de 200 kb, ce qui est insuffisant pour notre recherche de grands allèles. Dans un premier temps, nous avons entrepris la caractérisation d'un cas index, déjà caractérisé par peignage moléculaire à

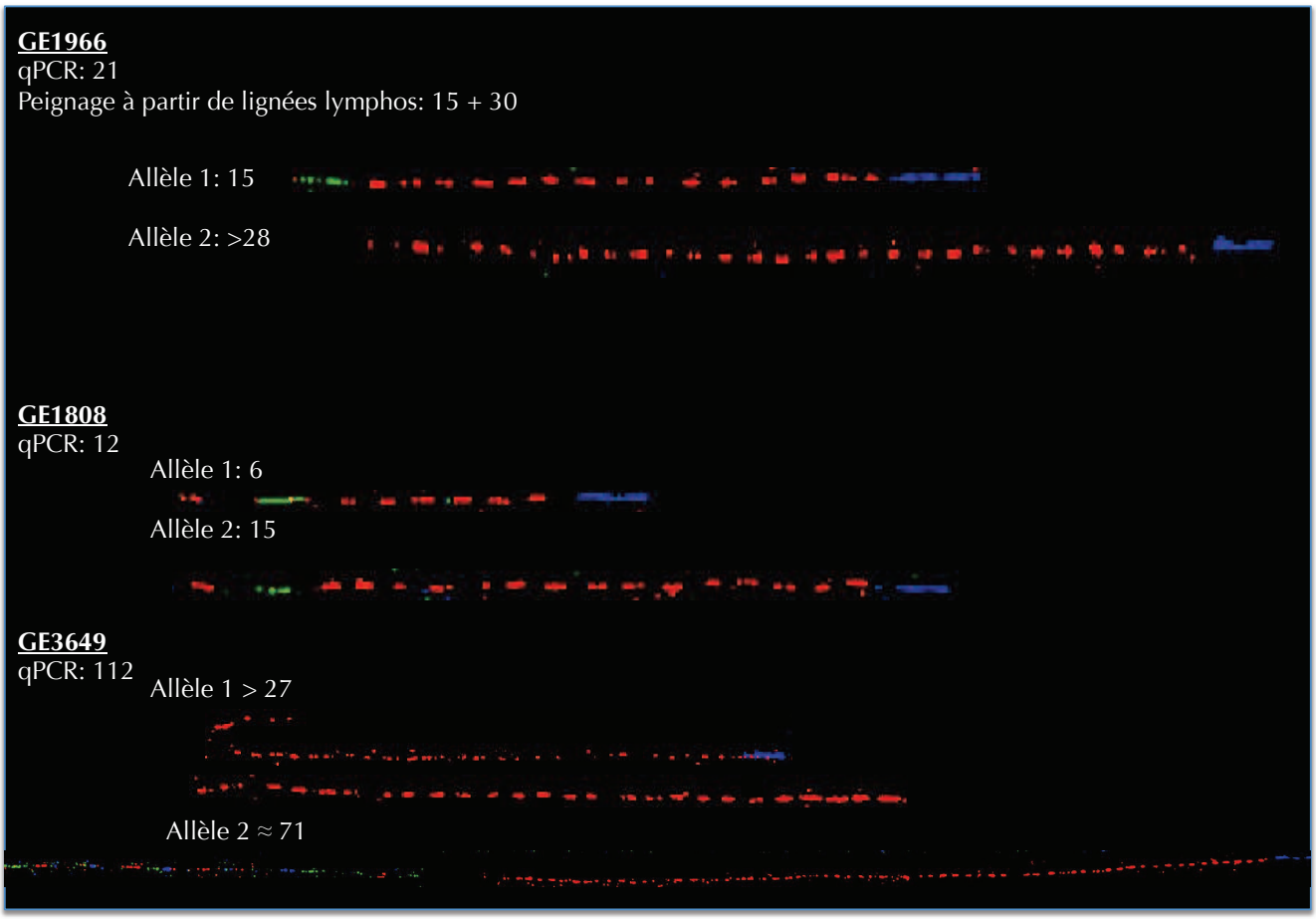


Figure 40 : Détermination du nombre allélique de copies par peignage moléculaire à partir de sang congelés de deux témoins et d'un cas index de l'étude GENESIS.
Pour chaque individu, les signaux intacts les plus longs sont présentés.

partir de lignées cellulaires (Allèle 1 : 15 copies, Allèle 2 : 30 copies). A partir de sang congelé, nous avons observé des signaux entiers portant 15 répétitions, et un signal incomplet portant 28 répétitions. Par ailleurs, nous avons analysé un témoin avec un NGC faible (GE1808, 12 en qPCR), pour lequel nous avons visualisé des signaux complets avec 6 répétitions, et d'autres avec 15 répétitions (Figure 40). Ce désaccord entre les deux techniques avait déjà été observé pour les cas index avec un NGC < 20, et provient d'une imprécision de la qPCR qui sous-estime le NGC pour les faibles valeurs. Pour finir, un témoin avec un NGC de 112 en qPCR (GE3649) a été également analysé. Nous avons identifié un signal complet portant environ 71 répétitions (l'hybridation étant imparfaite, il n'est pas possible de compter précisément le nombre de copies), et des signaux incomplets portant jusqu'à 27 répétitions.

Ce protocole utilisant du sang congelé donne donc des résultats très encourageants. Il permet de caractériser des allèles de petite taille, et donne des résultats concordants avec ceux obtenus à partir de lignées (voir résultats pour GE1966). Cependant, les allèles complets de grande taille sont extrêmement difficiles à obtenir. Pour le moment, nous pouvons inférer la taille minimum de chaque allèle en fonction du signal complet le plus long obtenu. Ainsi, pour le témoin GE3649, la somme des NACs obtenus par peignage à partir de sang congelés est égale à 98, ce qui paraît cohérent avec le résultat de qPCR (112). Nous sommes en train de caractériser un plus grand nombre de témoins par cette technique afin de déterminer si la répartition allélique du nombre de copies est la même que chez les cas présentant un NGC élevé.

- **Co-ségrégation d'un grand nombre de copies du CNV *RNU2* et du cancer du sein dans des familles de cancers du sein.**

Nous avons étudié la co-ségrégation du NGC avec le cancer du sein dans des familles remplissant les critères suivants, l'analyse de co-ségrégation dans les autres familles ne permettant pas de tirer de conclusions :

- au moins un individu avec un NGC > 100,

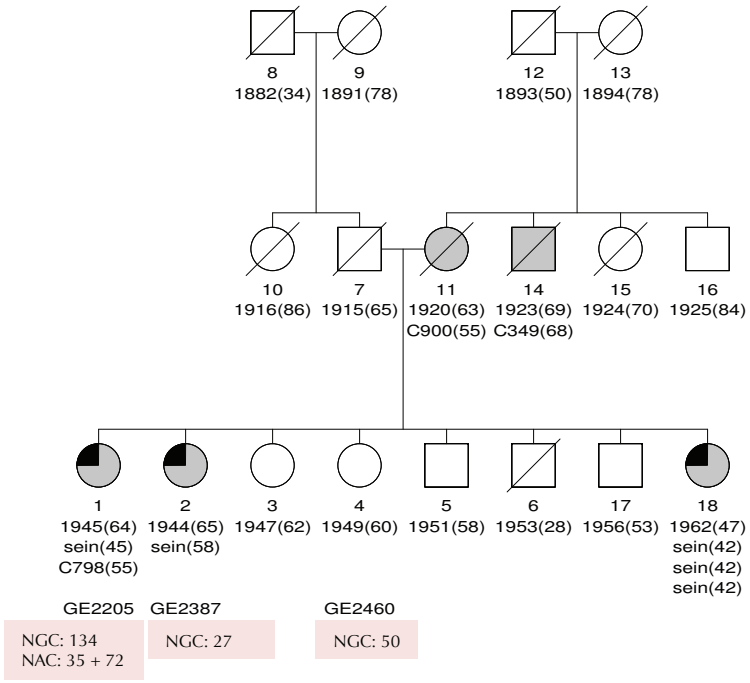


Figure 41 : Arbre généalogique de la famille du cas index GE2205.
C349: tumeur maligne des bronches ou du poumon; C798: tumeurs malignes d'autres sièges; C900: myélome multiple.
sein(45): âge de diagnostic du cancer
1945(64): date de naissance et âge au décès ou au dernier recensement
NGC: nombre global de copies (qPCR)
NAC: nombre allélique de copies (peignage moléculaire)

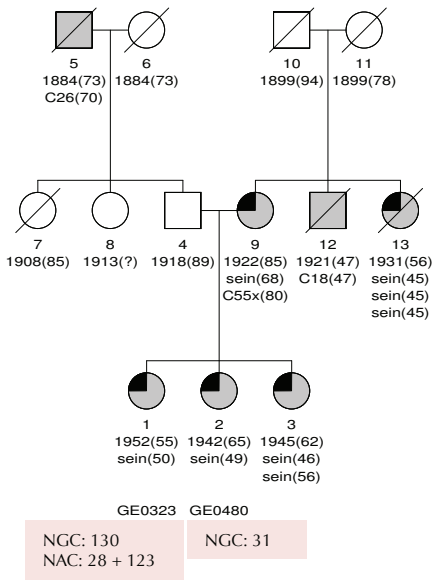


Figure 42 : Arbre généalogique de la famille du cas index GE323.
C26: tumeur maligne des organes digestifs, de sièges autres et mal définis ; C55x: tumeur maligne de l'utérus; C18: tumeur maligne du côlon;
sein(36): âge de diagnostic du cancer
1958(52): date de naissance et âge au décès ou au dernier recensement
NGC: nombre global de copies (qPCR)
NAC: nombre allélique de copies (peignage moléculaire)

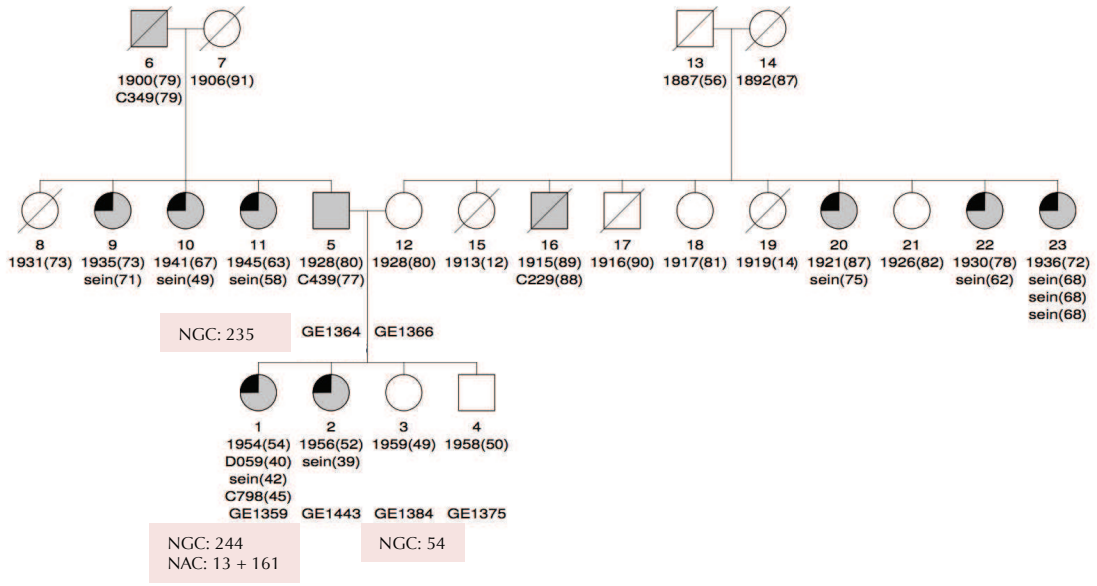


Figure 43 : Arbre généalogique de la famille du cas index GE1359.
D059: carcinome in situ du sein
C798: tumeurs malignes d'autres sièges; C349: tumeur maligne des bronches ou du poumon ; C229: tumeur maligne du foie; C439: tumeur maligne de la peau
sein(42): âge de diagnostic du cancer
1954(54): date de naissance et âge au décès ou au dernier recensement
NGC: nombre global de copies (qPCR)
NAC: nombre allélique de copies (peignage moléculaire)

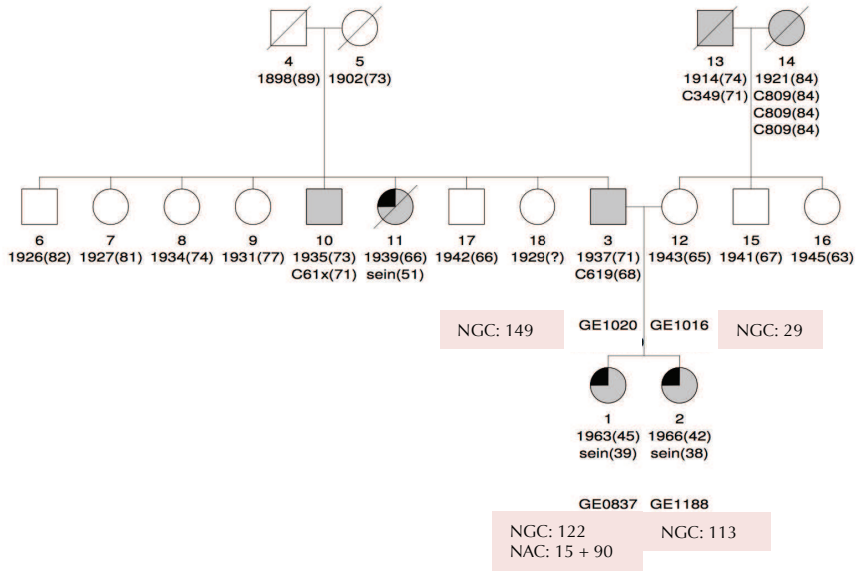


Figure 44 : Arbre généalogique de la famille du cas index GE837.
C619 / C61x: tumeur maligne de la prostate; C349: tumeur maligne des bronches ou du poumon , C809: tumeur maligne de siège non précisé
sein(39): âge de diagnostic du cancer
1963(45): date de naissance et âge au décès ou au dernier recensement
NGC: nombre global de copies (qPCR)
NAC: nombre allélique de copies (peignage moléculaire)

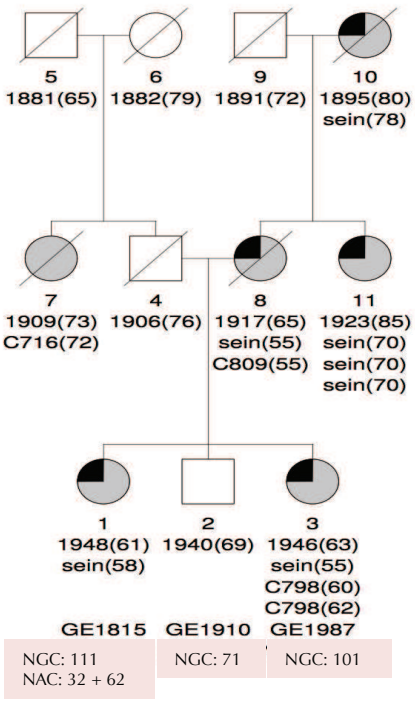


Figure 45 : Arbre généalogique de la famille du cas index GE1815.
C716: tumeur maligne du cervelet; C798: tumeurs malignes d'autres sièges; C809: tumeur maligne de siège non précisé
sein(58): âge de diagnostic du cancer
1948(61): date de naissance et âge au décès ou au dernier recensement
NGC: nombre global de copies (qPCR)
NAC: nombre allélique de copies (peignage moléculaire)

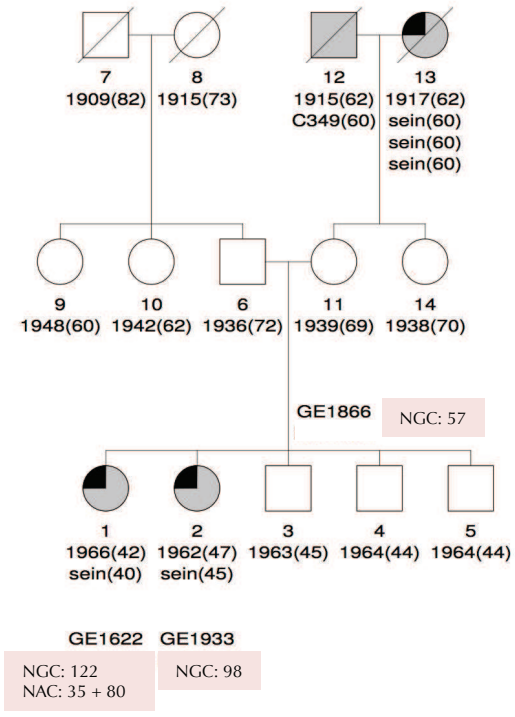


Figure 46 : Arbre généalogique de la famille du cas index GE1622.
C349: tumeur maligne des bronches ou du poumon
sein(40): âge de diagnostic du cancer
1966(42): date de naissance et âge au décès ou au dernier recensement
NGC: nombre global de copies (qPCR)
NAC: nombre allélique de copies (peignage moléculaire)

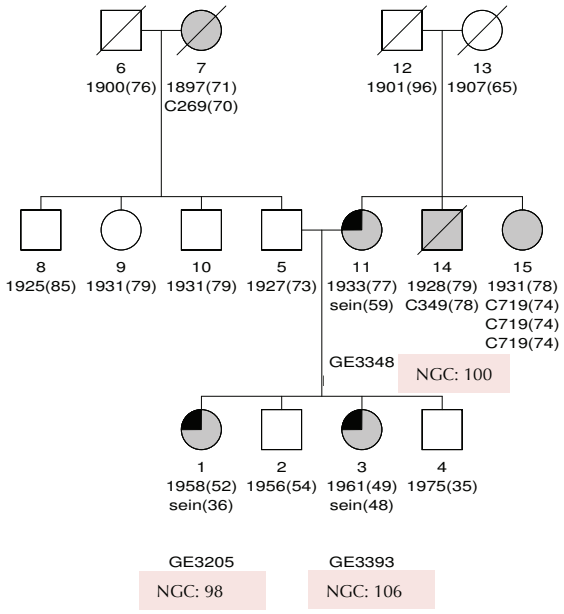


Figure 47 : Arbre généalogique de la famille du cas index GE3205.
C269: tumeur maligne de sièges mal définis de l'appareil digestif; C349: tumeur maligne des bronches ou du poumon; C719: tumeur maligne du cerveau
sein(36): âge de diagnostic du cancer
1958(52): date de naissance et âge au décès ou au dernier recensement
NGC: nombre global de copies (qPCR)
NAC: nombre allélique de copies (peignage moléculaire)

Tableau 12 : Caractéristiques et histoire familiale des contrôles présentant un nombre global de copies élevé en qPCR (<100) provenant de l'étude GENESIS.
Lignes grisées : témoin avec une histoire familiale de cancer du sein et/ou plus jeune que 50 ans.

Nombre global de copies	Âge (dernières informations)	Histoire familiale de cancer du sein / ovaires	Histoire familiale de cancer
157	57	-	5 individus atteints de côté maternelle (prostate, colon, cerveau, …)
154	59	Mère: sein (70 ans)	-
148	59	-	Tante paternelle: utérus (55 ans)
133	65	-	Sœur: de siège mal défini (29 ans) - Mère: vessie (79 ans)
131	66	Tante paternelle: sein (?) - Mère: sein (88 ans)	-
126	61	Grand-mère maternelle: sein (35 ans)	Mère: colon (64 ans)
119	65	-	Père: prostate (60 ans)
116	57	-	-
112	64	Père: sein (80 ans) - Mère: ovaire (50 ans) - Tante maternelle: sein (75 ans) - Grand-mère paternelle: sein (?)	-
112	44	-	Tante: peau - Grands-parents maternels: cerveau (69 ans)
108	52	-	-
107	46	Tante maternelle: sein (<50 ans)	Oncle maternel: estomac
107	68	-	-
103	44	-	Grand-père paternel: poumon - Père: lèvres/bouche - Tante maternelle: lèvres/bouche
102	39	-	Grand-mère maternelle: estomac + non-précisé - Grand-père paternel: pancréas
102	70	Sœur: sein (61 ans)	Grand-mère maternelle: utérus (68 ans) - Tante maternelle : prostate
102	34	Grand-mère paternelle: sein	Tante: plèvre
101	58	-	Père: foie et prostate
101	59	-	Mère: appareil digestif - Tante: tissus lymphos
101	60	-	Grand-mère paternelle: estomac - Oncle paternel: prostate

- au moins deux apparentés analysés en qPCR : soit un cas de cancer du sein et un parent (père ou mère), soit un cas de cancer du sein analysé par peignage moléculaire et un apparenté.

Nous avons identifié 7 familles remplissant ces critères. Pour deux familles, l'histoire familiale et les données de génotypage du CNV ne sont pas compatibles avec une co-ségrégation d'un NGC élevé avec le cancer du sein (Figures 41 & 42) mais elles le sont pour 5 autres familles (Figures 43 - 47). C'est le cas pour le cas index (GE1359) présentant le NGC le plus élevé (244) : étant donné qu'il possède lui aussi un NGC très élevé, l'allèle portant 161 copies a été vraisemblablement hérité du père, qui a trois sœurs atteintes d'un cancer du sein (49 ans, 58 ans, 71 ans) (Figure 43). Ce cas index a été atteint d'un carcinome *in situ* du sein à 40 ans, d'un cancer du sein à 42 ans, et enfin d'une tumeur maligne à d'autres sièges à 45 ans. Une de ses sœurs, non atteinte de cancer du sein, est porteuse d'un NGC moyen (54), et n'a donc vraisemblablement pas hérité de l'allèle avec 161 copies du père. Malheureusement, il n'a pas été possible d'analyser la sœur atteinte d'un cancer du sein. Nous observons également une co-ségrégation dans la famille du cas index GE837 : le père, dont une sœur a été atteinte d'un cancer du sein à 51 ans, a transmis un allèle avec 90 copies à ses deux filles atteintes d'un cancer du sein (38 et 39 ans) (Figure 44). Le cas index GE1815 (NGC 111) a été atteint d'un cancer du sein à 58 ans, et sa sœur, également porteuse d'un NGC élevé (101) à 55 ans (Figure 45). De même pour le cas index GE1622 (NGC 122, cancer du sein à 40 ans) et sa sœur (NGC 98, cancer du sein à 45 ans) (Figure 46); le cas index GE3205 (NGC 98, cancer du sein à 36 ans), sa sœur (NGC 106, cancer du sein à 48 ans) et sa mère (NGC 100, cancer du sein à 59 ans) (Figure 47).

J'ai également réalisé la même analyse pour les témoins présentant un NGC élevé (> 100) (Tableau 12). Parmi ces 20 familles, cinq présentent des antécédents de cancer du sein. Un témoin avec un NGC de 112 présente une histoire familiale de cancer particulièrement forte : sa mère a développé un cancer de l'ovaire à 50 ans, sa tante maternelle un cancer du sein à 75 ans, sa grand-mère paternelle un cancer du sein, et de manière très intéressante son père a également développé un cancer du sein à 80 ans, ce qui est particulièrement rare. On peut également noter que 4 témoins sont encore jeunes (< 45 ans).

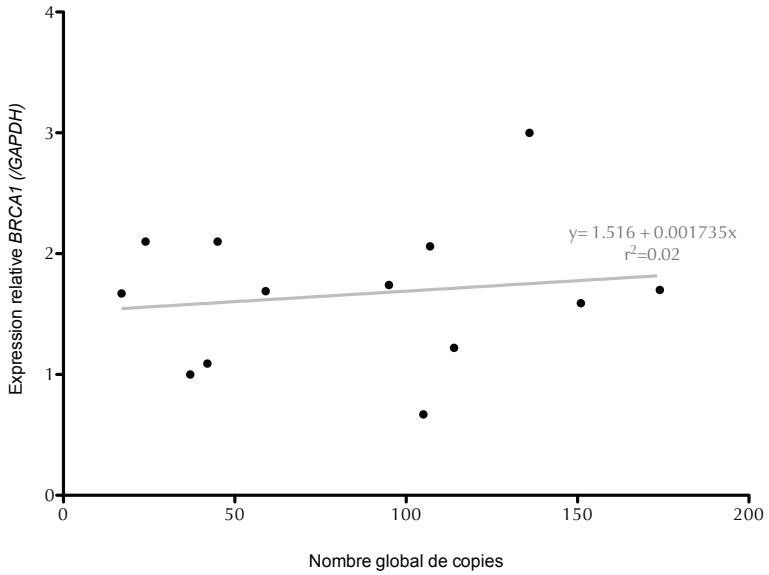


Figure 48 : Expression relative de *BRCA1*, par rapport à la *GAPDH*, mesurée par RT-qPCR chez 13 cas index provenant de l'étude GENESIS.

- **Etude de l'expression de *BRCA1* dans les lignées lymphoblastoïdes des cas index porteurs d'un NGC très élevé**

Les résultats obtenus sur un large nombre d'individus atteints de cancer du sein et témoins appariés ont mis en évidence une possible influence du nombre de copies du CNV *RNU2* sur la probabilité de développer un cancer du sein. Puisqu'il avait été montré que la présence d'un CNV peut modifier l'expression des gènes situés dans son voisinage, nous avons voulu tester l'effet de la variation du nombre de copies du CNV *RNU2* sur le niveau d'expression de *BRCA1*.

Pour cela, nous avons évalué le niveau relatif d'expression de *BRCA1* par RT-qPCR pour 13 lignées lymphoblastoïdes, établies à partir des lymphocytes B de 13 cas index porteurs d'un NGC variable (17-174 en peignage moléculaire), en comparant avec un gène de ménage , la *GAPDH*. Nous avons observé des variations d'un facteur 4 dans le taux de transcrits *BRCA1* (Minimum : 0,67, Maximum : 3,00), conformément à ce qui avait été décrit (Ribieras et al., 1997). Par contre cette expression n'est pas corrélée avec le nombre de copies du CNV *RNU2* (Figure 48 & Tableau 11). Pour affiner notre analyse, nous avons mesuré le niveau d'expression allélique de *BRCA1*. Nous avons pour cela sélectionné, parmi ceux étudiés précédemment, les individus hétérozygotes pour au moins un SNP exonique de *BRCA1*, afin de pouvoir différencier les deux allèles. Nous avons ensuite déterminé pour ces 7 individus le niveau d'expression de chacun des allèles de *BRCA1* par RT-qPCR (Tableau 11) en utilisant des sondes capables de reconnaitre spécifiquement l'un ou l'autre des deux allèles du SNP rs1799966. Pour chaque individu, nous avons observé un déséquilibre d'expression entre les deux allèles, *i.e.* un allèle est légèrement plus exprimé que l'autre. Cependant, le niveau de ce déséquilibre d'expression allélique semble indépendant de la présence ou non d'un NAC élevé du CNV *RNU2*.

Pour confirmer ces observations, nous avons étudié plus précisément l'individu porteur du plus grand NAC (GE1359) par séquençage, grâce à la présence d'un SNP présent à l'état hétérozygote (rs1799966) que nous avons identifié au sein de la séquence codante de *BRCA1*. Dans un premier temps, nous avons génotypé ce SNP

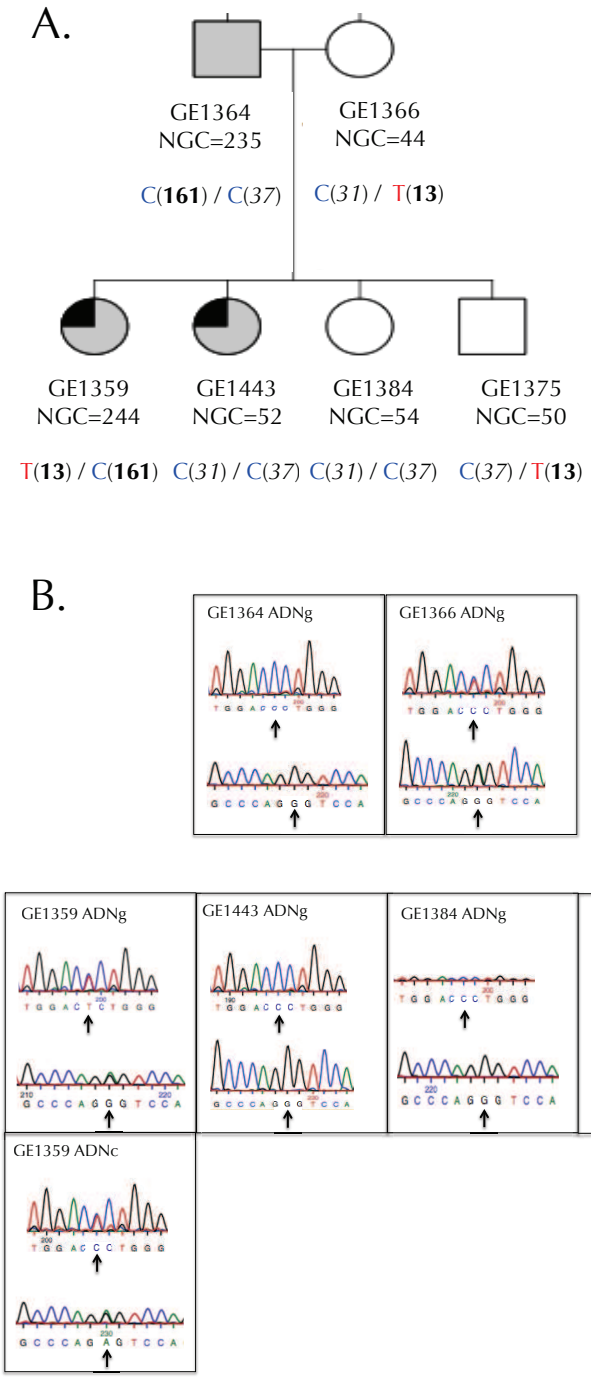


Figure 49 : Haplotype des individus de la famille du cas index GE1359.
A. Arbre généalogique simplifié. C/T: génotype pour le SNP rs1799966, **13**: nombre allélique de copies mesuré par peignage moléculaire, 31: nombre allélique de copies estimé par calcul.
B. Résultats de séquençage du SNP rs1799966 sur l'ADN des membres de la famille et sur l'ADNc du cas index GE1359.

chez plusieurs membres de la famille de cet individu. Cela nous a permis d'associer les deux allèles de ce SNP avec un nombre de copies *réel* (mesuré par peignage moléculaire) ou *estimé* (à partir des résultats de qPCR) et de confirmer la transmission mendélienne du nombre de copies dans cette famille (Figure 49). Nous avons également séquencé l'ADNc pour l'individu GE1359, et nous n'avons pas vu de différence d'intensité entre les deux allèles pour ce SNP. Cela signifie que les deux allèles de *BRCA1* sont exprimés de manière quasi-identique, confirmant l'absence d'un effet majeur du nombre élevé de copies du CNV *RNU2* sur la transcription de *BRCA1*.

- **Confirmation de l'effet du nombre global de copies sur la prédisposition au cancer du sein dans une seconde étude cas/témoins**

Afin de confirmer ou d'infirmer les résultats que nous avons obtenus avec les individus de l'étude GENESIS, nous avons répliqué cette analyse sur une seconde étude, BCFR, incluant des cas de cancers du sein diagnostiqués à un âge jeune (avant 45 ans) et des témoins. Nous avons de nouveau estimé le nombre global de copies du CNV *RNU2* par qPCR, en utilisant le même protocole et le même individu calibrateur que pour l'étude GENESIS, pour 1017 témoins et 1149 cas. Nous avons obtenu un nombre moyen de copies de 58,04 (IC : 56,41 – 59,67) pour les témoins, avec un minimum de 11 et un maximum de 260. Pour les cas, la moyenne est de 60,92 (IC : 59,38 – 62,46) avec un minimum de 10 et un maximum de 203.

Comme pour l'étude GENESIS, j'ai vérifié les valeurs obtenues pour les individus présentant un NGC faible ou élevé (≈ 5 %), donc facilement distinguables, en analysant de nouveau l'échantillon (en duplicat) à partir du prélèvement initial. Les résultats obtenus étaient cohérents pour la plupart des échantillons, sauf pour certains provenant tous de deux plaques de qPCR. N'ayant pas pu identifier l'origine de ce désaccord, nous avons décidé d'omettre l'ensemble des mesures provenant de ces deux plaques pour l'analyse, dans l'attente de la résolution du problème.

Nous avons donc conservé les résultats pour 744 témoins (Moyenne : 59,71, IC : 57,76-61,65, Min : 11, Max : 244) et 914 cas (Moyenne : 62,75, IC : 61,01-64,50,

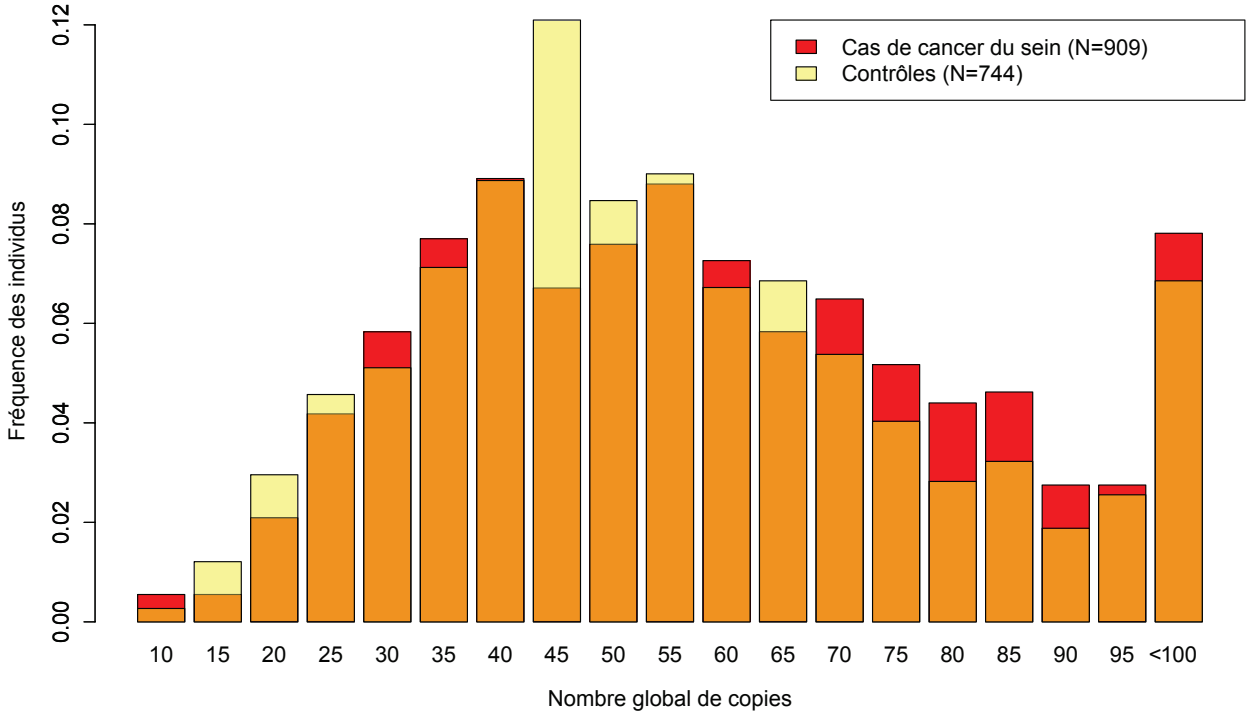


Figure 50 : Distribution du nombre global de copies du macrosatellite *RNU2* pour l'étude BCFR, estimé par qPCR chez 909 cas de cancer du sein et 744 témoins.

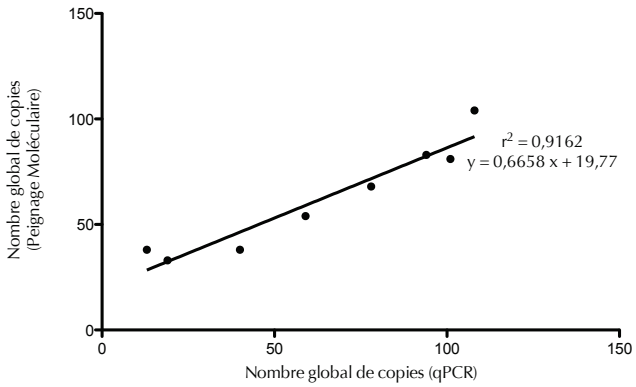


Figure 51 : Corrélation entre le nombre global de copies estimé par qPCR et mesuré par peignage moléculaire pour 8 individus de l'étude BCFR.

Min : 10, Max : 202) (Figure 50). Dans un second temps, nous avons également vérifié l'exactitude de ces mesures en peignant 8 individus présentant des nombres de copies variés et pour lesquels une lignée lymphoblastoïde était disponible (17 – 127 en qPCR) (Figure 51). Nous avons obtenu une très bonne corrélation entre les deux techniques (r^2= 0,9162, p-value = 0,0002).

Dans cette étude également, nous avons vu que le nombre global de copies est significativement différent entre les cas et les témoins (Anova à 2 facteurs, p-value = 0,037), après correction pour l'effet de la plaque. En combinant ces deux études, on obtient un effet plus marqué, après ajustement pour l'étude d'origine (Anova à 2 facteurs, p-value = 0,0045). La comparaison des valeurs médianes du nombre global de copies pour ces deux études révèle que les cas portent 2,5 copies de plus (53,29 vs 50,78).

3.4 Discussion

La localisation précise du CNV *RNU2* a confirmé sa proximité avec le gène *BRCA1*, un gène majeur de prédisposition au cancer du sein. Cependant, du fait de son absence du génome de référence, le lien potentiel du nombre de copies du CNV *RNU2* avec la prédisposition génétique au cancer du sein n'a jamais été investigué dans des études à large échelle. J'ai mis au point un test basé sur la qPCR permettant d'évaluer le NGC du locus, que j'ai validé en conduisant plusieurs expériences de validation. J'ai ensuite génotypé le CNV dans deux études de cas de cancer du sein et témoins appariés : l'étude GENESIS, qui est une étude de paires de sœurs atteintes d'un cancer du sein et non porteuses de mutation sur *BRCA1* et *BRCA2*, et l'étude BCFR, regroupant des cas précoces de cancer du sein (âge au diagnostic < 45 ans). Au total, nous avons pu mesurer le NGC pour 2070 cas et 1758 témoins. Cette mesure est imprécise, surtout pour des NGC faibles (< 20) ou élevés (> 100), comme nous l'avons vu après peignage de ces individus. Cependant, on peut remarquer que pour ces deux études, le NGC minimum est d'environ 10, ce qui est en accord avec les données sur le nombre allélique minimum (5). De plus, l'imprécision de la méthode n'apparait pas être un handicap pour des études cas-témoins. Pour chacune des études, nous observons une augmentation de la probabilité d'être atteint d'un cancer du sein proportionnelle au NGC. En couplant les deux études, et bien que les cas soient d'origine différente et n'aient pas été sélectionnés sur les mêmes critères (voir Matériels & Méthodes), l'effet du NGC sur la prédisposition génétique au cancer du sein est renforcé.

Il sera nécessaire de quantifier plus précisément le risque associé aux différents allèles du CNV *RNU2*. Malheureusement, l'analyse statistique de nos données se heurte à deux problèmes : le fait que le CNV possède un très grand nombre d'allèles (66 allèles différents ont été caractérisés au jour d'aujourd'hui) alors que ce sont d'habitude des polymorphismes bi-alléliques qui sont étudiés, et le fait que notre test ne permette actuellement que de mesurer le NGC. Malgré nos tentatives (données non montrées), nous n'avons pas pu mettre au point un test à haut débit permettant de

mesurer les NAC, ce qui complique considérablement l'analyse des données. En effet, nous ne pouvons pas prendre en compte uniquement la taille de l'allèle le plus grand pour notre étude statistique, ce qui engendre une perte de pouvoir statistique. Suivant l'hypothèse d'un effet, qu'il soit dominant ou récessif, d'un NAC élevé à partir d'un certain seuil, tel que cela a été montré pour les maladies à triplets, il est crucial de connaître la répartition allélique d'un NGC élevé. Tant que nous n'aurons pas déterminé cette valeur seuil, l'interprétation des données familiales restera complexe, et aucune conclusion sur la pénétrance d'un NGC élevé ne pourra être apportée. Dans le but de déterminer cette valeur seuil, nous réfléchissons actuellement à l'outil statistique le plus adapté avec l'aide de l'épidémiologiste co-coordinatrice de l'étude, N. Andrieu et de F. Lesueur, chercheur dans son équipe. Une hypothèse que nous envisageons est de regrouper les valeurs de NGC obtenues dans des groupes, afin de diminuer le nombre d'allèles. Dans ce cas, on pourrait imaginer pouvoir identifier des *tag*-SNPs qui pourraient être utilisés pour repérer les porteuses de groupes de nombres élevés de copies. Pour déterminer si cela est possible, nous allons utiliser les données de génotypage de la puce iCOGS. La puce iCOGS est une puce Infinium iSelect HD Custom Genotyping BeadChip (Illumina) comprenant 204000 SNPs qui ont été choisis en raison de leur localisation dans des régions identifiées lors d'études pan-génomiques et dans des gènes candidats afin d'identifier, entre autres, des facteurs génétique de risque de cancer du sein (Bahcall, 2013). La totalité des échantillons de l'étude GENESIS a été génotypée avec cette puce, qui comprend 117 SNPs du bloc de déséquilibre de liaison de *BRCA1*. Les résultats vont être disponibles très prochainement et vont être analysés à cette fin.

Bien que nous suspections fortement un effet du nombre de copies sur la prédisposition génétique au cancer du sein, nous n'avons pas vu d'influence du nombre de copies du CNV *RNU2* sur l'expression de *BRCA1* dans la lignée lymphoblastoïde de la porteuse du plus grand NGC identifié. Pour autant, ces résultats négatifs n'écartent pas totalement cette hypothèse car ces lignées ont été immortalisées, ce qui a pu modifier le contexte chromatinien de la région *BRCA1 – RNU2*. Par ailleurs, il est également envisageable que l'effet du nombre de copies soit tissu-spécifique. Il

faudrait pour vérifier cette hypothèse étudier différentes lignées cellulaires, et particulièrement des lignées cellulaires mammaires, la difficulté étant que les grands nombres de copies sont extrêmement rares. Grâce à une collaboration avec l'anatomopathologiste M.-E. Fondrevelle (Centre Léon Bérard), nous allons continuer cette analyse dans les tissus mammaires sains et tumoraux des femmes porteuses d'un grand nombre de copies provenant de l'étude GENESIS, en fonction de la disponibilité du matériel, à partir des tumeurs incluses en bloc de paraffine. Grâce au Centre Coordinateur de l'équipe GENESIS qui a recensé les différents hôpitaux dans lesquels ont été opérés les cas index et où sont stockés les blocs de tumeurs, nous avons d'ores-et-déjà pu en récupérer une vingtaine. A partir de ces blocs, nous allons extraire l'ARN et l'ADN et regarder le nombre de copies du CNV dans le tissu sain et tumoral par FISH, puis le niveau d'expression de *BRCA1* par RT-qPCR.

Nous envisageons également, parallèlement, une autre hypothèse pour tenter d'expliquer comment des NACs élevés pourraient jouer un rôle dans la prédisposition génétique au cancer du sein : une augmentation de la fragilité du locus *RNU2* pourrait porter atteinte à l'intégrité de cette région du chromosome 17 et donc conduire à la perte d'un allèle de *BRCA1*. Plusieurs études ont montré que le CNV est un site fragile de cassures de l'ADN, notamment suite à l'infection par l'Adénovirus 12 ou en réponse aux traitements par différentes drogues (Durnam et al., 1988; Li et al., 1993; Liao et al., 1999; Lindgren et al., 1985a; MacArthur et al., 1997; Yu et al., 1998). De plus, cette fragilité est dépendante de l'expression de p53 (Li et al., 1998a). Cependant, cette fragilité n'a pas été étudiée en fonction du nombre de copies du CNV, ce que nous souhaitons maintenant investiguer puisque nous disposons d'un grand nombre de lignées avec des nombres de copies variés. Pour cela, nous avons entrepris une collaboration avec l'équipe du Dr Coquelle (IRCM, Montpellier) qui va utiliser les techniques de FISH et de peignage moléculaire déjà mises au point.

Identifier le mécanisme reliant le nombre de copies du CNV *RNU2* à la prédisposition génétique au cancer du sein serait une avancée importante d'une part

pour confirmer l'analyse statistique sommaire de nos données, mais également pour permettre la mise en place d'un test prédictif basé sur notre test de qPCR. Les individus ayant un NGC élevé déterminé par qPCR pourraient ensuite être analysés plus finement en peignage moléculaire pour déterminer leur NAC. En ce qui concerne ce mécanisme, nous nous sommes focalisés dans un premier temps sur un effet sur le gène *BRCA1*, du fait de la proximité immédiate de ce gène, mais il est envisageable que l'effet soit indépendant de *BRCA1*. Dans ce cas, il serait intéressant de conduire une analyse sur puce pour déterminer les gènes qui sont différentiellement transcrits (suite à une modification du contexte chromatinien dû à l'augmentation du nombre de copies) ou différentiellement épissés (puisque l'ARNsn U2 est impliqué dans l'épissage) lorsque le NGC augmente.

4 Impact du nombre de copies du CNV *RNU2* sur l'expression de l'ARNsn U2

4.1 Introduction

Comme indiqué dans l'introduction de ce manuscrit, l'ARNsn U2 est un composant essentiel de la machinerie d'épissage. Ce complexe dynamique a un rôle clé dans la cellule, puisqu'il est responsable de l'excision des introns, une étape clé du processus de maturation des ARNm. Il a été décrit il y a de nombreuses années dans la littérature que les niveaux d'expression de l'ARNsn U2, tout comme ceux de l'ARNsn U1, sont soumis à un mécanisme de compensation de dosage, mais il faut noter que des arguments appuyant l'existence d'un tel mécanisme ont été apportés uniquement dans le cas de U1 (Bailey et al., 1995; Mangin et al., 1985). De plus, le niveau d'expression de ces deux ARNsn n'a été étudié que dans des lignées cellulaires comportant un nombre moyen de copies puisque des allèles comportant un NAC très élevé n'ont été identifiés que très récemment.

Dernièrement, l'étude du taux d'ARNsn U2 a suscité un nouvel intérêt, en lien avec différentes pathologies. Ainsi, une étude chez la souris a mis en lumière l'importance de la régulation de son expression, puisqu'une augmentation d'expression de l'une des copies parmi les cinq répétitions du gène peut entraîner un défaut de neurogenèse (Jia et al., 2012). Par ailleurs, le taux d'ARNsn U2 circulant dans le plasma semble constituer un biomarqueur de certains cancers (Baraniskin et al., 2013, 2014; Kuhlmann et al., 2014; Mazières et al., 2013). Il est donc important de déterminer l'influence précise du nombre de copies sur le niveau d'expression de cet ARNsn, et les mécanismes mis en jeu pour la régulation spatiale et temporelle de l'expression de ces copies.

4.2 Matériels et Méthodes

- Culture cellulaire

Les lignées lymphoblastoides ont été immortalisées avec le virus d'Epstein-Barr à partir de lymphocytes B provenant d'individus inclus dans l'étude GEMO ou GENESIS (décrites préalablement). Les lignées sont cultivées dans un milieu RPMI (Roswell Park Memorial Institute) supplémenté avec 20 % de sérum de veau fœtal (SVF) et 1 % de pénistreptomycine, dans une étude à 37°C avec 5 % de CO_2.

- Extraction de l'ARN et rétro-transcription

Les ARN totaux ont été isolés à partir des lignées lymphoblastoides avec le kit NucleoSpin miRNA (Macherey-Nagel) selon les recommandations du fournisseur. 50 ng d'ARN ont été rétrotranscrits avec le kit Expand Reverse Transcriptase (Roche Diagnostic) avec 1 µg d'OligodT et de Random Primers (Promega) en suivant les recommandations du fournisseur.

- PCR quantitative

La réaction a été faite sur 2 µl de la réaction de rétrotranscription diluée 50 fois avec le kit GoTaq qPCR Master Mix (Promega), dans un appareil StepOnePlus Real-Time PCR (Life Technologies) dans des plaques 96 puits en suivant les étapes suivantes : 95°C pendant 10 min, 40 cycles à 95°C pendant 15 min et 60°C pendant 1 min. Les amorces utilisées pour la le gène *RNU2* ont été designées avec le logiciel Primer3 v.0.4.0 (http://frodo.wi.mit.edu/primer3/) et synthétisées par Eurofins MWG Operon : 5′-CTCGGCCTTTTGGCTAAGAT-3′ et 5′-CGTTCCTGGAGGTACTGCAA-3′. L'expression du gène *RNU2* a été normalisée grâce aux valeurs obtenues pour le gène de ménage *GAPDH* : 5′-CGGAGTCAACGGATTTGGTCGTAT-3′ et 5′-AGCCTTCTCCATGGTGGTGAAGAC-3′. Chaque échantillon a été analysé en duplicat. L'expression de l'ARNsn U2 a été évaluée avec la méthode comparative du ΔΔCt en suivant les recommandations du fournisseur.

- Analyse de la méthylation par pyroséquençage

50 ng d'ADN ont été convertis avec le kit EZ DNA Methylation-Gold Kit (Zymo Research). La région DSE-PSE est amplifiée avec les amorces : 5'-GTTTYGGGGGYGGAGTTAA-3' (58°C) et 5'-CTCCTATTCCATCTCCCTACT-3' (58°C). La réaction de séquençage est réalisée avec les amorces 5'-GTTTYGGGGGYGGAGTTAA-3' (58°C) et 5'- TACCCCATTCCCTCTATCT-3' (58°C) et la sonde 5'- TTTTTTGTGAAAGGG-3' pour la région DSE, 5'-AGATAGAGGGAATGGGGTA-3' (58°C), 5'- CTCCTATTCCATCTCCCTACT-3' (58°C) et 5'-AGGTTGGGGTTTTTAT-3' pour la région PSE. L'amplification et la réaction de séquençage sont réalisées avec les amorces 5'- TTAGGAAGGAGGAAGGGA-3' (58°C) et 5'- CAACCCTTTATCTCCCRCTC-3' (59°C) ainsi que la sonde 5'-GAGGTAAATGTTGAGT-3' pour la région contrôle. L'amplification est réalisée avec l'enzyme HotStarTaq (Qiagen) en suivant les recommandations du fournisseur. Chaque amorce de PCR est couplée avec de la biotine, permettant de collecter le produit de PCR avec des billes recouvertes de Streptavidine. Après dénaturation, le produit de PCR est ensuite séquencé avec l'amorce de séquençage en utilisant le kit PyroMark Gold Q96 (Qiagen) dans le pyroséquenceur PSQ 96MA (Qiagen) en suivant les recommandations du fournisseur.

- Analyse de la méthylation sur puce

L'analyse du profil de méthylation de 47 individus appartenant au projet HapMap et au projet 1000 Génomes a été réalisée par nos collaborateurs (Equipe du Dr Sharp, Mount Sinai Hospital, New York). Brièvement, les échantillons ont été analysés sur la puce Infinium HumanMethylation450 BeadChip (Illumina). Cette puce permet d'analyser la méthylation sur plus de 485,000 sites CpG, couvrant ainsi 99 % des gènes (RefSeq Genes), avec une moyenne de 17 sites CpG par gène.

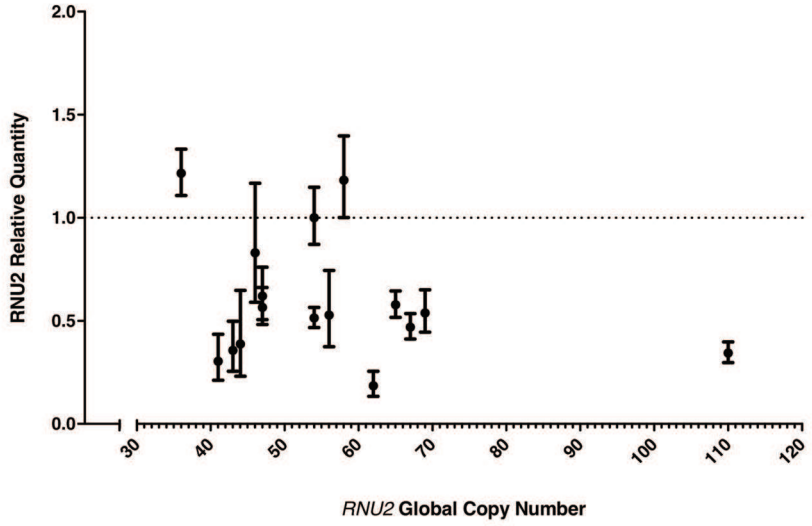


Figure 52 : Expression de l'ARNsn U2 en fonction du nombre global de copies du CNV *RNU2* dans 16 lignées lymphoblastoïdes issus d'individus de l'étude GEMO.

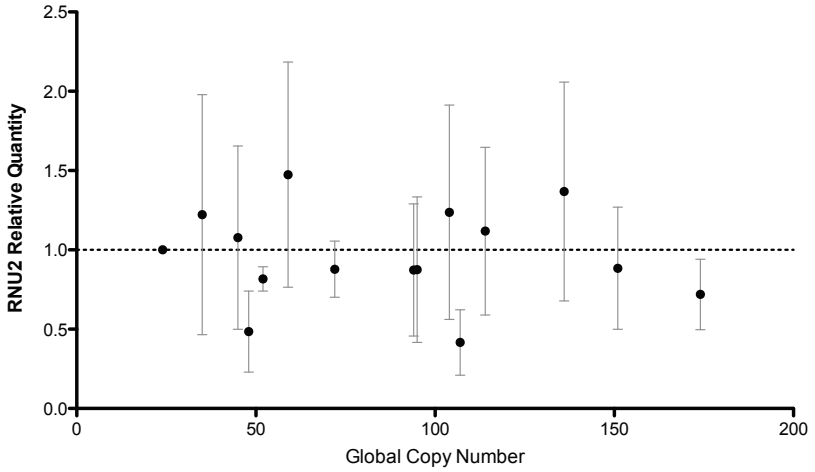


Figure 53 : Expression de l'ARNsn U2 en fonction du nombre global de copies du CNV *RNU2* dans 16 lignées lymphoblastoïdes issus d'individus de l'étude GENESIS.

4.3 Résultats

- Le niveau d'expression de l'ARNsn U2 n'est pas corrélé au NGC du CNV *RNU2*

Dans un premier temps, nous avons étudié le niveau d'expression de l'ARNsn U2 dans 16 lignées lymphoblastoïdes issues d'individus de la cohorte GEMO (voir Article 2), et dont le NGC, déterminé par peignage moléculaire, varie entre 36 et 110 (Figure 52). J'ai observé des variations du niveau d'expression de 6,6 fois entre les individus. Cette variation du niveau d'expression n'est pas corrélée au NGC ou au NAC du CNV.

J'ai confirmé cette absence de corrélation sur 15 individus provenant de l'étude GENESIS et présentant des nombres de copies plus variés (24-174) (Figure 53). En effet, cette fois encore, j'ai observé de légères variations du niveau d'expression de l'ARNsn U2, mais non corrélées au nombre de copies. Ces deux études confirment les résultats de la littérature suggérant un mécanisme de compensation de dosage, c'est-à-dire que toutes les copies du gène *RNU2* ne seraient pas transcrites au même niveau, certaines n'étant pas transcrites du tout.

- **La méthylation des séquences promotrices de l'ARNsn U2 augmente avec le NGC du CNV**

Etant donné que l'activation ou l'inactivation d'une région d'ADN découle de la mise en place de marques épigénétiques particulières, la méthylation de l'ADN étant l'une des marques inactivatrices la plus fréquemment impliquée, nous avons décidé d'étudier le profil de méthylation de l'unité répétée du CNV *RNU2*, en collaboration avec l'équipe de Zdenko Herceg (CIRC, Lyon).

J'ai donc étudié par pyroséquençage le statut de méthylation des cytosines des deux séquences promotrices du gène *RNU2*, DSE (5 cytosines) et PSE (4 cytosines), préalablement décrites comme étant hypométhylées (Jiang and Liao, 1999) chez 50 individus provenant de l'étude GENESIS. En parallèle, j'ai également étudié une séquence contrôle, située 200 pb en amont de la séquence du gène *RNU2*, qui avait été décrite comme étant hyperméthylée (3 cytosines). Pour cette région contrôle, les

Tableau 13 : Etude par pyroséquençage du pourcentage de méthylation de 3 cytosines de la région contrôle de l'unité répétée du macrosatellite *RNU2* chez 50 individus de l'étude GENESIS.

	Région Contrôle				
	CpG 1	**CpG 2**	**CpG 3**	**CpG 1-3**	**CpG 2-3**
Nombre de valeurs	50	50	50	50	50
Minimum	36,10	66,30	70,30	58,50	69,50
25% Percentile	39,80	77,18	83,53	66,45	80,48
Médiane	42,35	82,55	89,30	71,80	86,05
75% Percentile	44,13	86,80	92,23	74,25	89,27
Maximum	45,90	100,0	99,30	81,30	99,70
Moyenne	41,91	81,97	87,39	70,43	84,71
Déviation Standard	2,672	7,510	7,063	5,430	6,974
Erreur Standard	0,3778	1,062	0,9989	0,7680	0,9863
Minimum IC 95%	41,15	79,83	85,38	68,88	82,72
Maximum IC 95%	42,67	84,10	89,40	71,97	86,69
r^2	0,3362	0,4932	0,3887	0,4750	0,4814
p-value	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001

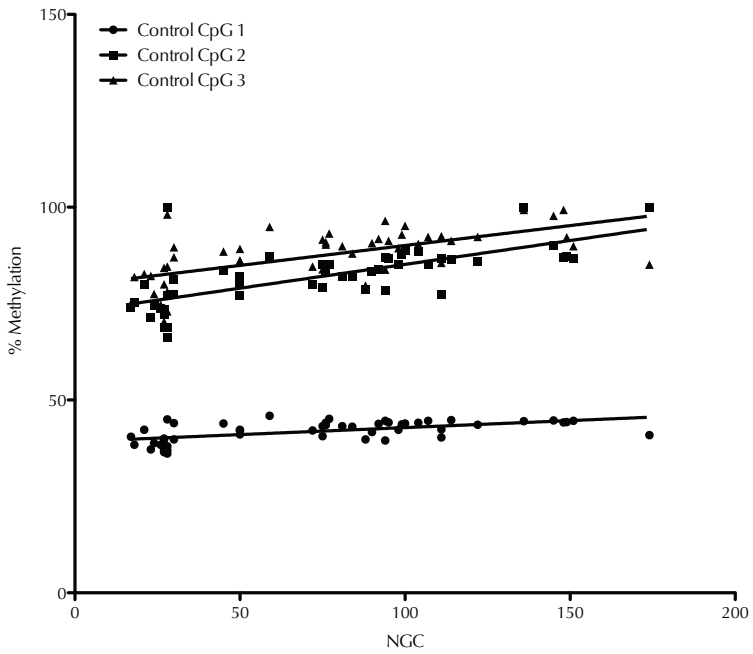


Figure 54 : Pourcentage de méthylation de 3 cytosines de la région contrôle de l'unité répétée du macrosatellite *RNU2* exprimé en fonction du nombre global de copies (NGC) chez 50 individus de l'étude GENESIS.

Tableau 14 : Etude par pyroséquençage du pourcentage de méthylation de 5 cytosines de la région DSE de l'unité répétée du macrosatellite *RNU2* chez 50 individus de l'étude GENESIS.
DSE: Distal Sequence Element

	Région DSE						
	CpG 1	**CpG 2**	**CpG 3**	**CpG 4**	**CpG 5**	**CpG 1-5**	**CpG 4-5**
Nombre de valeurs	50	50	50	50	50	50	50
Minimum	0,0	0,0	0,0	0,0	0,0	0,0	0,0
25% Percentile	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Médiane	0,0	0,0	0,0	0,0	0,0	0,0	0,0
75% Percentile	0,0	0,0	0,0	0,0	0,0	0,1750	0,0
Maximum	0,0	4,600	5,800	5,200	21,00	5,100	10,50
Moyenne	0,0	0,2320	0,3420	0,1040	1,912	0,5180	1,008
Déviation Standard	0,0	0,9364	1,368	0,7354	4,472	1,108	2,372
Erreur Standard	0,0	0,1324	0,1934	0,1040	0,6324	0,1568	0,3355
Minimum IC 95%	0,0000e+000	-0,03413	-0,04665	-0,1050	0,6411	0,2030	0,3338
Maximum IC 95%	0,0000e+000	0,4981	0,7307	0,3130	3,183	0,8330	1,682
r^2	ND	0,01946	0,05281	0,01028	0,4007	0,3655	0,3750
p-value	ND	0,2557	0,0478	0,3545	< 0.0001	< 0.0001	< 0.0001

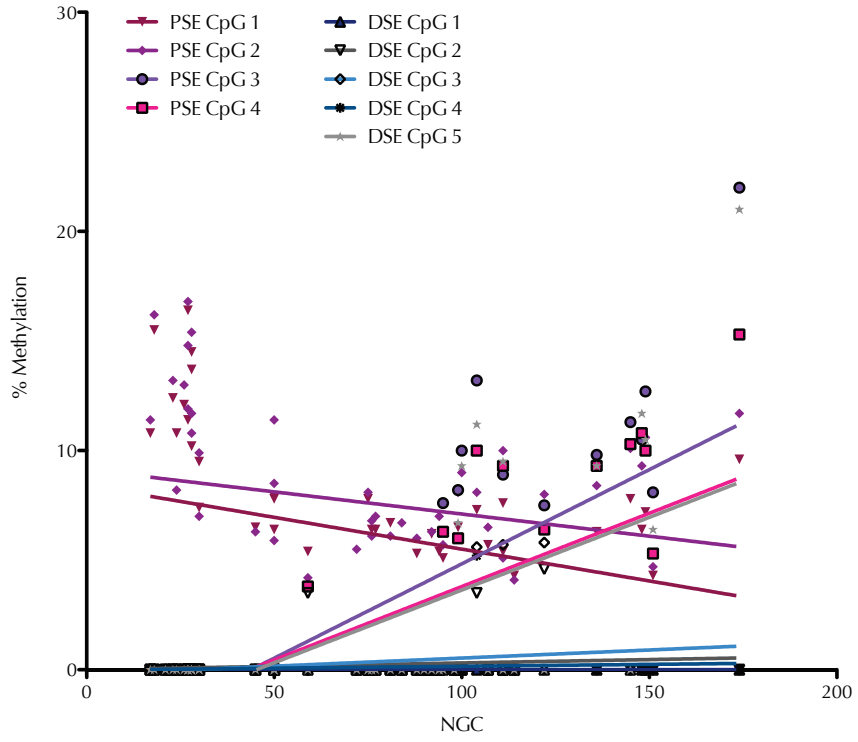


Figure 55 : Pourcentage de méthylation de 5 cytosines de la région DSE et 4 cytosines de la région PSE de l'unité répétée du macrosatellite *RNU2* exprimé en fonction du nombre global de copies (NGC) chez 50 individus de l'étude GENESIS.
DSE: Distal Sequence Element
PSE: Proximal Sequence Element

Tableau 15 : Etude par pyroséquençage du pourcentage de méthylation de 4 cytosines de la région PSE de l'unité répétée du macrosatellite *RNU2* chez 50 individus de l'étude GENESIS.
PSE: Proximal Sequence Element

	Région PSE					
	CpG 1	**CpG 2**	**CpG 3**	**CpG 4**	**CpG 1-4**	**CpG 3-4**
Nombre de valeurs	50	50	50	50	50	50
Minimum	0,0	0,0	0,0	0,0	0,0	0,0
25% Percentile	3,225	5,650	0,0	0,0	2,494	0,0
Médiane	6,400	7,500	0,0	0,0	4,038	0,0
75% Percentile	8,300	10,65	1,875	0,9500	6,575	2,675
Maximum	16,40	16,80	22,00	15,30	14,65	18,65
Moyenne	6,258	7,632	2,596	2,056	4,636	2,326
Déviation Standard	4,486	4,344	5,036	3,984	3,215	4,438
Erreur Standard	0,6344	0,6144	0,7122	0,5634	0,4547	0,6277
Minimum IC 95%	4,983	6,397	1,165	0,9239	3,722	1,065
Maximum IC 95%	7,533	8,867	4,027	3,188	5,549	3,587
r^2	0,07618	0,03946	0,5309	0,5139	0,1183	0,5403
p-value	0,0216	0,0771	< 0.0001	< 0.0001	0,0145	< 0.0001

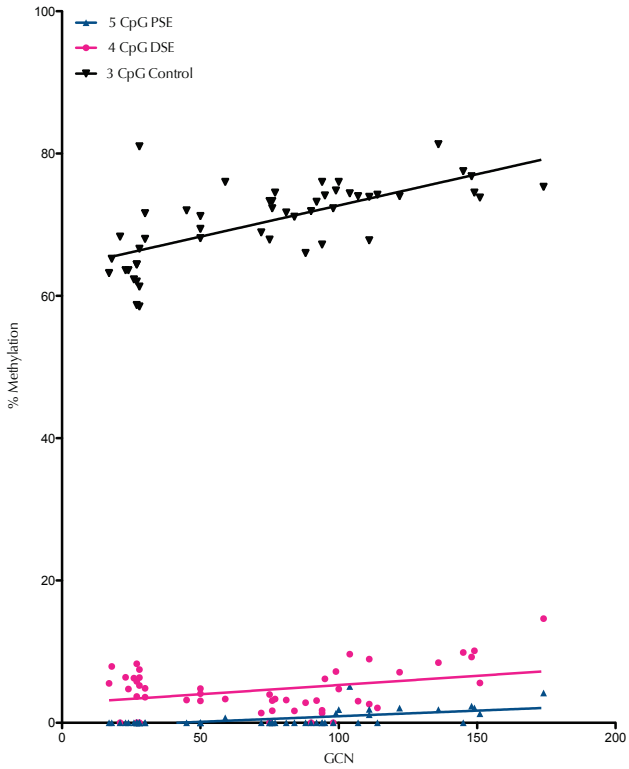


Figure 56 : Pourcentage moyen de méthylation de la région DSE, de la région PSE et de la région contrôle de l'unité répétée du macrosatellite *RNU2* exprimé en fonction du nombre global de copies (NGC) chez 50 individus de l'étude GENESIS.
DSE: Distal Sequence Element
PSE: Proximal Sequence Element

cytosines 2 et 3 sont en moyenne fortement méthylées chez les 50 individus (81,97 % et et 87,39 % respectivement) (Tableau 13 & Figure 54). Etonnement, la cytosine 1 est en moyenne plus faiblement méthylée : 41,91 %. Ainsi, le pourcentage de méthylation pour la région contrôle varie pour les 50 individus entre 58,5 % et 81,30 % si on considère les 3 cytosines et de 69,5 % à 99,70 % si on ne considère que les $2^{ème}$ et $3^{ème}$ cytosines.

Pour la région DSE, les 4 premières cytosines analysées ne sont que très rarement méthylées chez les individus analysés : la cytosine 1 n'est jamais méthylée, tandis que les trois autres le sont chez quelques individus (n = 3, n = 3 et n = 1) mais très faiblement (Maximum de méthylation : 4,6 %, 5,8 % et 5,2 % respectivement) (Tableau 14 & Figure 55). Pour ces trois cytosines, le niveau moyen de méthylation pour les 50 individus est proche de 0 (0,230 %, 0,342 % et 0,104 %). *A contrario*, pour la cytosine 5, 9 des 12 individus présentant un NGC supérieur à 100 sont méthylés, avec un pourcentage de méthylation variant de 6,7 à 21.

Pour la région PSE, les cytosines 1 et 2 présentent un profil de méthylation différent de celui des cytosines 3 et 4 (Tableau 15 & Figure 55). Les premières sont certes faiblement méthylées, mais le sont pour tous les individus : elles présentent un niveau « basal » faible de méthylation (en moyenne 6,26 % et 7,63 %). A l'inverse, les cytosines 3 et 4 ne sont méthylées que chez des individus présentant un NGC supérieur à 50.

Ainsi, pour ces 50 individus, en considérant l'ensemble des cytosines analysées, la région contrôle est en moyenne méthylée à 70,43 %, la région PSE à 4,64 % et la région DSE à 0,51 %. Ces observations confirment la structure bimodale de la méthylation de l'unité répétée du CNV : une région contrôle hyperméthylée et une région contenant les séquences promotrices et le gène *RNU2* hypométhylée. De manière très intéressante, on remarque que le statut de méthylation pour chacune de ces régions est corrélé au NGC du CNV : la région contrôle (r^2 = 0,4750, p-value < 0,0001), la région PSE (r^2 = 0,118, p-value = 0,0145) et la région DSE (r^2 = 0,3655, p-value < 0,0001) (Figure 56). Pour la région PSE, la méthylation des deux premières cytosines étudiées n'étant pas corrélée au NGC, contrairement aux cytosines 3 et 4 (r^2

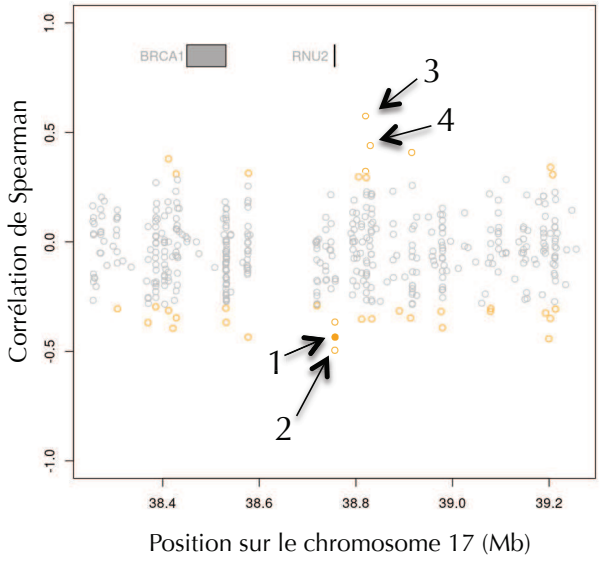


Figure 57 : Valeur de corrélation de Spearman entre le niveau de méthylation et le nombre haploïdique de copies pour toutes les sondes de méthylation situés à ± 500 kb du locus *RNU2* (Puce 450 BeadChip) chez 47 individus du projet 1000 Génomes.
1,2: deux sondes localisées sur le CNV *RNU2*, et présentées en Figure 58.
3,4: deux sondes dans les régions avoisinantes du CNV *RNU2*, présentées en Figure 59.

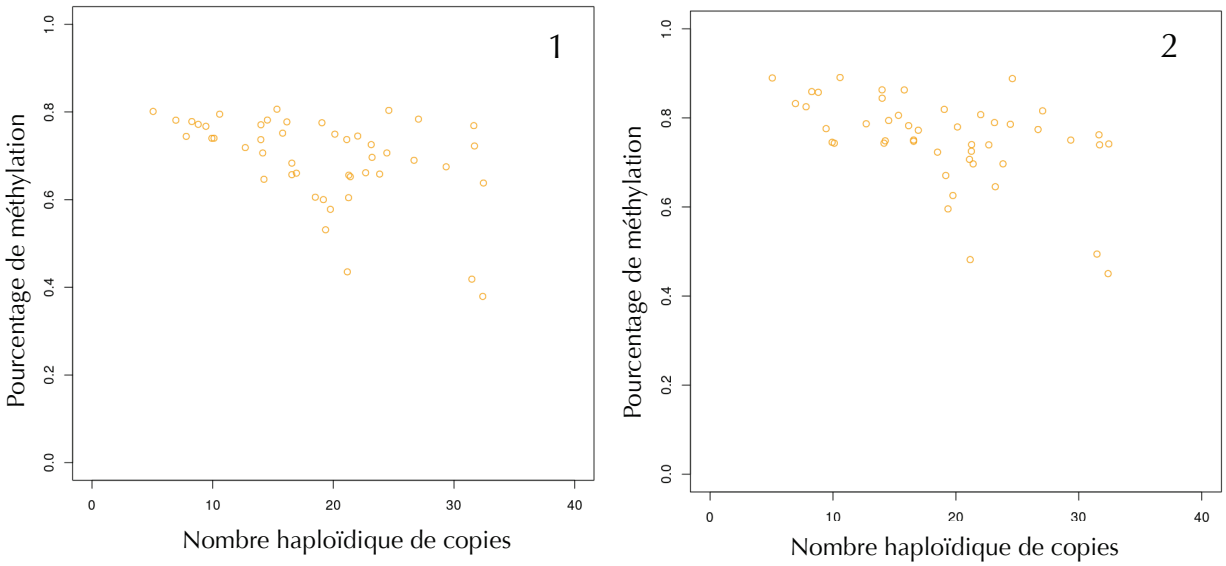


Figure 58 : Pourcentage de méthylation de deux sondes du locus *RNU2* en fonction du nombre haploïdique de copies chez 47 individus du projet 1000 Génomes.
1. Sonde cg17378535, r = -0,49, p-value = 4,19E-04
2. Sonde cg24047905, r = -0,44, p-value = 2,21E-03
La valeur de corrélation négative (r) entre le nombre haploidique de copies et le niveau de méthylation pour ces deux sondes est illustrée en Figure 57.

= 0,6733, p-value < 0,0001), l'effet est plus fort en ne prenant en compte que ces deux dernières (r^2 = 0,5403, p-value < 0,0001).

- **L'augmentation du nombre global de copies s'accompagne de modifications épigénétiques au niveau du locus *RNU2*, et potentiellement au niveau des séquences avoisinantes**

De façon intéressante, 47 individus analysés dans le cadre du projet 1000 Génomes et pour lesquels nous avons pu estimer le nombre haploïdique de copies (NHC), ont également été analysés sur une puce de méthylation par l'équipe du Dr Sharp (Mount Sinai Hospital, New York). Cette puce permet de regarder le statut de méthylation de 450 000 sites CpG répartis dans le génome (17 sites CpG par gène en moyenne). Nos collaborateurs ont analysé plusieurs CpG situés dans une fenêtre de 500 kb autour de la position attendue du locus *RNU2* (telle que nous avons pu l'estimer dans le Chapitre 2 – Section 1) sur la version hg18 du génome de référence. Pour chaque sonde, nos collaborateurs ont regardé le niveau de corrélation (rho de Spearman) entre le NHC et le niveau de méthylation (Figure 57).

Pour les trois sondes couvrant le CNV *RNU2*, on observe une corrélation négative entre le NHC et le statut de méthylation (sonde cg17378535, r = -0,49, p-value = 4,19E-04 ; sonde cg24047905, r = -0,44, p-value = 2,21E-03 ; sonde cg13726456, r = -0,37, p-value=1,14E-02) (Figures 57 & 58). Effectivement, lorsqu'on regarde plus précisément chez les 47 individus le niveau de méthylation des deux sondes donnant les plus fortes valeurs d'association (sonde cg17378535 et sonde cg24047905), on observe que celui-ci diminue lorsque le NHC augmente (Figure 58).

Cette analyse a permis d'identifier une corrélation positive entre le NHC du CNV *RNU2* et la méthylation des CpG situés dans ou à proximité des gènes *ARL4D* (sonde cg23635580, r = 0,44, p-value = 1,84E-03) et *LOC100130581* (sonde cg21506159, r = 0,58, p-value = 2,18E-05) (Figure 59). Pour ces deux sondes, situées respectivement à 75 et 65 kb du locus *RNU2*, on observe une augmentation de la méthylation, certes faible (pente de la droite de régression), corrélée à l'augmentation du NHC. Ces deux gènes sont situés en position télomérique par rapport au CNV

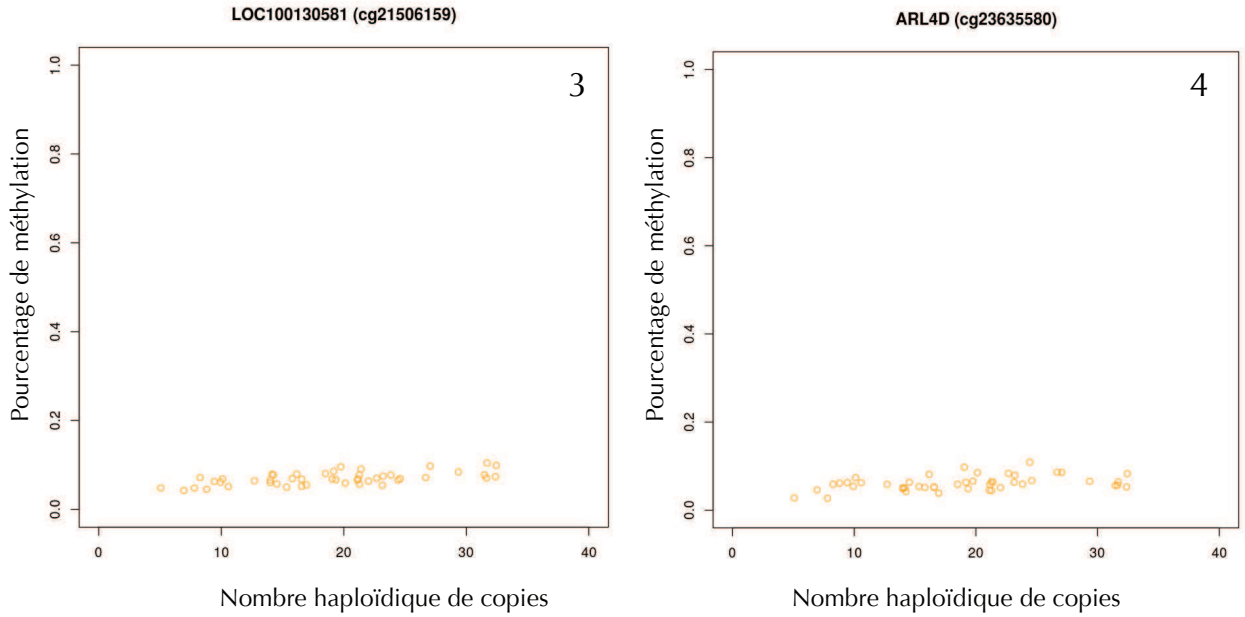


Figure 59 : Pourcentage de méthylation de deux sondes de méthylation reconnaissant des gènes dans la région environnante du locus *RNU2* en fonction du nombre haploïdique de copies chez 47 individus du projet 1000 Génomes.

3. *LOC100130581 :* sonde cg21506159, r = 0,58, p-value = 2,18E-05
4. *ARL4D :* sonde cg23635580, r = 0,44, p-value = 1,84E-03

La valeur de corrélation positive (r) entre le nombre haploïdique de copies et le niveau de méthylation pour ces deux sondes est illustrée en Figure 57.

(Figures 57 et 59). Le locus *LOC100130581* contient le pseudogène *RNU2-4P,* et présente de fortes homologies avec l'unité de base du CNV. Le gène *ARL4D* code pour l'ADP-ribosylation factor-like 4D.

Pour d'autres sondes situées autour du locus *RNU2*, on observe une potentielle corrélation, positive ou négative, entre le NHC et le niveau de méthylation ($r > 0,3$), sans que celle-ci ne soit significativement significative (p-value > 0,005). La faible taille de notre échantillonage ($n = 47$), ajoutée au fait que ces individus ne portent pas des nombres très élevés de copies, peuvent limiter considérablement notre pouvoir statistique. En effet, le NHC pour ces individus varie entre 5 et 32, avec un nombre moyen de 19.

Des analyses complémentaires sont nécessaires pour conclure sur les répercussions d'un changement du nombre de copies sur la méthylation des régions avoisinantes. Cependant, cette étude nous apporte de nouvelles preuves que le changement du nombre de copies a un effet sur la méthylation du locus lui-même, et qu'il pourrait également en avoir sur des séquences avoisinantes.

4.4 Discussion

Suite à l'analyse d'un grand nombre de cas de cancer du sein et de témoins de l'étude GENESIS, j'ai pu identifier des individus porteurs de NGC variés, dont certains avec un NGC particulièrement élevé. La caractérisation et l'accès à des lignées cellulaires de ces individus ont fourni un outil inestimable pour l'étude des répercussions fonctionnelles de la variabilité du nombre de copies du locus *RNU2*. Dans un premier temps, je me suis intéressée au niveau d'expression de l'ARNsn U2, et j'ai montré dans 28 lignées lymphoblastoïdes que celui-ci n'est pas corrélé au NGC, confirmant les données de la littérature suggérant l'existence d'un mécanisme de compensation de dosage. Pour déterminer les bases moléculaires de ce mécanisme, j'ai donc investigué une marque épigénétique que l'on sait être inhibitrice : la méthylation des cytosines de l'ADN. J'ai confirmé les données de la littérature sur le profil de méthylation bimodal de l'unité répétée du CNV : une région contrôle hyperméthylée, et une région contenant les séquences promotrices et le gène *RNU2* hypométhylée. J'ai montré par analyse de l'ADN traité au bisulfite de 50 individus que le niveau de méthylation de ces deux régions augmente avec le NGC.

Pour cette analyse, nous avons regardé plusieurs cytosines sur chacune de ces régions. J'ai pu remarquer que les niveaux de méthylation des cytosines d'une même région variaient, certains étant dépendant du NGC, d'autres non, bien que ces cytosines ne soient espacées que de quelques nucléotides. Ceci suggère que certaines ont une fonction régulatrice alors que d'autres non. Il serait intéressant d'en tester d'autres afin d'établir clairement quelles sont les cytosines potentiellement *fonctionnelles*. De même, il serait intéressant de séquencer les transcrits pour déterminer si les copies qui s'expriment sont porteuses de SNP particulier en commun.

- **Le locus *RNU2*, une répétition de dosage**

L'ARNsn U2 est une molécule requise en grande quantité dans la cellule, tout comme d'autres molécules telles que les ARN ribosomaux, les ARN de transfert, les ubiquitines et les histones. Ces molécules ont en commun d'être codées par des gènes

présents au sein de répétitions en tandem. La transcription d'un seul gène ne pouvant produire qu'un nombre limité de molécules d'ARN, la présence de plusieurs copies du gène est une façon pour la cellule de produire un plus grand nombre de ces molécules, tout comme la transcription simultanée par plusieurs polymérases d'une même copie. Ces répétitions ont été appelées répétitions en tandem *de dosage*, et se distinguent des familles multigéniques et des gènes dupliqués à différents loci. Par exemple, chez l'Homme, on retrouve plus de 30 à 40 répétitions des 5 clusters contenant l'ADN ribosomal (ADNr), plus de 400 gènes codant pour des ARN de transfert (ARNt), plus de 10 à 20 fois chaque gène codant pour une histone.

Aucune étude n'a pour le moment identifié chez un individu un nombre allélique minimum inférieur à 5 pour le locus *RNU2* (Pavelitz et al., 1995; Schaap et al., 2013). Une diminution drastique du nombre de copies, qu'une plus grande activité de la polymérase ne pourrait compenser, aboutirait vraisemblablement à un déficit en ARNsn U2 et se traduirait par de graves conséquences au niveau de la cellule, entraînant sans doute un phénotype létal. Pour étayer cette hypothèse, pour certains macrosatellites décrits jusqu'à présent, le nombre allélique minimum identifié est de 1, comme pour le locus *D4Z4* chez les patients atteints de FSHD (Tableau 4).

A l'inverse, selon les macrosatellites et selon les individus, on remarque une grande variabilité pour le nombre allélique maximum de copies, soulignant une tendance à l'augmentation du nombre de copies (Schaap et al., 2013; Tessereau et al., 2013; Tremblay et al., 2010, 2011). Cette augmentation se traduit par un nombre allélique de copies extrêmement varié selon les individus, et donc par la mise en place d'un mécanisme de compensation de dosage comme nous l'avons vu. Pour les individus porteurs d'un faible NGC, il est vraisemblable de penser que toutes les copies sont exprimées. Or, il a été montré que les copies non exprimées du locus de l'ADNr jouent un rôle majeur dans la protection de ce locus contre une fragilité induite par une forte concentration de gènes hautement transcrits (Ide et al., 2010). Il sera alors très intéressant d'étudier la fragilité du locus chez les individus présentant un NGC faible (< 20).

A l'inverse, il est difficile d'imaginer quelles seraient les conséquences d'un défaut de compensation de dosage chez des individus porteurs d'un NGC élevé, c'est-

à-dire si toutes leurs copies étaient exprimées. La cellule devrait alors faire face à un excès d'ARNsn U2, qui se traduirait peut-être par un excès d'épissage ou un épissage incorrect. Il serait intéressant dans les lignées porteuses d'un NGC élevé de déméthyler l'ADN par un traitement à l'azacytidine et d'en étudier les conséquences. Il a été montré que, suite à ce traitement, les régions hétérochromatiques perdent leur haut degré de compaction et se décondensent (de Capoa et al., 1996; Haaf and Schmid, 2000).

- **Localisation des copies non actives et hétérochromatinisation de la région**

De nombreuses études ont également permis d'associer la variation du nombre de copies de répétitions en tandem chez l'Homme avec des changements épigénétiques dans la région : l'expansion des répétitions CGG à l'origine du syndrome du X fragile est associée à l'hyperméthylation du promoteur du gène *FMR1* (Bardoni and Mandel, 2002) ; les contractions des répétitions *D4Z4* à l'origine du syndrome FSHD sont associées à une perte de méthylation de la région (van Overveld et al., 2003). Ces observations ont été également faites par l'équipe du Dr Sharp après l'étude de 186 répétitions en tandem (Brahmachary et al., soumis). Par ailleurs, ils ont également identifié 138 gènes dont l'expression varie en fonction du nombre de copies de ces répétitions en tandem. La majorité de ces gènes sont localisés à l'extérieur des répétitions, à une distance médiane de 208 kb (Brahmachary et al., soumis). Ceci suggère qu'un mécanisme similiaire au Repeat-Induced Gene Silencing, initialement découvert chez des organismes transgéniques (Assaad et al., 1993; Dorer and Henikoff, 1994; Garrick et al., 1998; Ye and Signer, 1996), pourrait avoir lieu dans le génome humain (Brahmachary et al., soumis).

Il a été montré chez l'Homme par FISH que l'ADN satellite colocalise avec l'hétérochromatine constitutive, et possède des caractéristiques de repliement qui lui sont propres. Par ailleurs, chez la souris, la Drosophile et les plantes, des transgènes présents en multiple copies sont très faiblement exprimés, voire pas du tout, et ceci de manière indépendante d'une localisation au niveau de l'hétérochromatine centromérique (Assaad et al., 1993; Dorer and Henikoff, 1994; Garrick et al., 1998; Ye and Signer, 1996). Ces différentes observations suggèrent que les répétitions en tandem

adoptent des structures particulières et peuvent ainsi conduire à la formation directe d'hétérochromatine et à sa propagation. L'hétérochromatinisation de la région conduit alors à une perte d'expression des gènes dans la région. Ce mécanisme est communément appelé *Repeat-Induced Gene Silencing* (RIGS).

Pour le locus *RNU2*, la variation de méthylation que nous observons confirme que, chez les individus porteurs d'un grand NGC, certaines copies sont hyperméthylées et donc vraisemblablement transcriptionnellement inactives. Nous nous intéressons actuellement à la localisation de ces copies réprimées : sont-elles aléatoirement réprimées ou peut-on observer une localisation préférentielle au sein de l'allèle ? Une hypothèse est que les copies non exprimées soient localisées à proximité les unes des autres à une des extrémités du locus, par exemple du côté du promoteur *BRCA1*, aboutissant à une hétérochromatinisation de la région. Pour tester cette hypothèse, la technique du peignage moléculaire pourrait une fois de plus être utile. Chez l'individu avec le NGC le plus élevé, j'ai testé un nouveau code-barres *RNU2* ne comportant que 2 couleurs, et j'ai ajouté un anticorps reconnaissant les cytosines méthylées. Je me suis pour le moment heurtée à un problème de niveau de résolution de la technique : la région contrôle étant hyperméthylée, il est impossible de distinguer spécifiquement les cytosines correspondant aux séquences promotrices. Nous réfléchissons à des améliorations du protocole.

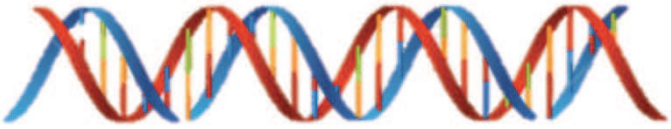
Les données préliminaires obtenues par le Dr Sharp suggèrent effectivement que la variation du nombre de copies du CNV s'accompagne de modifications épigénétiques sur le locus lui-même, mais pourrait également s'accompagner de modifications épigénétiques sur les régions aux alentours. Pour confirmer cela, nous allons prochainement regarder par pyroséquençage le niveau de méthylation de cytosines localisées au sein du promoteur *BRCA1* et des gènes *NBR1* et *NBR2*.

Enfin, la méthylation d'une région d'ADN varie en fonction du type cellulaire étudié, mais également au cours des divisions cellulaires. Lorsque c'était possible, nous avons récupéré des tumeurs incluses en bloc de paraffine pour les cas index de l'étude GENESIS avec un NGC élevé (> 100). Nous allons très prochainement regarder sur l'ADN tumoral le niveau de méthylation par pyroséquençage sur les cytosines déjà testées, afin de mettre en évidence d'éventuelles différences entre les tissus. Nous envisageons également d'étudier d'autres marques épigénétiques répressives, telles que

la tri-méthylation de la lysine 9 sur l'histone H3 (H3K9Me3), par ChIP sur les lignées lymphoblastoïdes d'individus avec des NGC variés. Ces autres marques pourraient également être nécessaires pour le mécanisme de compensation de dosage.

Cette étude nous a permis de confirmer l'existence d'un mécanisme de compensation de dosage au niveau du locus *RNU2*, permettant à la cellule de générer une quantité importante, mais ne variant pas au-delà d'un facteur 7, d'ARNsn U2, alors que j'ai montré que le nombre de copies du locus pouvait varier d'un facteur 20. Nous avons montré que ce mécanisme requiert la méthylation des cytosines de l'ADN sur l'ensemble de l'unité répétée (la région contrôle et les régions promotrices). Nous avons mis en évidence que la variation du nombre de copies du CNV *RNU2* a des répercussions sur la méthylation du locus lui-même, et qu'elle pourrait également en avoir sur les régions avoisinantes. Ceci pourrait se traduire par une hétérochromatinisation de la région et donc un changement d'expression des gènes voisins. Il est bien évidemment nécessaire de confirmer ces observations sur un plus grand nombre d'individus. Pour cela, nous aimerions réaliser une analyse quantitative de la méthylation de l'ADN sur l'ensemble de la région, voire sur le génome complet, des individus de l'étude GENESIS porteurs de NGC variés.

Chapitre 3
Discussion générale et Perspectives



L'objectif de ma thèse a été de caractériser le macrosatellite *RNU2*, et d'étudier son éventuelle implication dans la prédisposition génétique au cancer du sein.

L'un des principaux prérequis de l'étude des répercussions des macrosatellites est l'estimation précise des variations du nombre de copies dans la population générale. Comme nous l'avons vu dans l'introduction, les séquences répétées en tandem sont difficiles à assembler, et sont de fait souvent absentes du génome de référence. Une étude récente, non publiée, conduite sur les données de séquençage du génome de référence par l'étude du Dr Sharp a permis d'identifier 178 répétitions en tandem, dont 89 qui n'avaient pas été identifiées par l'étude la plus complète publiée à ce jour (Conrad et al., 2010). 34 % d'entre elles sont situées dans des régions non assemblées. Par ailleurs, le nombre de copies présent au sein de l'assemblage est une sous-représentation du nombre de copies *réel*. Ainsi, la compréhension des conséquences fonctionnelles des macrosatellites passe par une meilleure évaluation de la diversité génétique *réelle*. Pour cela, il faudra attendre de récolter des données de qualité sur un plus grand nombre d'individus, bien que le projet de séquençage de 1000 Génomes humains y ait d'ores-et-déjà considérablement contribué.

Au cours de ma thèse, j'ai pu mettre en place trois approches permettant d'obtenir des indicateurs de la diversité génétique *réelle* au locus *RNU2* : le nombre allélique de copies par peignage moléculaire, le nombre haploidique de copies par mesure de la profondeur de couverture de séquençage, et enfin le nombre global de copies par PCR quantitative. J'ai montré pour chacune de ces approches que les mesures réalisées étaient cohérentes. Ce travail, ainsi qu'une étude publiée l'année dernière, ont permis d'identifier 66 allèles différents contenant de 5 à 48 répétitions, de 50 à 58 répétitions, de 60 à 64 répétitions, de 73 à 75 répétitions, 79, 82, 89, 93, 123 ou 161 répétitions. Parmi ces allèles, 13 n'avaient jamais été caractérisés jusqu'à présent, dont tous ceux contenant plus de 63 répétitions identifiés chez des cas de cancer du sein, et couvrant donc plus de 380 kb. Le plus grand allèle que nous avons observé contient 163 répétitions, qui couvrent plus de 1 Mb d'ADN. Mon travail a donc permis de préciser le

niveau de polymorphisme du locus *RNU2*, posant ainsi les bases nécessaires pour l'étude des répercussions de la variation du nombre de copies.

En analysant plus de 2000 cas de cancer du sein et 2000 témoins, j'ai pu observer une association entre le nombre global de copies du CNV et la prédisposition génétique au cancer du sein. Dans les lignées lymphoblastoïdes, je n'ai pas observé de corrélation entre le nombre de copies et le niveau d'expression de *BRCA1*. Cependant, j'ai pu mettre en évidence une augmentation de la méthylation sur le locus *RNU2*. D'après une analyse préliminaire, ces modifications épigénétiques pourraient se répercuter sur l'ensemble de la région, confirmant l'existence d'un mécanisme de *Repeat-Induced Gene Silencing* mis en évidence sur d'autres répétitions en tandem. Or, les modifications épigénétiques ainsi que le contexte chromatinien sont extrêmement variables d'un tissu à un autre. L'étude prochaine de tumeurs provenant de femmes porteuses de nombre de copies variés nous permettra donc de déterminer si ces modifications épigénétiques se propagent jusqu'au promoteur *BRCA1*, entraînant une diminution de son expression, et constituant ainsi un évènement précoce dans le processus tumoral.

Par ailleurs, l'extrême variabilité de la taille du locus soulève la question du mécanisme à l'origine d'un changement du nombre de copies. En suivant au cours du temps plusieurs allèles du CNV *RNU2* associés à des mutations ancestrales de *BRCA1*, nous avons pu évaluer le taux de mutation de ce locus, et exclure un changement du nombre de copies par crossing-over. Des recherches supplémentaires sont nécessaires pour identifier clairement le mécanisme. Cependant, nous pouvons nous interroger sur la dynamique de mutation de chaque allèle du CNV *RNU2* : est-ce que chaque allèle a la même probabilité de muter (*i.e.* gagner ou perdre une ou plusieurs copies) ? Est-ce qu'une mutation conduit à un seul évènement de gain ou de perte de copies ou bien est-ce que le nombre de copies gagné ou perdu est aléatoire ? Comme pour les répétitions de trinucléotides, on peut imaginer que le risque d'expansion croît en

fonction de la taille de l'allèle muté. Il faudrait pour tester cela suivre plusieurs allèles de tailles variées au cours d'un grand nombre de divisions cellulaires.

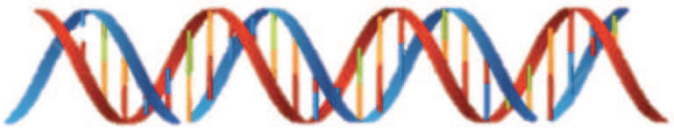
Selon cette hypothèse, on peut imaginer que l'expansion du nombre de copies du locus participe à l'augmentation de sa fragilité, tel que cela a été montré pour le locus *FRAXA* dans le syndrome du X fragile. Le CNV *RNU2* a été caractérisé comme un site fragile de l'ADN, suite à l'infection par l'Adénovirus 12 ou après un traitement par l'Actinomycine D. Cependant, ces expériences n'ont été conduites que sur un nombre restreint de lignées cellulaires transfectées avec des constructions du locus *RNU2,* et pour lesquelles le nombre de copies n'était que peu variable. Ainsi, l'impact de la variation du nombre de copies sur la fragilité du locus n'a jamais été étudié. Nous possédons aujourd'hui les outils nécessaires pour conduire cette analyse.

La fragilité induite par un grand nombre de copies pourrait conduire dans certains tissus, tels que le tissu mammaire, à une cassure au niveau de la région *RNU2*, qui pourrait lors de la réparation de cette cassure endommager le gène *BRCA1*, étant donné sa proximité, et constituerait ainsi un nouveau mécanisme d'inactivation du gène *BRCA1*. La fragilité d'un macrosatellite n'a encore jamais été impliquée dans la prédisposition génétique à un cancer.

L'identification du mécanisme moléculaire à l'origine du lien entre la variation du nombre de copies du macrosatellite *RNU2* et la prédisposition génétique au cancer du sein constituerait une avancée majeure pour la prise en charge des femmes prédispoées. Ceci permettrait de mettre en place un test diagnostique qui pourra être proposé aux femmes appartenant à des familles de cancer du sein et pour lesquelles aucune mutation sur *BRCA1* ou *BRCA2* n'est identifiée. Dans un premier temps, la quantification du nombre global de copies par qPCR permettrait de réaliser un premier crible afin d'identifier des femmes potentiellement porteuses d'un nombre anormalement élevé de copies. Ces femmes pourront ensuite être analysées par peignage moléculaire pour déterminer précisément la distribution allélique. Les femmes

porteuses d'un nombre allélique de copies supérieur à une valeur seuil, qu'il faudra déterminer, se verront alors proposer un suivi personnalisé, comme les porteuses de mutations *BRCA1/2*.

Chapitre 4
Autres applications du code-barres *RNU2*



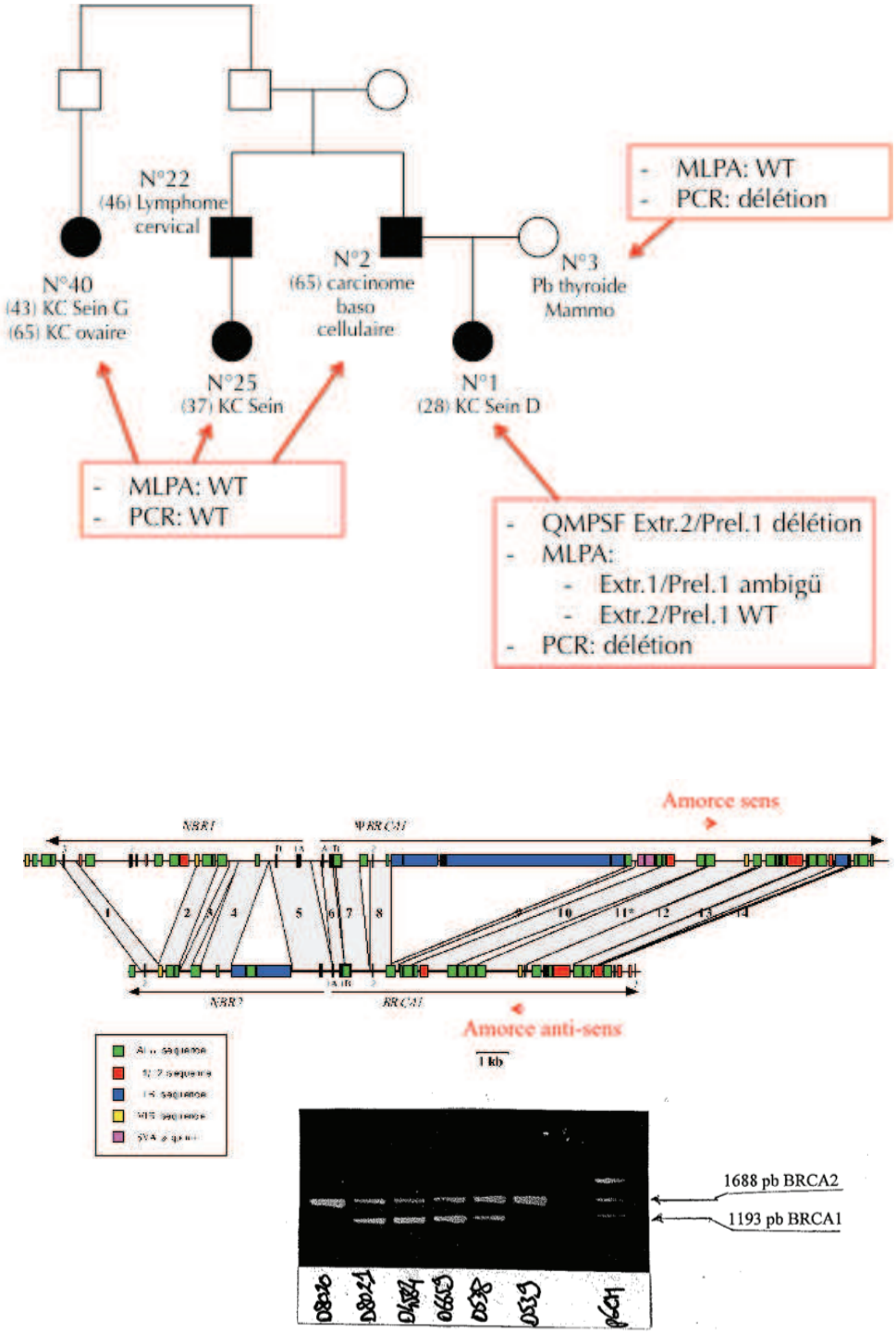


Figure 60 : Résultat discordant sur la présence d'une déléttion de 37 kb au niveau du promoteur *BRCA1* chez une patiente.

A. Arbre généalogique de la famille de la patiente et résultats obtenus avec les principales techniques

B. Résultat positif pour le test PCR décrit ci-dessus à partir d'ADN extrait de deux prélèvements sanguins.

Le test utilise une amorce sens complémentaire d'une séquence du pseudogène et une amorce antisens complémentaire d'une séquence du gène *BRCA1*.

1 Première identification d'une conversion génique entre le gène *BRCA1* et le pseudogène *BRCA1P1* ?

Les réarrangements génomiques issus d'évènements de recombinaison homologue non allélique sont à l'origine de nombreux cas de pathologies héréditaires. Ce type de réarrangement est favorisé par les régions d'ADN présentes dans le génome en plusieurs copies et ayant une homologie de séquence très élevée (> 95 %). Ainsi, mon équipe a identifié il y a une dizaine d'années un point chaud de recombinaison dans le locus *BRCA1*. Ce point-chaud résulte de la duplication chez les primates des exons 1a, 1b et 2 de *BRCA1*, ce qui a conduit à l'apparition d'un pseudogène (*BRCA1P1*) situé à environ 30 kb de *BRCA1* en position télomérique (voir Chapitre 4). Des évènements distincts de recombinaison entre l'intron 2 de *BRCA1* et l'intron 2 de *BRCA1P1* conduisent à des délétions de 37 kb qui ont été identifiées dans de nombreuses familles de cancer du sein d'origines diverses (Puget et al., 2002). On observe dans les allèles mutés un remplacement des 2 premiers exons de *BRCA1* par ceux du pseudogène, conduisant à la perte du promoteur et du codon d'initiation de la traduction (situé dans l'exon 2) puisque l'exon 2 de *BRCA1P1* diverge au niveau de ce codon. Ce type de réarrangement est maintenant systématiquement recherché dans les tests diagnostiques de routine à l'aide d'un essai PCR dédié utilisant une amorce sens complémentaire d'une séquence du pseudogène et une amorce antisens complémentaire d'une séquence du gène *BRCA1* et permettant donc l'amplification d'un petit fragment uniquement lorsque ces séquences sont rapprochées.

Bien que nous ayons obtenu pour une patiente ayant une forte histoire familiale de cancer du sein un résultat positif pour le test PCR décrit ci-dessus à partir d'ADN extrait de deux prélèvements sanguins, le résultat du test MLPA n'était pas en faveur d'une délétion des exons 1 et 2. Le séquençage du fragment amplifié a confirmé la juxtaposition de séquences homologues au pseudogène avec des séquences

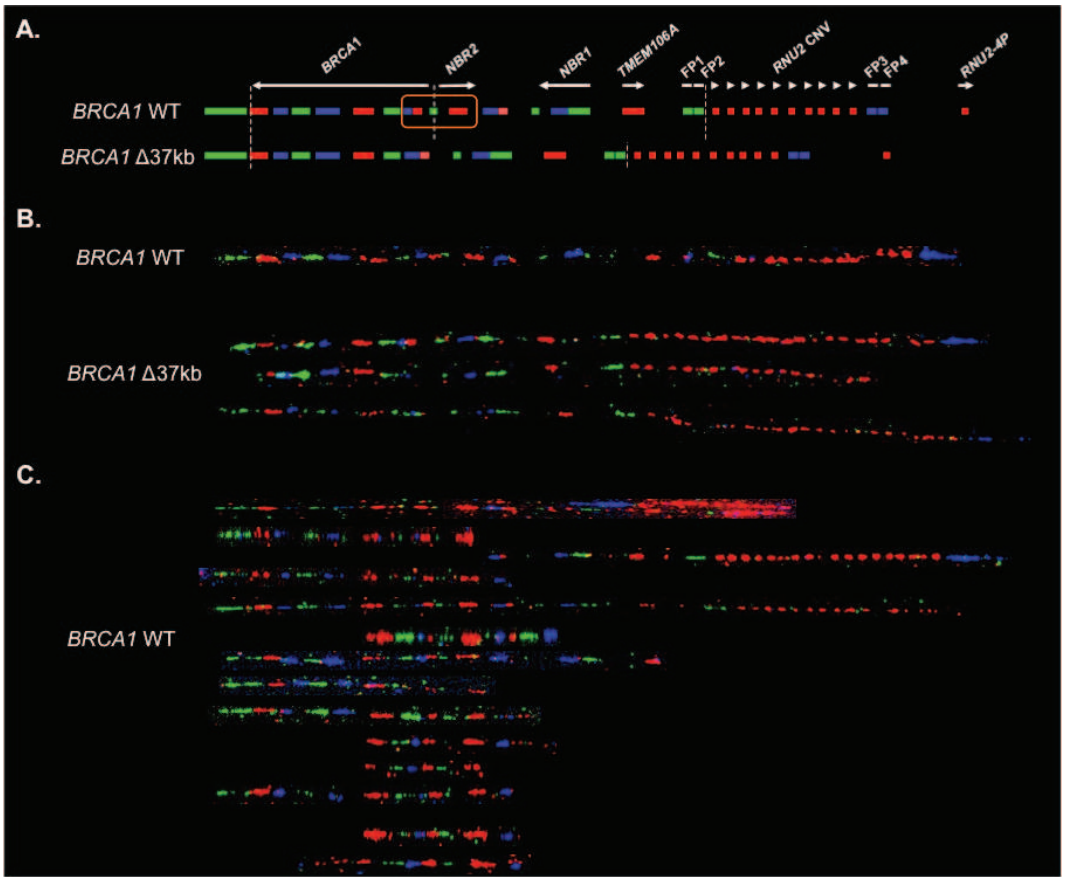


Figure 61: Analyse par peignage moléculaire avec des sondes couvrant la région (A), d'un individu avec la délétion de 37 kb (B) et de la patiente N°1 (C).
B. Observation de l'allèle *BRCA1* WT et de l'allèle *BRCA1* muté;
C. Observation uniquement de l'allèle *BRCA1* WT.

homologues au gène, tout comme ce qui est observé dans le cas des délétions de 37 kb (Figure 60).

Pour réconcilier ces deux observations, nous avons utilisé la technique du peignage moléculaire. L'analyse de l'ADN de la patiente par FISH sur ADN étiré a montré un profil semblable à celui des témoins, confirmant l'absence de délétion et rendant vraisemblable l'hypothèse d'une conversion génique (Figure 61). Le caractère pathogène de cet évènement reste à démontrer.

En conclusion, l'utilisation d'au moins deux techniques différentes permettant de mettre en évidence des réarrangements est indispensable, surtout pour ceux concernant la région située en amont du gène *BRCA1* où la présence d'un pseudogène favorise non seulement les délétions et les duplications, mais peut également favoriser les conversions.

Figure 62: Analyse par peignage moléculaire de 6 individus identifiés par Schaap *et al.* comme présentant une instabilité au niveau du locus *RNU2*.

2 Mise en évidence d'un mosaïcisme au niveau du locus *RNU2* et Identification d'allèles complexes du CNV *RNU2*

L'étude du nombre allélique de copies du CNV *RNU2*, conduite par Schaap et ses collaborateurs sur 210 individus et utilisant l'électrophorèse en champs pulsés, leur a permis d'identifier 6 individus présentant un profil particulier et donc de mettre en évidence une instabilité de ce locus :

- 1 individu ayant hérité d'un allèle maternel avec une expansion *de novo* d'une copie (14 à 15 copies)
- 4 individus porteurs de trois allèles avec des nombres de copies différents.
- 1 individu pour lequel le CNV *RNU2* présente un profil complexe avec plus de quatre allèles pouvant être identifiés.

Ces individus appartenant tous au projet HapMap, et donc au projet 1000 Génomes, des lignées lymphoblastoides sont disponibles pour chacun d'entre eux. Nous avons donc décidé de tirer profit de notre code-barres *RNU2* et de la précision de la technique du peignage moléculaire pour valider ses observations (Figure 62).

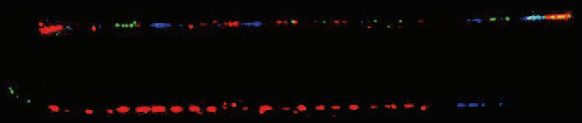
D'après Schaap *et al.*, l'individu NA19103 est porteur d'un allèle à 6 copies et d'un allèle à 15 copies. Par peignage moléculaire, j'ai retrouvé l'allèle à 6 copies, cependant je n'ai pas identifié l'allèle à 15 copies. A la place, j'ai identifié deux allèles : un premier, majoritaire, porteur de 13 copies et un second porteur de 16 copies. Pour confirmer l'expansion *de novo* d'une seule copie, il est nécessaire d'analyser par peignage moléculaire l'ADN de la mère.

Pour l'individu NA18862, mes observations sont en accord avec celles de Schaap *et al.* : cet individu est porteur de trois allèles avec respectivement 8, 12 et 13

- GM18593 – Lame 374 et lame 378
PFGE: 20, 22 et 25 (mosaic 22 75% et 25 25%)

Allèle 1: 20

Allèle 2: 22

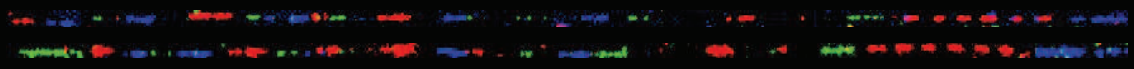


Allèle ?: 15

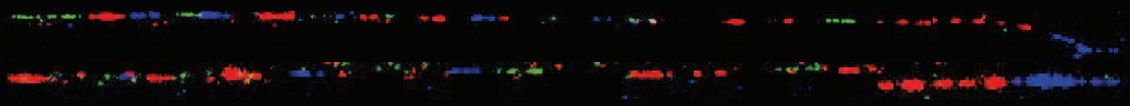
Allèle ? : 19

- GM18949 – Lame 373 et 390
PFGE: 7, 7 et 28 (mosaic 28 10% et 7 90%)

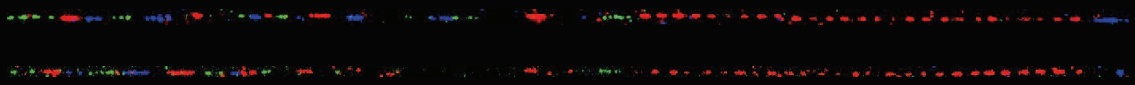
Allèle 1: 6 (56%)



Allèle 2: 7 (13%)

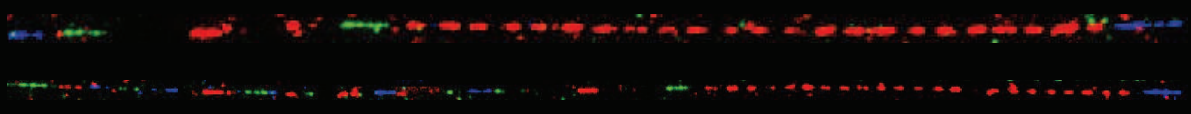


Allèle 3: 27 (30%)

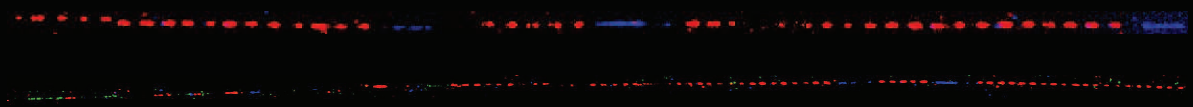


- GM18960 – Lame 38-2
PFGE: 5, 19, 22, 34 (complex)
DOC: 47,98

Allèle 1: 23



Allèle 2: 35/5/20

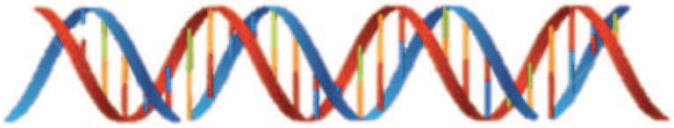


copies. Pour l'individu NA18949, j'ai également retrouvé trois allèles (6, 7 et 27 copies) précisant ainsi les données obtenues par Schaap *et al.* (7, 7 et 28 copies). De même, l'individu NA19142 est porteur de 23, 24 et 36 copies (contre 24, 25 et 37). Pour l'individu NA18593, j'ai également retrouvé les allèles porteurs de 20 et 22 copies qu'avait identifié Schaap *et al.*, mais je n'ai pas retrouvé l'allèle à 25 copies. J'ai cependant observé deux autres signaux : l'un porteur de 15 copies et l'autre de 19 copies.

Enfin, chez l'individu NA18960, Schaap *et al.* ont identifié quatre allèles : 5, 19, 22 et 34 copies. Par peignage moléculaire, j'ai identifié un allèle porteur de 23 répétitions. Le second allèle est porteur de 35 répétitions, puis de 5 répétitions et enfin de 20 répétitions, chacun de ces *sous-CNV* étant séparé par des séquences uniques (sonde bleue).

Cette analyse confirme la complémentarité des deux approches. L'électrophorèse en champs pulsés permet d'identifier rapidement des individus présentant un mosaïcisme au locus *RNU2*, ainsi que le nombre différent d'allèles, ce qui est plus délicat par peignage moléculaire. En effet, il faut analyser un nombre considérable de lamelles peignées avant de récolter suffisamment de signaux provenant d'un allèle rare. Cependant, le peignage moléculaire permet une quantification plus précise du nombre allélique de copies, et surtout l'identification d'allèles complexes au locus *RNU2*. Il serait maintenant très intéressant d'étudier le locus complexe identifié chez l'individu NA18960, afin de comprendre quels évènements ont conduit à la *triplication* du CNV.

Références



1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Abel, K.J., Xu, J., Yin, G.Y., Lyons, R.H., Meisler, M.H., and Weber, B.L. (1995). Mouse Brca1: localization sequence analysis and identification of evolutionarily conserved domains. Hum. Mol. Genet. *4*, 2265–2273.

Adams, D.J., Dermitzakis, E.T., Cox, T., Smith, J., Davies, R., Banerjee, R., Bonfield, J., Mullikin, J.C., Chung, Y.J., Rogers, J., et al. (2005). Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. Nat. Genet. *37*, 532–536.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Autom. Control *19*, 716–723.

Aldhous, M.C., Abu Bakar, S., Prescott, N.J., Palla, R., Soo, K., Mansfield, J.C., Mathew, C.G., Satsangi, J., and Armour, J.A.L. (2010). Measurement methods and accuracy in copy number variation: failure to replicate associations of beta-defensin copy number with Crohn's disease. Hum. Mol. Genet. *19*, 4930–4938.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Allen, N.E., Beral, V., Casabonne, D., Kan, S.W., Reeves, G.K., Brown, A., Green, J., and Million Women Study Collaborators (2009). Moderate alcohol intake and cancer incidence in women. J. Natl. Cancer Inst. *101*, 296–305.

Anderson, S.F., Schlegel, B.P., Nakajima, T., Wolpin, E.S., and Parvin, J.D. (1998). BRCA1 protein is linked to the RNA polymerase II holoenzyme complex via RNA helicase A. Nat. Genet. *19*, 254–256.

Antoniou, A.C., Pharoah, P.D.P., McMullan, G., Day, N.E., Stratton, M.R., Peto, J., Ponder, B.J., and Easton, D.F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. Br. J. Cancer *86*, 76–83.

Antoniou, A.C., Spurdle, A.B., Sinilnikova, O.M., Healey, S., Pooley, K.A., Schmutzler, R.K., Versmold, B., Engel, C., Meindl, A., Arnold, N., et al. (2008). Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. Am. J. Hum. Genet. *82*, 937–948.

Aprelikova, O.N., Fang, B.S., Meissner, E.G., Cotter, S., Campbell, M., Kuthiala, A., Bessho, M., Jensen, R.A., and Liu, E.T. (1999). BRCA1-associated growth arrest is RB-dependent. Proc. Natl. Acad. Sci. U. S. A. *96*, 11866–11871.

Ares, M., Jr, Chung, J.S., Giglio, L., and Weiner, A.M. (1987). Distinct factors with Sp1 and NF-A specificities bind to adjacent functional elements of the human U2 snRNA gene enhancer. Genes Dev. *1*, 808–817.

Armour, J.A.L., Palla, R., Zeeuwen, P.L.J.M., Heijer, M. d., Schalkwijk, J., and Hollox, E.J. (2007). Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats.

Nucleic Acids Res. *35*, e19–e19.

Van Arsdell, S.W., and Weiner, A.M. (1984). Human genes for U2 small nuclear RNA are tandemly repeated. Mol. Cell. Biol. *4*, 492–499.

Assaad, F.F., Tucker, K.L., and Signer, E.R. (1993). Epigenetic repeat-induced gene silencing (RIGS) in Arabidopsis. Plant Mol. Biol. *22*, 1067–1085.

Atlas, E., Stramwasser, M., Whiskin, K., and Mueller, C.R. (2000). GA-binding protein alpha/beta is a critical regulator of the BRCA1 promoter. Oncogene *19*, 1933–1940.

Atlas, E., Stramwasser, M., and Mueller, C.R. (2001). A CREB site in the BRCA1 proximal promoter acts as a constitutive transcriptional element. Oncogene *20*, 7110–7114.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types : induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J. Exp. Med. *79*, 137–158.

Bahcall, O.G. (2013). iCOGS collection provides a collaborative model. Foreword. Nat. Genet. *45*, 343.

Bailey, A.D., Li, Z., Pavelitz, T., and Weiner, A.M. (1995). Adenovirus type 12-induced fragility of the human RNU2 locus requires U2 small nuclear RNA transcriptional regulatory elements. Mol. Cell. Biol. *15*, 6246–6255.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003–1007.

Baillat, D., Hakimi, M.-A., Näär, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell *123*, 265–276.

Balog, J., Miller, D., Sanchez-Curtailles, E., Carbo-Marques, J., Block, G., Potman, M., de Knijff, P., Lemmers, R.J.L.F., Tapscott, S.J., and van der Maarel, S.M. (2012). Epigenetic regulation of the X-chromosomal macrosatellite repeat encoding for the cancer/testis gene CT47. Eur. J. Hum. Genet. EJHG *20*, 185–191.

Baraniskin, A., Nöpel-Dünnebacke, S., Ahrens, M., Jensen, S.G., Zöllner, H., Maghnouj, A., Wos, A., Mayerle, J., Munding, J., Kost, D., et al. (2013). Circulating U2 small nuclear RNA fragments as a novel diagnostic biomarker for pancreatic and colorectal adenocarcinoma. Int. J. Cancer J. Int. Cancer *132*, E48–57.

Baraniskin, A., Nöpel-Dünnebacke, S., Schumacher, B., Gerges, C., Bracht, T., Sitek, B., Meyer, H.E., Gerken, G., Dechene, A., Schlaak, J.F., et al. (2014). Analysis of U2 Small Nuclear RNA Fragments in the Bile Differentiates Cholangiocarcinoma from Primary Sclerosing Cholangitis and Other Benign Biliary Disorders. Dig. Dis. Sci.

Bardoni, B., and Mandel, J.-L. (2002). Advances in understanding of fragile X pathogenesis and FMRP function, and in identification of X linked mental retardation genes. Curr. Opin. Genet. Dev. *12*, 284–293.

Bauman, J.G., Wiegant, J., Borst, P., and van Duijn, P. (1980). A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochromelabelled RNA. Exp. Cell Res. *128*, 485–490.

Bellanné-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufils, S., Bertrand, S., Georges, I., Glibert, F., Gros, I., and Lucotte, G. (1992). Mapping the whole human genome by fingerprinting yeast artificial chromosomes. Cell *70*, 1059–1068.

Bensimon, A., Simon, A., Chiffaudel, A., Croquette, V., Heslot, F., and Bensimon, D. (1994). Alignment and sensitive detection of DNA by a moving interface. Science *265*, 2096–2098.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Bentley, R.W., Pearson, J., Gearry, R.B., Barclay, M.L., McKinney, C., Merriman, T.R., and Roberts, R.L. (2010). Association of higher DEFB4 genomic copy number with Crohn's disease. Am. J. Gastroenterol. *105*, 354–359.

Bernstein, L.B., Manser, T., and Weiner, A.M. (1985). Human U1 small nuclear RNA genes: extensive conservation of flanking sequences suggests cycles of gene amplification and transposition. Mol. Cell. Biol. *5*, 2159–2171.

Bhangale, T.R., Stephens, M., and Nickerson, D.A. (2006). Automating resequencing-based detection of insertion-deletion polymorphisms. Nat. Genet. *38*, 1457–1462.

Binder-Foucard, F., Belot, A., Delafosse, P., Remontet, L., Woronoff, A., and Bossard, N. (2013). Estimation nationale de l'incidence et de la mortalité par cancer en France entre 1980 et 2012. Partie 1 – Tumeurs solides. 122 p.

Blunt, T., Steers, F., Daniels, G., and Carritt, B. (1994). Lack of RH C/E expression in the Rhesus D--phenotype is the result of a gene deletion. Ann. Hum. Genet. *58*, 19–24.

Bochar, D.A., Wang, L., Beniya, H., Kinev, A., Xue, Y., Lane, W.S., Wang, W., Kashanchi, F., and Shiekhattar, R. (2000). BRCA1 is associated with a human SWI/SNF-related complex: linking chromatin remodeling to breast cancer. Cell *102*, 257–265.

Bois, P., Stead, J.D., Bakshi, S., Williamson, J., Neumann, R., Moghadaszadeh, B., and Jeffreys, A.J. (1998). Isolation and characterization of mouse minisatellites. Genomics *50*, 317–330.

Bojesen, S.E., Pooley, K.A., Johnatty, S.E., Beesley, J., Michailidou, K., Tyrer, J.P., Edwards, S.L., Pickett, H.A., Shen, H.C., Smart, C.E., et al. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat. Genet. *45*, 371–384, 384e1–2.

Bonnen, P.E., Wang, P.J., Kimmel, M., Chakraborty, R., and Nelson, D.L. (2002). Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res. *12*, 1846–1853.

Boutry-Kryza, N., Labalme, A., Till, M., Schluth-Bolard, C., Langue, J., Turleau, C., Edery, P., and Sanlaville, D. (2012). An 800 kb deletion at 17q23.2 including the MED13 (THRAP1) gene, revealed by aCGH in a patient with a SMC 17p. Am. J. Med. Genet. A. *158A*, 400–405.

Branlant, C., Krol, A., Ebel, J.P., Lazar, E., Haendler, B., and Jacob, M. (1982). U2 RNA shares a

structural domain with U1, U4, and U5 RNAs. EMBO J. *1*, 1259–1265.

Bray, F., Ren, J.-S., Masuyer, E., and Ferlay, J. (2013). Global estimates of cancer prevalence for 27 sites in the adult population in 2008. Int. J. Cancer *132*, 1133–1145.

Brennwald, P., Porter, G., and Wise, J.A. (1988). U2 small nuclear RNA is remarkably conserved between Schizosaccharomyces pombe and mammals. Mol. Cell. Biol. *8*, 5575–5580.

Broca, P. (1866). Traité des Tumeurs (P. Asselin).

Brock, G.J., and Bird, A. (1997). Mosaic methylation of the repeat unit of the human ribosomal RNA genes. Hum. Mol. Genet. *6*, 451–456.

Brown, K.H., Dobrinski, K.P., Lee, A.S., Gokcumen, O., Mills, R.E., Shi, X., Chong, W.W.S., Chen, J.Y.H., Yoo, P., David, S., et al. (2012). Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. Proc. Natl. Acad. Sci. U. S. A. *109*, 529–534.

Brown, M.A., Xu, C.F., Nicolai, H., Griffiths, B., Chambers, J.A., Black, D., and Solomon, E. (1996). The 5′ end of the BRCA1 gene lies within a duplicated region of human chromosome 17q21. Oncogene *12*, 2507–2513.

Bruce, H.A., Sachs, N., Rudnicki, D.D., Lin, S.G., Willour, V.L., Cowell, J.K., Conroy, J., McQuaid, D.E., Rossi, M., Gaile, D.P., et al. (2009). Long tandem repeats as a form of genomic copy number variation: structure and length polymorphism of a chromosome 5p repeat in control and schizophrenia populations. Psychiatr. Genet. *19*, 64–71.

Brunet, J., Vazquez-Martin, A., Colomer, R., Graña-Suarez, B., Martin-Castillo, B., and Menendez, J.A. (2008). BRCA1 and acetyl-CoA carboxylase: the metabolic syndrome of breast cancer. Mol. Carcinog. *47*, 157–163.

Buist, D.S., LaCroix, A.Z., Barlow, W.E., White, E., Cauley, J.A., Bauer, D.C., and Weiss, N.S. (2001). Bone mineral density and endogenous hormones and risk of breast cancer in postmenopausal women (United States). Cancer Causes Control CCC *12*, 213–222.

Burrows, J.F., Scott, C.J., and Johnston, J.A. (2010). The DUB/USP17 deubiquitinating enzymes: a gene family within a tandemly repeated sequence, is also embedded within the copy number variable beta-defensin cluster. BMC Genomics *11*, 250.

Buysse, K., Antonacci, F., Callewaert, B., Loeys, B., Fränkel, U., Siu, V., Mortier, G., Speleman, F., and Menten, B. (2009). Unusual 8p inverted duplication deletion with telomere capture from 8q. Eur. J. Med. Genet. *52*, 31–36.

Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y., and Gabellini, D. (2012). A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in FSHD muscular dystrophy. Cell *149*, 819–831.

Le Calvez-Kelm, F., Oliver, J., Damiola, F., Forey, N., Robinot, N., Durand, G., Voegele, C., Vallée, M.P., Byrnes, G., Registry, B.C.F., et al. (2012). RAD51 and breast cancer susceptibility: no evidence for rare variant association in the Breast Cancer Family Registry study. PloS One *7*, e52374.

Campbell, N.A., Reece, Mitchell, L.G., and Benjamin/Cummings Publishing Company (1999).

Biology. [Gr. 9-12] [Gr. 9-12].

Campbell, R.D., Dunham, I., Kendall, E., and Sargent, C.A. (1990). Polymorphism of the human complement component C4. Exp. Clin. Immunogenet. *7*, 69–84.

De Capoa, A., Menendez, F., Poggesi, I., Giancotti, P., Grappelli, C., Marotta, M.R., Di Leandro, M., Reynaud, C., and Niveleau, A. (1996). Cytological evidence for 5-azacytidine-induced demethylation of the heterochromatic regions of human chromosomes. Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. *4*, 271–276.

Card, C.O., Morris, G.F., Brown, D.T., and Marzluff, W.F. (1982). Sea urchin small nuclear RNA genes are organized in distinct tandemly repeating units. Nucleic Acids Res. *10*, 7677–7688.

Carter, N.P. (2004). As normal as normal can be? Nat. Genet. *36*, 931–932.

Carter, N.P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. *39*, S16–21.

Centre International de Recherche sur le Cancer, L'Académie Nationale de Médecine, L'Académie Nationale des Sciences, and La Fédération Nationale des Cent (2007). Les Causes du Cancer en France.

Chadwick, B.P. (2008). DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. Genome Res. *18*, 1259–1269.

Chargaff, E., Magasanik, B., Vischer, E., Green, C., Doniger, R., and Elson, D. (1950). Nucleotide composition of pentose nucleic acids from yeast and mammalian tissues. J. Biol. Chem. *186*, 51–67.

Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature *371*, 215–220.

Cheeseman, K., Rouleau, E., Vannier, A., Thomas, A., Briaux, A., Lefol, C., Walrafen, P., Bensimon, A., Lidereau, R., Conseiller, E., et al. (2012). A diagnostic genetic test for the physical mapping of germline rearrangements in the susceptibility breast cancer genes BRCA1 and BRCA2. Hum. Mutat. *33*, 998–1009.

Chen, Y.-T., Iseli, C., Venditti, C.A., Old, L.J., Simpson, A.J.G., and Jongeneel, C.V. (2006). Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. Genes. Chromosomes Cancer *45*, 392–400.

Choy, K.W., Setlur, S.R., Lee, C., and Lau, T.K. (2010). The impact of human copy number variation on a new era of genetic testing. BJOG Int. J. Obstet. Gynaecol. *117*, 391–398.

Claus, E.B., Risch, N., and Thompson, W.D. (1991). Genetic analysis of breast cancer in the cancer and steroid hormone study. Am. J. Hum. Genet. *48*, 232–242.

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat. Genet. *37*, 1243–1246.

Coene, E.D., Hollinshead, M.S., Waeytens, A.A.T., Schelfhout, V.R.J., Eechaute, W.P., Shaw, M.K., Van Oostveldt, P.M.V., and Vaux, D.J. (2005). Phosphorylated BRCA1 is predominantly located in

the nucleus and mitochondria. Mol. Biol. Cell *16*, 997–1010.

Cole, C., Sobala, A., Lu, C., Thatcher, S.R., Bowman, A., Brown, J.W.S., Green, P.J., Barton, G.J., and Hutvagner, G. (2009). Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA N. Y. N *15*, 2147–2160.

Collaborative Group on Hormonal Factors in Breast Cancer (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. Lancet *360*, 187–195.

COMPLEXO, Southey, M.C., Park, D.J., Nguyen-Dumont, T., Campbell, I., Thompson, E., Trainer, A.H., Chenevix-Trench, G., Simard, J., Dumont, M., et al. (2013). COMPLEXO: identifying the missing heritability of breast cancer via next generation collaboration. Breast Cancer Res. BCR *15*, 402.

Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. Nat. Genet. *38*, 75–81.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

Cooper, G.M., Nickerson, D.A., and Eichler, E.E. (2007). Mutational and selective effects on copy-number variants in the human genome. Nat. Genet. *39*, S22–29.

Cox, A., Dunning, A.M., Garcia-Closas, M., Balasubramanian, S., Reed, M.W.R., Pooley, K.A., Scollen, S., Baynes, C., Ponder, B.A.J., Chanock, S., et al. (2007). A common coding variant in CASP8 is associated with breast cancer risk. Nat. Genet. *39*, 352–358.

Cuello, P., Boyd, D.C., Dye, M.J., Proudfoot, N.J., and Murphy, S. (1999). Transcription of the human U2 snRNA genes continues beyond the 3′ box in vivo. EMBO J. *18*, 2867–2877.

Dacheux, E., Vincent, A., Nazaret, N., Combet, C., Wierinckx, A., Mazoyer, S., Diaz, J.-J., Lachuer, J., and Venezia, N.D. (2013). BRCA1-Dependent Translational Regulation in Breast Cancer Cells. PloS One *8*, e67313.

Dahlberg, J.E., and Lund, E. (1988). The Genes and Transcription of the Major Small Nuclear RNAs. In Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles, M.L. Birnstiel, ed. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 38–70.

Darzacq, X., Jády, B.E., Verheggen, C., Kiss, A.M., Bertrand, E., and Kiss, T. (2002). Cajal body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylation guide RNAs. EMBO J. *21*, 2746–2756.

Dawson, E., Chen, Y., Hunt, S., Smink, L.J., Hunt, A., Rice, K., Livingston, S., Bumpstead, S., Bruskiewich, R., Sham, P., et al. (2001). A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. Genome Res. *11*, 170–178.

Delgrange, O., and Rivals, E. (2004). STAR: an algorithm to Search for Tandem Approximate Repeats. Bioinforma. Oxf. Engl. *20*, 2812–2820.

Denison, R.A., Van Arsdell, S.W., Bernstein, L.B., and Weiner, A.M. (1981). Abundant

pseudogenes for small nuclear RNAs are dispersed in the human genome. Proc. Natl. Acad. Sci. U. S. A. *78*, 810–814.

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., et al. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature *380*, 152–154.

Dopman, E.B., and Hartl, D.L. (2007). A portrait of copy-number polymorphism in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *104*, 19920–19925.

Dorer, D.R., and Henikoff, S. (1994). Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. Cell *77*, 993–1002.

Le Dran, H. (1757). Mémoire avec un précis de plusieurs observations sur le cancer.

Durnam, D.M., Menninger, J.C., Chandler, S.H., Smith, P.P., and McDougall, J.K. (1988). A fragile site in the human U2 small nuclear RNA gene cluster is revealed by adenovirus type 12 infection. Mol. Cell. Biol. *8*, 1863–1867.

Van Duyvenvoorde, H.A., Lui, J.C., Kant, S.G., Oostdijk, W., Gijsbers, A.C., Hoffer, M.J., Karperien, M., Walenkamp, M.J., Noordam, C., Voorhoeve, P.G., et al. (2013). Copy number variants in patients with short stature. Eur. J. Hum. Genet. EJHG.

Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D.P., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R., et al. (2007a). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature *447*, 1087–1093.

Easton, D.F., Deffenbaugh, A.M., Pruss, D., Frye, C., Wenstrup, R.J., Allen-Brady, K., Tavtigian, S.V., Monteiro, A.N.A., Iversen, E.S., Couch, F.J., et al. (2007b). A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am. J. Hum. Genet. *81*, 873–883.

Egan, C.M., Sridhar, S., Wigler, M., and Hall, I.M. (2007). Recurrent DNA copy number variation in the laboratory mouse. Nat. Genet. *39*, 1384–1389.

Egloff, S., O'Reilly, D., and Murphy, S. (2008). Expression of human snRNA genes from beginning to end. Biochem. Soc. Trans. *36*, 590–594.

Egloff, S., Al-Rawaf, H., O'Reilly, D., and Murphy, S. (2009). Chromatin structure is implicated in "late" elongation checkpoints on the U2 snRNA and beta-actin genes. Mol. Cell. Biol. *29*, 4002–4013.

Eichler, E.E., Clark, R.A., and She, X. (2004). An assessment of the sequence gaps: unfinished business in a finished human genome. Nat. Rev. Genet. *5*, 345–354.

Eliceiri, G.L., and Sayavedra, M.S. (1976). Small RNAs in the nucleus and cytoplasm of HeLa cells. Biochem. Biophys. Res. Commun. *72*, 507–512.

Emerson, J.J., Cardoso-Moreira, M., Borevitz, J.O., and Long, M. (2008). Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. Science *320*, 1629–1631.

Evans, D.G.R., Birch, J.M., Thorneycroft, M., McGown, G., Lalloo, F., and Varley, J.M. (2002). Low rate of TP53 germline mutations in breast cancer/sarcoma families not fulfilling classical criteria for

Li-Fraumeni syndrome. J. Med. Genet. *39*, 941–944.

Ewertz, M., Duffy, S.W., Adami, H.O., Kvåle, G., Lund, E., Meirik, O., Mellemgaard, A., Soini, I., and Tulinius, H. (1990). Age at first birth, parity and risk of breast cancer: a meta-analysis of 8 studies from the Nordic countries. Int. J. Cancer J. Int. Cancer *46*, 597–603.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., and Bray, F. (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Http://globocan.iarc.fr.

Fernandez-Jimenez, N., Castellanos-Rubio, A., Plaza-Izurieta, L., Gutierrez, G., Irastorza, I., Castaño, L., Vitoria, J.C., and Bilbao, J.R. (2011). Accuracy in Copy Number Calling by qPCR and PRT: A Matter of DNA. PLoS ONE *6*, e28910.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat. Rev. Genet. *7*, 85–97.

Flores, M., Morales, L., Gonzaga-Jauregui, C., Dominguez-Vidana, R., Zepeda, C., Yanez, O., Gutierrez, M., Lemus, T., Valle, D., Avila, M.C., et al. (2007). Recurrent DNA inversion rearrangements in the human genome. Proc. Natl. Acad. Sci. *104*, 6099–6106.

Florijn, R.J., Bonden, L.A., Vrolijk, H., Wiegant, J., Vaandrager, J.W., Baas, F., den Dunnen, J.T., Tanke, H.J., van Ommen, G.J., and Raap, A.K. (1995). High-resolution DNA Fiber-FISH for genomic DNA mapping and colour bar-coding of large genes. Hum. Mol. Genet. *4*, 831–836.

Freedman, M.L., Penney, K.L., Stram, D.O., Riley, S., McKean-Cowdin, R., Le Marchand, L., Altshuler, D., and Haiman, C.A. (2005). A haplotype-based case-control study of BRCA1 and sporadic breast cancer risk. Cancer Res. *65*, 7516–7522.

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. (2006). Copy number variation: new insights in genome diversity. Genome Res. *16*, 949–961.

Friedenreich, C.M., and Cust, A.E. (2008). Physical activity and breast cancer risk: impact of timing, type and dose of activity and population subgroup effects. Br. J. Sports Med. *42*, 636–647.

Fu, H., Feng, J., Liu, Q., Sun, F., Tie, Y., Zhu, J., Xing, R., Sun, Z., and Zheng, X. (2009). Stress induces tRNA cleavage by angiogenin in mammalian cells. FEBS Lett. *583*, 437–442.

Gad, S., Scheuner, M.T., Pages-Berhouet, S., Caux-Moncoutier, V., Bensimon, A., Aurias, A., Pinto, M., and Stoppa-Lyonnet, D. (2001a). Identification of a large rearrangement of the BRCA1 gene using colour bar code on combed DNA in an American breast/ovarian cancer family previously studied by direct sequencing. J. Med. Genet. *38*, 388–392.

Gad, S., Aurias, A., Puget, N., Mairal, A., Schurra, C., Montagna, M., Pages, S., Caux, V., Mazoyer, S., Bensimon, A., et al. (2001b). Color bar coding the BRCA1 gene on combed DNA: a useful strategy for detecting large gene rearrangements. Genes. Chromosomes Cancer *31*, 75–84.

Gad, S., Klinger, M., Caux-Moncoutier, V., Pages-Berhouet, S., Gauthier-Villars, M., Coupier, I., Bensimon, A., Aurias, A., and Stoppa-Lyonnet, D. (2002a). Bar code screening on combed DNA for large rearrangements of the BRCA1 and BRCA2 genes in French breast cancer families. J. Med. Genet. *39*,

817–821.

Gad, S., Caux-Moncoutier, V., Pagès-Berhouet, S., Gauthier-Villars, M., Coupier, I., Pujol, P., Frénay, M., Gilbert, B., Maugard, C., Bignon, Y.-J., et al. (2002b). Significant contribution of large BRCA1 gene rearrangements in 120 French breast and ovarian cancer families. Oncogene *21*, 6841–6847.

Gai, X., Perin, J.C., Murphy, K., O'Hara, R., D'arcy, M., Wenocur, A., Xie, H.M., Rappaport, E.F., Shaikh, T.H., and White, P.S. (2010). CNV Workshop: an integrated platform for high-throughput copy number variation discovery and clinical diagnostics. BMC Bioinformatics *11*, 74.

Ganley, A.R.D., and Kobayashi, T. (2007). Highly efficient concerted evolution in the ribosomal DNA repeats: total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res. *17*, 184–191.

Gargano, S., Wang, P., Rusanganwa, E., and Bacchetti, S. (1995). The transcriptionally competent U2 gene is necessary and sufficient for adenovirus type 12 induction of the fragile site at 17q21-22. Mol. Cell. Biol. *15*, 6256–6261.

Garrick, D., Fiering, S., Martin, D.I.K., and Whitelaw, E. (1998). Repeat-induced gene silencing in mammals. Nat. Genet. *18*, 56–59.

Gatchel, J.R., and Zoghbi, H.Y. (2005). Diseases of unstable repeat expansion: mechanisms and common principles. Nat. Rev. Genet. *6*, 743–755.

Van Geel, M., Heather, L.J., Lyle, R., Hewitt, J.E., Frants, R.R., and de Jong, P.J. (1999). The FSHD region on human chromosome 4q35 contains potential coding regions among pseudogenes and a high density of repeat elements. Genomics *61*, 55–65.

Gelfand, Y., Rodriguez, A., and Benson, G. (2007). TRDB--the Tandem Repeats Database. Nucleic Acids Res. *35*, D80–87.

Gilles, F., Goy, A., Remache, Y., Manova, K., and Zelenetz, A.D. (2000). Cloning and characterization of a Golgin-related gene from the large-scale polymorphism linked to the PML gene. Genomics *70*, 364–374.

Glade, M.J. (1999). Food, nutrition, and the prevention of cancer: a global perspective. American Institute for Cancer Research/World Cancer Research Fund, American Institute for Cancer Research, 1997. Nutr. Burbank Los Angel. Cty. Calif *15*, 523–526.

Gondo, Y., Okada, T., Matsuyama, N., Saitoh, Y., Yanagisawa, Y., and Ikeda, J.E. (1998). Human megasatellite DNA RS447: copy-number polymorphisms and interspecies conservation. Genomics *54*, 39–49.

Gonzalez, I.L., and Sylvester, J.E. (2001). Human rDNA: evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. Genomics *73*, 255–263.

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science *307*, 1434–1440.

Gowen, L.C., Johnson, B.L., Latour, A.M., Sulik, K.K., and Koller, B.H. (1996). Brca1 deficiency results in early embryonic lethality characterized by neuroepithelial abnormalities. Nat. Genet. *12*, 191–

194.

De Greef, J.C., Frants, R.R., and van der Maarel, S.M. (2008). Epigenetic mechanisms of facioscapulohumeral muscular dystrophy. Mutat. Res. *647*, 94–102.

Gripenberg, U. (1964). SIZE VARIATION AND ORIENTATION OF THE HUMAN Y CHROMOSOME. Chromosoma *15*, 618–629.

Gudbjartsson, D.F., Walters, G.B., Thorleifsson, G., Stefansson, H., Halldorsson, B.V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., et al. (2008). Many sequence variants affecting diversity of adult human height. Nat. Genet. *40*, 609–615.

Guescini, M., Sisti, D., Rocchi, M.B., Stocchi, L., and Stocchi, V. (2008). A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. BMC Bioinformatics *9*, 326.

Gulcher, J. (2012). Microsatellite markers for linkage and association studies. Cold Spring Harb. Protoc. *2012*, 425–432.

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S.A.A.C., Cook, S., Pravenec, M., Aitman, T., Jacob, H., Shull, J.D., et al. (2008). Distribution and functional impact of DNA copy number variation in the rat. Nat. Genet. *40*, 538–545.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M., and Weissenbach, J. (1994). The 1993-94 Généthon human genetic linkage map. Nat. Genet. *7*, 246–339.

Haaf, T., and Schmid, M. (2000). Experimental condensation inhibition in constitutive and facultative heterochromatin of mammalian chromosomes. Cytogenet. Cell Genet. *91*, 113–123.

Hakem, R., de la Pompa, J.L., Sirard, C., Mo, R., Woo, M., Hakem, A., Wakeham, A., Potter, J., Reitmair, A., Billia, F., et al. (1996). The tumor suppressor gene Brca1 is required for embryonic cellular proliferation in the mouse. Cell *85*, 1009–1023.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., and King, M.C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science *250*, 1684–1689.

Hamada, H., Seidman, M., Howard, B.H., and Gorman, C.M. (1984). Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. Mol. Cell. Biol. *4*, 2622–2630.

Hammarström, K., Westin, G., Bark, C., Zabielski, J., and Petterson, U. (1984). Genes and pseudogenes for human U2 RNA. Implications for the mechanism of pseudogene formation. J. Mol. Biol. *179*, 157–169.

Hannan, A.J. (2010). Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for "missing heritability."Trends Genet. TIG *26*, 59–65.

Hashizume, R., Fukuda, M., Maeda, I., Nishikawa, H., Oyake, D., Yabuki, Y., Ogata, H., and Ohta, T. (2001). The RING heterodimer BRCA1-BARD1 is a ubiquitin ligase inactivated by a breast cancer-derived mutation. J. Biol. Chem. *276*, 14537–14540.

Hassold, T., Abruzzo, M., Adkins, K., Griffin, D., Merrill, M., Millie, E., Saker, D., Shen, J., and Zaragoza, M. (1996). Human aneuploidy: incidence, origin, and etiology. Environ. Mol. Mutagen. *28*,

167–175.

Hastings, P.J., Ira, G., and Lupski, J.R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet. *5*, e1000327.

Hedrick, P.W. (2011). Population genetics of malaria resistance in humans. Heredity *107*, 283–304.

Henderson, B.R. (2005). Regulation of BRCA1, BRCA2 and BARD1 intracellular trafficking. BioEssays News Rev. Mol. Cell. Dev. Biol. *27*, 884–893.

Hernandez, N. (2001). Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. J. Biol. Chem. *276*, 26733–26736.

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. *36*, 39–56.

Hillis, D.M., and Dixon, M.T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. Q. Rev. Biol. *66*, 411–453.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U. S. A. *106*, 9362–9367.

Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072–1079.

Hinds, D.A., Kloek, A.P., Jen, M., Chen, X., and Frazer, K.A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat. Genet. *38*, 82–85.

Hollox, E.J. (2008). Copy number variation of beta-defensins and relevance to disease. Cytogenet. Genome Res. *123*, 148–155.

Hollox, E.J., Armour, J.A.L., and Barber, J.C.K. (2003). Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. Am. J. Hum. Genet. *73*, 591–600.

Htun, H., Lund, E., Westin, G., Pettersson, U., and Dahlberg, J.E. (1985). Nuclease S1-sensitive sites in multigene families: human U2 small nuclear RNA genes. EMBO J. *4*, 1839–1845.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. *9*, 868–877.

Huang, Q., Jacobson, M.R., and Pederson, T. (1997). 3′ processing of human pre-U2 small nuclear RNA: a base-pairing interaction between the 3′ extension of the precursor and an internal region. Mol. Cell. Biol. *17*, 7178–7185.

Huber, J., Cronshagen, U., Kadokura, M., Marshallsay, C., Wada, T., Sekine, M., and Lührmann, R. (1998). Snurportin1, an m3G-cap-specific nuclear import receptor with a novel domain structure. EMBO J. *17*, 4114–4126.

Hurles, M. (2005). How homologous recombination generates a mutable genome. Hum. Genomics *2*, 179–186.

Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and

Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*, 949–951.

Ide, S., Miyazaki, T., Maki, H., and Kobayashi, T. (2010). Abundance of Ribosomal RNA Gene Copies Maintains Genome Integrity. Science *327*, 693–696.

International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789–796.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Jacobs, E.Y., Ogiwara, I., and Weiner, A.M. (2004). Role of the C-terminal domain of RNA polymerase II in U2 snRNA transcription and 3′ processing. Mol. Cell. Biol. *24*, 846–855.

Jeffreys, A.J., Wilson, V., and Thein, S.L. (1985). Hypervariable "minisatellite" regions in human DNA. Nature *314*, 67–73.

Jeffreys, A.J., Allen, M.J., Armour, J.A., Collick, A., Dubrova, Y., Fretwell, N., Guram, T., Jobling, M., May, C.A., and Neil, D.L. (1995). Mutation processes at human minisatellites. Electrophoresis *16*, 1577–1585.

Jia, Y., Mu, J.C., and Ackerman, S.L. (2012). Mutation of a U2 snRNA gene causes global disruption of alternative splicing and neurodegeneration. Cell *148*, 296–308.

Jiang, C., and Liao, D. (1999). Striking bimodal methylation of the repeat unit of the tandem array encoding human U2 snRNA (the RNU2 locus). Genomics *62*, 508–518.

John, E.M., Hopper, J.L., Beck, J.C., Knight, J.A., Neuhausen, S.L., Senie, R.T., Ziogas, A., Andrulis, I.L., Anton-Culver, H., Boyd, N., et al. (2004). The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. Breast Cancer Res. BCR *6*, R375–389.

Kallioniemi, O.P., Kallioniemi, A., Kurisu, W., Thor, A., Chen, L.C., Smith, H.S., Waldman, F.M., Pinkel, D., and Gray, J.W. (1992). ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. Proc. Natl. Acad. Sci. U. S. A. *89*, 5321–5325.

Kan, Y.W., and Dozy, A.M. (1978). Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proc. Natl. Acad. Sci. U. S. A. *75*, 5631–5635.

Kawai, S., and Amano, A. (2012). BRCA1 regulates microRNA biogenesis via the DROSHA microprocessor complex. J. Cell Biol. *197*, 201–208.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am. J. Hum. Genet. *76*, 8–32.

Kogi, M., Fukushige, S., Lefevre, C., Hadano, S., and Ikeda, J.E. (1997). A novel tandem repeat sequence located on human chromosome 4p: isolation and characterization. Genomics *42*, 278–283.

Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. (2002). A high-resolution recombination map of the human genome. Nat. Genet. *31*, 241–247.

De Koning, P.J.A., Kummer, J.A., de Poot, S.A.H., Quadir, R., Broekhuizen, R., McGettrick, A.F., Higgins, W.J., Devreese, B., Worrall, D.M., and Bovenschen, N. (2011). Intracellular serine protease inhibitor SERPINB4 inhibits granzyme M-induced cell death. PloS One *6*, e22645.

Korbel, J.O., Kim, P.M., Chen, X., Urban, A.E., Weissman, S., Snyder, M., and Gerstein, M.B. (2008). The current excitement about copy-number variation: how it relates to gene duplications and protein families. Curr. Opin. Struct. Biol. *18*, 366–374.

Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S.T., Schlessinger, D., Sutherland, G.R., and Richards, R.I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science *252*, 1711–1714.

Kruglyak, L., and Nickerson, D.A. (2001). Variation is the spice of life. Nat. Genet. *27*, 234–236.

Krum, S.A., Miranda, G.A., Lin, C., and Lane, T.F. (2003). BRCA1 associates with processive RNA polymerase II. J. Biol. Chem. *278*, 52012–52020.

Ku, C.S., Loy, E.Y., Salim, A., Pawitan, Y., and Chia, K.S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. J. Hum. Genet. *55*, 403–415.

Kuhlmann, J.D., Baraniskin, A., Hahn, S.A., Mosel, F., Bredemeier, M., Wimberger, P., Kimmig, R., and Kasimir-Bauer, S. (2014). Circulating U2 small nuclear RNA fragments as a novel diagnostic tool for patients with epithelial ovarian cancer. Clin. Chem. *60*, 206–213.

Kuhn, L., Schramm, D.B., Donninger, S., Meddows-Taylor, S., Coovadia, A.H., Sherman, G.G., Gray, G.E., and Tiemessen, C.T. (2007). African infants' CCL3 gene copies influence perinatal HIV transmission in the absence of maternal nevirapine. AIDS Lond. Engl. *21*, 1753–1761.

Kuiper, R.P., Ligtenberg, M.J.L., Hoogerbrugge, N., and Geurts van Kessel, A. (2010). Germline copy number variation and cancer risk. Curr. Opin. Genet. Dev. *20*, 282–289.

Kulski, J.K., Shiina, T., Anzai, T., Kohara, S., and Inoko, H. (2002). Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. Immunol. Rev. *190*, 95–122.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature *413*, 519–523.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Ledet, E.M., Hu, X., Sartor, O., Rayford, W., Li, M., and Mandal, D. (2013). Characterization of germline copy number variation in high-risk African American families with prostate cancer. The Prostate *73*, 614–623.

Lee, J.A., and Lupski, J.R. (2006). Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. Neuron *52*, 103–121.

Lee, A.J., Cunningham, A.P., Kuchenbaecker, K.B., Mavaddat, N., Easton, D.F., Antoniou, A.C., Consortium of Investigators of Modifiers of BRCA1/2, and Breast Cancer Association Consortium (2014). BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. Br. J. Cancer *110*, 535–545.

Lee, C., Wevrick, R., Fisher, R.B., Ferguson-Smith, M.A., and Lin, C.C. (1997). Human centromeric DNAs. Hum. Genet. *100*, 291–304.

Lee, J.A., Carvalho, C.M.B., and Lupski, J.R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell *131*, 1235–1247.

Lemmers, R.J.L.F., Van Overveld, P.G.M., Sandkuijl, L.A., Vrieling, H., Padberg, G.W., Frants, R.R., and van der Maarel, S.M. (2004). Mechanism and timing of mitotic rearrangements in the subtelomeric D4Z4 repeat involved in facioscapulohumeral muscular dystrophy. Am. J. Hum. Genet. *75*, 44–53.

Lemmers, R.J.L.F., van der Vliet, P.J., Klooster, R., Sacconi, S., Camaño, P., Dauwerse, J.G., Snider, L., Straasheijm, K.R., van Ommen, G.J., Padberg, G.W., et al. (2010). A unifying genetic model for facioscapulohumeral muscular dystrophy. Science *329*, 1650–1653.

Leroy, E., Boyer, R., Auburger, G., Leube, B., Ulm, G., Mezey, E., Harta, G., Brownstein, M.J., Jonnalagada, S., Chernova, T., et al. (1998). The ubiquitin pathway in Parkinson's disease. Nature *395*, 451–452.

Lettre, G., Jackson, A.U., Gieger, C., Schumacher, F.R., Berndt, S.I., Sanna, S., Eyheramendy, S., Voight, B.F., Butler, J.L., Guiducci, C., et al. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. Nat. Genet. *40*, 584–591.

Levinson, G., and Gutman, G.A. (1987). High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. Nucleic Acids Res. *15*, 5323–5338.

Levinson, D.F., Duan, J., Oh, S., Wang, K., Sanders, A.R., Shi, J., Zhang, N., Mowry, B.J., Olincy, A., Amin, F., et al. (2011). Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. Am. J. Psychiatry *168*, 302–316.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.

Li, J., Jiang, T., Mao, J.-H., Balmain, A., Peterson, L., Harris, C., Rao, P.H., Havlak, P., Gibbs, R., and Cai, W.-W. (2004). Genomic segmental polymorphisms in inbred mouse strains. Nat. Genet. *36*, 952–954.

Li, Y.P., Tomanin, R., Smiley, J.R., and Bacchetti, S. (1993). Generation of a new adenovirus type 12-inducible fragile site by insertion of an artificial U2 locus in the human genome. Mol. Cell. Biol. *13*, 6064–6070.

Li, Z., Yu, A., and Weiner, A.M. (1998a). Adenovirus type 12-induced fragility of the human RNU2 locus requires p53 function. J. Virol. *72*, 4183–4191.

Li, Z., Bailey, A.D., Buchowski, J., and Weiner, A.M. (1998b). A tandem array of minimal U1 small nuclear RNA genes is sufficient to generate a new adenovirus type 12-inducible chromosome fragile site. J. Virol. *72*, 4205–4211.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S.L., and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. Nat. Genet. *25*, 239–240.

Liao, D. (1999). Concerted evolution: molecular mechanism and biological implications. Am. J. Hum. Genet. *64*, 24–30.

Liao, D., Pavelitz, T., Kidd, J.R., Kidd, K.K., and Weiner, A.M. (1997). Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. EMBO J. *16*, 588–598.

Liao, D., Yu, A., and Weiner, A.M. (1999). Coexpression of the adenovirus 12 E1B 55 kDa oncoprotein and cellular tumor suppressor p53 is sufficient to induce metaphase fragility of the human RNU2 locus. Virology *254*, 11–23.

Lieber, M.R., Ma, Y., Pannicke, U., and Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. Nat. Rev. Mol. Cell Biol. *4*, 712–720.

Lindgren, V., Ares, M., Bernstein, L.B., Weiner, A.M., and Francke, U. (1984). Mapping of human small nuclear RNA genes by in situ hybridization. 101S.

Lindgren, V., Ares, M., Jr, Weiner, A.M., and Francke, U. (1985a). Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. Nature *314*, 115–116.

Lindgren, V., Bernstein, L.B., Weiner, A.M., and Francke, U. (1985b). Human U1 small nuclear RNA pseudogenes do not map to the site of the U1 genes in 1p36 but are clustered in 1q12-q22. Mol. Cell. Biol. *5*, 2172–2180.

Liu, X., and Barker, D.F. (1999). Evidence for effective suppression of recombination in the chromosome 17q21 segment spanning RNU2-BRCA1. Am. J. Hum. Genet. *64*, 1427–1439.

Liu, C.Y., Flesken-Nikitin, A., Li, S., Zeng, Y., and Lee, W.H. (1996). Inactivation of the mouse Brca1 gene leads to failure in the morphogenesis of the egg cylinder in early postimplantation development. Genes Dev. *10*, 1835–1843.

Locke, D.P., Archidiacono, N., Misceo, D., Cardone, M.F., Deschamps, S., Roe, B., Rocchi, M., and Eichler, E.E. (2003). Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster. Genome Biol. *4*, R50.

Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M., et al. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am. J. Hum. Genet. *79*, 275–290.

Löhr, H.F., Gerken, G., Michel, G., Braun, H.B., and Meyer zum Büschenfelde, K.H. (1994). In vitro secretion of anti-GOR protein and anti-hepatitis C virus antibodies in patients with chronic hepatitis C. Gastroenterology *107*, 1443–1448.

López-Flores, I., and Garrido-Ramos, M.A. (2012). The repetitive DNA content of eukaryotic

genomes. Genome Dyn. *7*, 1–28.

Lu, Q., Wallrath, L.L., Granok, H., and Elgin, S.C. (1993). (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. Mol. Cell. Biol. *13*, 2802–2814.

Lund, E., and Dahlberg, J.E. (1984). True genes for human U1 small nuclear RNA. Copy number, polymorphism, and methylation. J. Biol. Chem. *259*, 2013–2021.

Lunt, P.W. (1998). 44th ENMC International Workshop: Facioscapulohumeral Muscular Dystrophy: Molecular Studies 19-21 July 1996, Naarden, The Netherlands. Neuromuscul. Disord. NMD *8*, 126–130.

Lunt, P.W., and Harper, P.S. (1991). Genetic counselling in facioscapulohumeral muscular dystrophy. J. Med. Genet. *28*, 655–664.

Lunt, P.W., Jardine, P.E., Koch, M., Maynard, J., Osborn, M., Williams, M., Harper, P.S., and Upadhyaya, M. (1995). Phenotypic-genotypic correlation will assist genetic counseling in 4q35-facioscapulohumeral muscular dystrophy. Muscle Nerve. Suppl. S103–109.

Luo, J., Margolis, K.L., Wactawski-Wende, J., Horn, K., Messina, C., Stefanick, M.L., Tindle, H.A., Tong, E., and Rohan, T.E. (2011). Association of active and passive smoking with risk of breast cancer among postmenopausal women: a prospective cohort study. BMJ *342*, d1016.

Lupski, J.R. (2007). Genomic rearrangements and sporadic disease. Nat. Genet. *39*, S43–47.

Lupski, J.R., and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. *1*, e49.

Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A., et al. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell *66*, 219–232.

Lyle, R., Wright, T.J., Clark, L.N., and Hewitt, J.E. (1995). The FSHD-associated repeat, D4Z4, is a member of a dispersed family of homeobox-containing repeats, subsets of which are clustered on the short arms of the acrocentric chromosomes. Genomics *28*, 389–397.

Van der Maarel, S.M., and Frants, R.R. (2005). The D4Z4 repeat-mediated pathogenesis of facioscapulohumeral muscular dystrophy. Am. J. Hum. Genet. *76*, 375–386.

MacArthur, H.L., Agarwal, M.L., and Bacchetti, S. (1997). Induction of fragility at the human RNU2 locus by cytosine arabinoside is dependent upon a transcriptionally competent U2 small nuclear RNA gene and the expression of p53. Somat. Cell Mol. Genet. *23*, 379–389.

MacLachlan, T.K., Takimoto, R., and El-Deiry, W.S. (2002). BRCA1 directs a selective p53-dependent transcriptional response towards growth arrest and DNA repair targets. Mol. Cell. Biol. *22*, 4280–4292.

Magdinier, F., Dalla Venezia, N., Lenoir, G.M., Frappart, L., and Dante, R. (1999). BRCA1 expression during prenatal development of the human mammary gland. Oncogene *18*, 4039–4043.

Magdinier, F., Billard, L.M., Wittmann, G., Frappart, L., Benchaïb, M., Lenoir, G.M., Guérin, J.F., and Dante, R. (2000). Regional methylation of the 5′ end CpG island of BRCA1 is associated with

reduced gene expression in human somatic cells. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *14*, 1585–1594.

Maher, B. (2008). Personal genomes: The case of the missing heritability. Nature *456*, 18–21.

Majumdar, A., and Patel, D.J. (2002). Identifying hydrogen bond alignments in multistranded DNA architectures by NMR. Acc. Chem. Res. *35*, 1–11.

Malkin, D., Li, F.P., Strong, L.C., Fraumeni, J.F., Jr, Nelson, C.E., Kim, D.H., Kassel, J., Gryka, M.A., Bischoff, F.Z., and Tainsky, M.A. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science *250*, 1233–1238.

Mallery, D.L., Vandenberg, C.J., and Hiom, K. (2002). Activation of the E3 ligase function of the BRCA1/BARD1 complex by polyubiquitin chains. EMBO J. *21*, 6755–6762.

Mandel, J.L. (1997). [Genetic diseases and unstable expansions of trinucleotide repeats]. Rev. Prat. *47*, 155–161.

Mangin, M., Ares, M., Jr, and Weiner, A.M. (1985). U1 small nuclear RNA genes are subject to dosage compensation in mouse cells. Science *229*, 272–275.

Manke, I.A., Lowery, D.M., Nguyen, A., and Yaffe, M.B. (2003). BRCT repeats as phosphopeptide-binding modules involved in protein targeting. Science *302*, 636–639.

Marshall, C.R., and Scherer, S.W. (2012). Detection and characterization of copy number variation in autism spectrum disorder. Methods Mol. Biol. Clifton NJ *838*, 115–135.

Matera, A.G., Weiner, A.M., and Schmid, C.W. (1990). Structure and evolution of the U2 small nuclear RNA multigene family in primates: gene amplification under natural selection? Mol. Cell. Biol. *10*, 5876–5882.

Mattaj, I.W. (1986). Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding. Cell *46*, 905–911.

Mattaj, I.W., and De Robertis, E.M. (1985). Nuclear segregation of U2 snRNA requires binding of specific snRNP proteins. Cell *40*, 111–118.

Mattaj, I.W., and Zeller, R. (1983). Xenopus laevis U2 snRNA genes: tandemly repeated transcription units sharing 5′ and 3′ flanking homology with other RNA polymerase II transcribed genes. EMBO J. *2*, 1883–1891.

Maxam, A.M., and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. *65*, 499–560.

Mazières, J., Catherinne, C., Delfour, O., Gouin, S., Rouquette, I., Delisle, M.-B., Prévot, G., Escamilla, R., Didier, A., Persing, D.H., et al. (2013). Alternative processing of the U2 small nuclear RNA produces a 19-22nt fragment with relevance for the detection of non-small cell lung cancer in human serum. PloS One *8*, e60134.

Mazoyer, S. (2005). Genomic rearrangements in the BRCA1 and BRCA2 genes. Hum. Mutat. *25*, 415–422.

McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., et al. (2006). Common deletion polymorphisms in the human genome.

Nat. Genet. *38*, 86–92.

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I.W., Maller, J.B., Kirby, A., et al. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. *40*, 1166–1174.

McCormack, V.A., and dos Santos Silva, I. (2006). Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol. *15*, 1159–1169.

McLaughlin, C.R., and Chadwick, B.P. (2011). Characterization of DXZ4 conservation in primates implies important functional roles for CTCF binding, array expression and tandem repeat organization on the X chromosome. Genome Biol. *12*, R37.

McNaught, K.S., Olanow, C.W., Halliwell, B., Isacson, O., and Jenner, P. (2001). Failure of the ubiquitin-proteasome system in Parkinson's disease. Nat. Rev. Neurosci. *2*, 589–594.

Meza, J.E., Brzovic, P.S., King, M.C., and Klevit, R.E. (1999). Mapping the functional domains of BRCA1. Interaction of the ring finger domains of BRCA1 and BARD1. J. Biol. Chem. *274*, 5659–5665.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. *45*, 353–361, 361e1–2.

Michel, G., Ritter, A., Gerken, G., Meyer zum Büschenfelde, K.H., Decker, R., and Manns, M.P. (1992). Anti-GOR and hepatitis C virus in autoimmune liver diseases. Lancet *339*, 267–269.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., and Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66–71.

Miller, D.T., Adam, M.P., Aradhya, S., Biesecker, L.G., Brothman, A.R., Carter, N.P., Church, D.M., Crolla, J.A., Eichler, E.E., Epstein, C.J., et al. (2010). Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am. J. Hum. Genet. *86*, 749–764.

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S., and Devine, S.E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res. *16*, 1182–1190.

Montgomery, S.B., Goode, D.L., Kvikstad, E., Albers, C.A., Zhang, Z.D., Mu, X.J., Ananda, G., Howie, B., Karczewski, K.J., Smith, K.S., et al. (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res. *23*, 749–761.

Mullaney, J.M., Mills, R.E., Pittard, W.S., and Devine, S.E. (2010). Small insertions and deletions (INDELs) in human genomes. Hum. Mol. Genet. *19*, R131–136.

Mullikin, J.C., Hunt, S.E., Cole, C.G., Mortimore, B.J., Rice, C.M., Burton, J., Matthews, L.H., Pavitt, R., Plumb, R.W., Sims, S.K., et al. (2000). An SNP map of human chromosome 22. Nature *407*, 516–520.

Narayanan, S. (1991). Applications of restriction fragment length polymorphism. Ann. Clin. Lab.

Sci. *21*, 291–296.

Näslund, K., Saetre, P., von Salomé, J., Bergström, T.F., Jareborg, N., and Jazin, E. (2005). Genome-wide prediction of human VNTRs. Genomics *85*, 24–35.

Neuhausen, S.L., Swensen, J., Miki, Y., Liu, Q., Tavtigian, S., Shattuck-Eidens, D., Kamb, A., Hobbs, M.R., Gingrich, J., and Shizuya, H. (1994). A P1-based physical map of the region from D17S776 to D17S78 containing the breast cancer susceptibility gene BRCA1. Hum. Mol. Genet. *3*, 1919–1926.

Newman, B., Austin, M.A., Lee, M., and King, M.C. (1988). Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. Proc. Natl. Acad. Sci. U. S. A. *85*, 3044–3048.

Nguyen, D.-Q., Webber, C., and Ponting, C.P. (2006). Bias of selection on human copy-number variants. PLoS Genet. *2*, e20.

Nohga, K., Reddy, R., and Busch, H. (1981). Comparison of RNase T1 fingerprints of U1, U2, and U3 small nuclear RNA's of HeLa cells, human normal fibroblasts, and Novikoff hepatoma cells. Cancer Res. *41*, 2215–2220.

Nojima, H., and Kornberg, R.D. (1983). Genes and pseudogenes for mouse U1 and U2 small nuclear RNAs. J. Biol. Chem. *258*, 8151–8155.

O'Reilly, D., Kuznetsova, O.V., Laitem, C., Zaborowska, J., Dienstbier, M., and Murphy, S. (2014). Human snRNA genes use polyadenylation factors to promote efficient transcription termination. Nucleic Acids Res. *42*, 264–275.

Okada, T., Gondo, Y., Goto, J., Kanazawa, I., Hadano, S., and Ikeda, J.-E. (2002). Unstable transmission of the RS447 human megasatellite tandem repetitive sequence that contains the USP17 deubiquitinating enzyme gene. Hum. Genet. *110*, 302–313.

Ouchi, T., Monteiro, A.N., August, A., Aaronson, S.A., and Hanafusa, H. (1998). BRCA1 regulates p53-dependent gene expression. Proc. Natl. Acad. Sci. U. S. A. *95*, 2302–2306.

Van Overveld, P.G.M., Lemmers, R.J.F.L., Sandkuijl, L.A., Enthoven, L., Winokur, S.T., Bakels, F., Padberg, G.W., van Ommen, G.-J.B., Frants, R.R., and van der Maarel, S.M. (2003). Hypomethylation of D4Z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. Nat. Genet. *35*, 315–317.

Padberg, G.W., Lunt, P.W., Koch, M., and Fardeau, M. (1991). Diagnostic criteria for facioscapulohumeral muscular dystrophy. Neuromuscul. Disord. NMD *1*, 231–234.

Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J., Rafiq, M.A., Conrad, D.F., Park, H., Hurles, M.E., Lee, C., Venter, J.C., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. Genome Biol. *11*, R52.

Pao, G.M., Janknecht, R., Ruffner, H., Hunter, T., and Verma, I.M. (2000). CBP/p300 interact with and function as transcriptional coactivators of BRCA1. Proc. Natl. Acad. Sci. U. S. A. *97*, 1020–1025.

Pardue, M.L., Lowenhaupt, K., Rich, A., and Nordheim, A. (1987). (dC-dA)n.(dG-dT)n sequences have evolutionarily conserved chromosomal locations in Drosophila with implications for roles in

chromosome structure and function. EMBO J. *6*, 1781–1789.

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. Nucleic Acids Res. *37*, 289–297.

Pavelitz, T., Rusché, L., Matera, A.G., Scharf, J.M., and Weiner, A.M. (1995). Concerted evolution of the tandem array encoding primate U2 snRNA occurs in situ, without changing the cytological context of the RNU2 locus. EMBO J. *14*, 169–177.

Pavelitz, T., Liao, D., and Weiner, A.M. (1999). Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. EMBO J. *18*, 3783–3792.

Pavelitz, T., Bailey, A.D., Elco, C.P., and Weiner, A.M. (2008). Human U2 snRNA genes exhibit a persistently open transcriptional state and promoter disassembly at metaphase. Mol. Cell. Biol. *28*, 3573–3588.

Perrin-Vidoz, L., Sinilnikova, O.M., Stoppa-Lyonnet, D., Lenoir, G.M., and Mazoyer, S. (2002). The nonsense-mediated mRNA decay pathway triggers degradation of most BRCA1 mRNAs bearing premature termination codons. Hum. Mol. Genet. *11*, 2805–2814.

Perry, G.H., Tchinda, J., McGrath, S.D., Zhang, J., Picker, S.R., Cáceres, A.M., Iafrate, A.J., Tyler-Smith, C., Scherer, S.W., Eichler, E.E., et al. (2006). Hotspots for copy number variation in chimpanzees and humans. Proc. Natl. Acad. Sci. U. S. A. *103*, 8006–8011.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.

Perry, G.H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C.W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N.A., et al. (2008a). The fine-scale and complex architecture of human copy-number variation. Am. J. Hum. Genet. *82*, 685–695.

Perry, G.H., Yang, F., Marques-Bonet, T., Murphy, C., Fitzgerald, T., Lee, A.S., Hyland, C., Stone, A.C., Hurles, M.E., Tyler-Smith, C., et al. (2008b). Copy number variation and evolution in humans and chimpanzees. Genome Res. *18*, 1698–1710.

Pharoah, P.D.P., Antoniou, A., Bobrow, M., Zimmern, R.L., Easton, D.F., and Ponder, B.A.J. (2002). Polygenic susceptibility to breast cancer and implications for prevention. Nat. Genet. *31*, 33–36.

Phillippy, A.M., Schatz, M.C., and Pop, M. (2008). Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. *9*, R55.

Puget, N., Gad, S., Perrin-Vidoz, L., Sinilnikova, O.M., Stoppa-Lyonnet, D., Lenoir, G.M., and Mazoyer, S. (2002). Distinct BRCA1 rearrangements involving the BRCA1 pseudogene suggest the existence of a recombination hot spot. Am. J. Hum. Genet. *70*, 858–865.

Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., et al. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat. Genet. *39*, 165–167.

Rao, Y., Hoffmann, E., Zia, M., Bodin, L., Zeman, M., Sellers, E.M., and Tyndale, R.F. (2000).

Duplications and defects in the CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. Mol. Pharmacol. *58*, 747–755.

Ray, H., Moreau, K., Dizin, E., Callebaut, I., and Venezia, N.D. (2006). ACCA phosphopeptide recognition by the BRCT repeats of BRCA1. J. Mol. Biol. *359*, 973–982.

Reddy, P.S., and Housman, D.E. (1997). The complex pathology of trinucleotide repeats. Curr. Opin. Cell Biol. *9*, 364–372.

Reddy, R., Henning, D., Epstein, P., and Busch, H. (1981). Primary and secondary structure of U2 snRNA. Nucleic Acids Res. *9*, 5645–5658.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

Reeves, G.K., Pirie, K., Beral, V., Green, J., Spencer, E., Bull, D., and Million Women Study Collaboration (2007). Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. BMJ *335*, 1134.

Ribieras, S., Magdinier, F., Leclerc, D., Lenoir, G., Frappart, L., and Dante, R. (1997). Abundance of BRCA1 transcripts in human cancer and lymphoblastoid cell lines carrying BRCA1 germ-line alterations. Int. J. Cancer J. Int. Cancer *73*, 715–718.

Richards, R.I., and Sutherland, G.R. (1992). Dynamic mutations: a new class of mutations causing human disease. Cell *70*, 709–712.

Richards, R.I., Holman, K., Yu, S., and Sutherland, G.R. (1993). Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. Hum. Mol. Genet. *2*, 1429–1435.

Riley, B., Williamson, M., Collier, D., Wilkie, H., and Makoff, A. (2002). A 3-Mb map of a large Segmental duplication overlapping the alpha7-nicotinic acetylcholine receptor gene (CHRNA7) at human 15q13-q14. Genomics *79*, 197–209.

Rincon, J.C., Engler, S.K., Hargrove, B.W., and Kunkel, G.R. (1998). Molecular cloning of a cDNA encoding human SPH-binding factor, a conserved protein that binds to the enhancer-like region of the U6 small nuclear RNA gene promoter. Nucleic Acids Res. *26*, 4846–4852.

Robitaille, Y., Lopes-Cendes, I., Becher, M., Rouleau, G., and Clark, A.W. (1997). The neuropathology of CAG repeat diseases: review and update of genetic and molecular features. Brain Pathol. Zurich Switz. *7*, 901–926.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928–933.

Sadowski, C.L., Henry, R.W., Lobo, S.M., and Hernandez, N. (1993). Targeting TBP to a non-TATA box cis-regulatory element: a TBP-containing complex activates transcription from snRNA promoters through the PSE. Genes Dev. *7*, 1535–1548.

Saitoh, Y., Miyamoto, N., Okada, T., Gondo, Y., Showguchi-Miyata, J., Hadano, S., and Ikeda,

J.E. (2000). The RS447 human megasatellite tandem repetitive sequence encodes a novel deubiquitinating enzyme with a functional promoter. Genomics *67*, 291–300.

Samonte, R.V., Conte, R.A., and Verma, R.S. (1999). Localization of human midisatellite and macrosatellite DNA sequences on chromosomes 1 and X in the great apes. J. Hum. Genet. *44*, 57–59.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U. S. A. *74*, 5463–5467.

Schaap, M., Lemmers, R.J.L.F., Maassen, R., van der Vliet, P.J., Hoogerheide, L.F., van Dijk, H.K., Baştürk, N., de Knijff, P., and van der Maarel, S.M. (2013). Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. BMC Genomics *14*, 143.

Schlegel, B.P., Green, V.J., Ladias, J.A., and Parvin, J.D. (2000). BRCA1 interaction with RNA polymerase II reveals a role for hRPB2 and hRPB10alpha in activated transcription. Proc. Natl. Acad. Sci. U. S. A. *97*, 3148–3153.

Schluth-Bolard, C., Delobel, B., Sanlaville, D., Boute, O., Cuisset, J.-M., Sukno, S., Labalme, A., Duban-Bedu, B., Plessis, G., Jaillard, S., et al. (2009). Cryptic genomic imbalances in de novo and inherited apparently balanced chromosomal rearrangements: array CGH study of 47 unrelated cases. Eur. J. Med. Genet. *52*, 291–296.

Schluth-Bolard, C., Labalme, A., Cordier, M.-P., Till, M., Nadeau, G., Tevissen, H., Lesca, G., Boutry-Kryza, N., Rossignol, S., Rocas, D., et al. (2013). Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. J. Med. Genet. *50*, 144–150.

Schramayr, S., Caporossi, D., Mak, I., Jelinek, T., and Bacchetti, S. (1990). Chromosomal damage induced by human adenovirus type 12 requires expression of the E1B 55-kilodalton viral protein. J. Virol. *64*, 2090–2095.

Scully, R., Anderson, S.F., Chao, D.M., Wei, W., Ye, L., Young, R.A., Livingston, D.M., and Parvin, J.D. (1997). BRCA1 is a component of the RNA polymerase II holoenzyme. Proc. Natl. Acad. Sci. U. S. A. *94*, 5605–5610.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., et al. (2004). Large-scale copy number polymorphism in the human genome. Science *305*, 525–528.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449.

Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. *77*, 78–88.

She, X., Cheng, Z., Zöllner, S., Church, D.M., and Eichler, E.E. (2008). Mouse segmental

duplication and copy number variation. Nat. Genet. *40*, 909–914.

Shlien, A., Tabori, U., Marshall, C.R., Pienkowska, M., Feuk, L., Novokmet, A., Nanda, S., Druker, H., Scherer, S.W., and Malkin, D. (2008). Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. Proc. Natl. Acad. Sci. U. S. A. *105*, 11264–11269.

Silver, H. (1989). Paternity testing. Crit. Rev. Clin. Lab. Sci. *27*, 391–408.

Sinsheimer, R.L. (1989). The Santa Cruz Workshop--May 1985. Genomics *5*, 954–956.

Skuzeski, J.M., and Jendrisak, J.J. (1985). A family of wheat embryo U2 snRNAs. Plant Mol. Biol. *4*, 181–193.

Smith, H.O., and Wilcox, K.W. (1970). A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J. Mol. Biol. *51*, 379–391.

Smith, T.M., Lee, M.K., Szabo, C.I., Jerome, N., McEuen, M., Taylor, M., Hood, L., and King, M.C. (1996). Complete genomic sequence and analysis of 117 kb of human DNA containing the gene BRCA1. Genome Res. *6*, 1029–1049.

Snijders, A.M., Nowak, N.J., Huey, B., Fridlyand, J., Law, S., Conroy, J., Tokuyasu, T., Demir, K., Chiu, R., Mao, J.-H., et al. (2005). Mapping segmental and sequence variations among laboratory mice using BAC array CGH. Genome Res. *15*, 302–311.

Sørensen, P.D., Lomholt, B., Frederiksen, S., and Tommerup, N. (1991). Fine mapping of human 5S rRNA genes to chromosome 1q42.11----q42.13. Cytogenet. Cell Genet. *57*, 26–29.

Southard, A.E., Edelmann, L.J., and Gelb, B.D. (2012). Role of copy number variants in structural birth defects. Pediatrics *129*, 755–763.

Srivastava, S., Zou, Z.Q., Pirollo, K., Blattner, W., and Chang, E.H. (1990). Germ-line transmission of a mutated p53 gene in a cancer-prone family with Li-Fraumeni syndrome. Nature *348*, 747–749.

Stankiewicz, P., and Lupski, J.R. (2002). Molecular-evolutionary mechanisms for genomic disorders. Curr. Opin. Genet. Dev. *12*, 312–319.

Stein, L.D. (2004). Human genome: end of the beginning. Nature *431*, 915–916.

Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavaré, S., et al. (2005). Genome-wide associations of gene expression variation in humans. PLoS Genet. *1*, e78.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., and Inouye, M. (1966). Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. Cold Spring Harb. Symp. Quant. Biol. *31*, 77–84.

Szabo, C.I., Wagner, L.A., Francisco, L.V., Roach, J.C., Argonza, R., King, M.C., and Ostrander, E.A. (1996). Human, canine and murine BRCA1 genes: sequence comparison among species. Hum. Mol. Genet. *5*, 1289–1298.

Tammi, M.T., Arner, E., and Andersson, B. (2003). TRAP: Tandem Repeat Assembly Program produces improved shotgun assemblies of repetitive sequences. Comput. Methods Programs Biomed. *70*, 47–59.

Tautz, D. (1993). Notes on the definition and nomenclature of tandemly repetitive DNA sequences. EXS *67*, 21–28.

Tautz, D., and Schlötterer (1994). Simple sequences. Curr. Opin. Genet. Dev. *4*, 832–837.

Tavtigian, S.V., Simard, J., Rommens, J., Couch, F., Shattuck-Eidens, D., Neuhausen, S., Merajver, S., Thorlacius, S., Offit, K., Stoppa-Lyonnet, D., et al. (1996). The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. Nat. Genet. *12*, 333–337.

Tawil, R., Forrester, J., Griggs, R.C., Mendell, J., Kissel, J., McDermott, M., King, W., Weiffenbach, B., and Figlewicz, D. (1996). Evidence for anticipation and association of deletion size with severity in facioscapulohumeral muscular dystrophy. The FSH-DY Group. Ann. Neurol. *39*, 744–748.

Tawil, R., Figlewicz, D.A., Griggs, R.C., and Weiffenbach, B. (1998). Facioscapulohumeral dystrophy: a distinct regional myopathy with a novel molecular pathogenesis. FSH Consortium. Ann. Neurol. *43*, 279–282.

Taylor, M.S., Ponting, C.P., and Copley, R.R. (2004). Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. Genome Res. *14*, 555–566.

Tessereau, C., Buisson, M., Monnet, N., Imbert, M., Barjhoux, L., Schluth-Bolard, C., Sanlaville, D., Conseiller, E., Ceppi, M., Sinilnikova, O.M., et al. (2013). Direct visualization of the highly polymorphic RNU2 locus in proximity to the BRCA1 gene. PloS One *8*, e76054.

Thakur, S., and Croce, C.M. (1999). Positive regulation of the BRCA1 promoter. J. Biol. Chem. *274*, 8837–8843.

The BRCA1 Exon 13 Duplication Screening Group (2000). The exon 13 duplication in the BRCA1 gene is a founder mutation present in geographically diverse populations. Am. J. Hum. Genet. *67*, 207–212.

Tjio, J.H., and Nichols, W.W. (1985). History and present status of human chromosome studies. Vitro Cell. Dev. Biol. J. Tissue Cult. Assoc. *21*, 305–313.

Travis, L.B., Hill, D., Dores, G.M., Gospodarowicz, M., van Leeuwen, F.E., Holowaty, E., Glimelius, B., Andersson, M., Pukkala, E., Lynch, C.F., et al. (2005). Cumulative absolute breast cancer risk for young women treated for Hodgkin lymphoma. J. Natl. Cancer Inst. *97*, 1428–1437.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. *13*, 36–46.

Tremblay, D.C., Alexander, G., Jr, Moseley, S., and Chadwick, B.P. (2010). Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. BMC Genomics *11*, 632.

Tremblay, D.C., Moseley, S., and Chadwick, B.P. (2011). Variation in array size, monomer composition and expression of the macrosatellite DXZ4. PloS One *6*, e18969.

Tucker, B.A., Scheetz, T.E., Mullins, R.F., DeLuca, A.P., Hoffmann, J.M., Johnston, R.M., Jacobson, S.G., Sheffield, V.C., and Stone, E.M. (2011). Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of

retinitis pigmentosa. Proc. Natl. Acad. Sci. U. S. A. *108*, E569–576.

Turner, D.J., Miretti, M., Rajan, D., Fiegler, H., Carter, N.P., Blayney, M.L., Beck, S., and Hurles, M.E. (2008). Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. Nat. Genet. *40*, 90–95.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D., et al. (2005). Fine-scale structural variation of the human genome. Nat. Genet. *37*, 727–732.

Vandeweyer, G., Reyniers, E., Wuyts, W., Rooms, L., and Kooy, R.F. (2011). CNV-WebStore: online CNV analysis, storage and interpretation. BMC Bioinformatics *12*, 4.

Varesco, L., Viassolo, V., Viel, A., Gismondi, V., Radice, P., Montagna, M., Alducci, E., Della Puppa, L., Oliani, C., Tommasi, S., et al. (2013). Performance of BOADICEA and BRCAPRO genetic models and of empirical criteria based on cancer family history for predicting BRCA mutation carrier probabilities: a retrospective study in a sample of Italian cancer genetics clinics. Breast Edinb. Scotl. *22*, 1130–1135.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Vergnaud, G., and Denoeud, F. (2000). Minisatellites: mutability and genome architecture. Genome Res. *10*, 899–907.

Warburton, P.E., Hasson, D., Guillem, F., Lescale, C., Jin, X., and Abrusan, G. (2008). Analysis of the largest tandemly repeated DNA families in the human genome. BMC Genomics *9*, 533.

Warren, S.T., Zhang, F., Licameli, G.R., and Peters, J.F. (1987). The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. Science *237*, 420–423.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weber, J.L. (1990). Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics *7*, 524–530.

Weber, J.L., and Wong, C. (1993). Mutation of human short tandem repeats. Hum. Mol. Genet. *2*, 1123–1128.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., and Lathrop, M. (1992). A second-generation linkage map of the human genome. Nature *359*, 794–801.

Wells, R.D. (1988). Unusual DNA structures. J. Biol. Chem. *263*, 1095–1098.

Westin, G., Monstein, H.J., Zabielski, J., Philipson, L., and Pettersson, U. (1981). Human DNA sequences complementary to the small nuclear RNA U2. Nucleic Acids Res. *9*, 6323–6338.

Westin, G., Zabielski, J., Hammarström, K., Monstein, H.J., Bark, C., and Pettersson, U. (1984). Clustered genes for human U2 RNA. Proc. Natl. Acad. Sci. U. S. A. *81*, 3811–3815.

Wijmenga, C., Padberg, G.W., Moerer, P., Wiegant, J., Liem, L., Brouwer, O.F., Milner, E.C., Weber, J.L., van Ommen, G.B., and Sandkuyl, L.A. (1991). Mapping of facioscapulohumeral muscular dystrophy gene to chromosome 4q35-qter by multipoint linkage analysis and in situ hybridization.

Genomics *9*, 570–575.

Wilson, G.M., Flibotte, S., Missirlis, P.I., Marra, M.A., Jones, S., Thornton, K., Clark, A.G., and Holt, R.A. (2006). Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. Genome Res. *16*, 173–181.

Wooster, R., Neuhausen, S.L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., and Averill, D. (1994). Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science *265*, 2088–2090.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789–792.

Wyman, A.R., and White, R. (1980). A highly polymorphic locus in human DNA. Proc. Natl. Acad. Sci. U. S. A. *77*, 6754–6758.

Xu, C.F., Brown, M.A., Chambers, J.A., Griffiths, B., Nicolai, H., and Solomon, E. (1995). Distinct transcription start sites generate two forms of BRCA1 mRNA. Hum. Mol. Genet. *4*, 2259–2264.

Xu, C.F., Brown, M.A., Nicolai, H., Chambers, J.A., Griffiths, B.L., and Solomon, E. (1997). Isolation and characterisation of the NBR2 gene which lies head to head with the human BRCA1 gene. Hum. Mol. Genet. *6*, 1057–1062.

Yang, Q., Rasmussen, S.A., and Friedman, J.M. (2002). Mortality associated with Down's syndrome in the USA from 1983 to 1997: a population-based study. Lancet *359*, 1019–1025.

Yang, Y., Chung, E.K., Wu, Y.L., Savelli, S.L., Nagaraja, H.N., Zhou, B., Hebert, M., Jones, K.N., Shu, Y., Kitzmiller, K., et al. (2007). Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. Am. J. Hum. Genet. *80*, 1037–1054.

Yasmeen, A., Liu, W., Dekhil, H., Kassab, A., Aloyz, R., Foulkes, W.D., and Al Moustafa, A.-E. (2008). BRCA1 mutations contribute to cell motility and invasion by affecting its main regulators. Cell Cycle Georget. Tex *7*, 3781–3783.

Ye, F., and Signer, E.R. (1996). RIGS (repeat-induced gene silencing) in Arabidopsis is transcriptional and alters chromatin configuration. Proc. Natl. Acad. Sci. U. S. A. *93*, 10881–10886.

Yee, H.A., Wong, A.K., van de Sande, J.H., and Rattner, J.B. (1991). Identification of novel single-stranded d(TC)n binding proteins in several mammalian species. Nucleic Acids Res. *19*, 949–953.

Yu, A., Bailey, A.D., and Weiner, A.M. (1998). Metaphase fragility of the human RNU1 and RNU2 loci is induced by actinomycin D through a p53-dependent pathway. Hum. Mol. Genet. *7*, 609–617.

Yu, A., Fan, H.Y., Liao, D., Bailey, A.D., and Weiner, A.M. (2000). Activation of p53 or loss of the Cockayne syndrome group B repair protein causes metaphase fragility of human U1, U2, and 5S genes. Mol. Cell *5*, 801–810.

Yu, X., Chini, C.C.S., He, M., Mer, G., and Chen, J. (2003). The BRCT domain is a phospho-

protein binding domain. Science *302*, 639–642.

Yuo, C.Y., Ares, M., Jr, and Weiner, A.M. (1985). Sequences required for 3′ end formation of human U2 small nuclear RNA. Cell *42*, 193–202.

Zahnleiter, D., Uebe, S., Ekici, A.B., Hoyer, J., Wiesener, A., Wieczorek, D., Kunstmann, E., Reis, A., Doerr, H.-G., Rauch, A., et al. (2013). Rare copy number variants are a common cause of short stature. PLoS Genet. *9*, e1003365.

Zatz, M., Marie, S.K., Passos-Bueno, M.R., Vainzof, M., Campiotto, S., Cerqueira, A., Wijmenga, C., Padberg, G., and Frants, R. (1995). High proportion of new mutations and possible anticipation in Brazilian facioscapulohumeral muscular dystrophy families. Am. J. Hum. Genet. *56*, 99–105.

Zhang, H., Somasundaram, K., Peng, Y., Tian, H., Zhang, H., Bi, D., Weber, B.L., and El-Deiry, W.S. (1998). BRCA1 physically associates with p53 and stimulates its transcriptional activity. Oncogene *16*, 1713–1721.

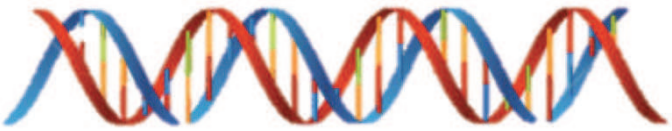
Zhao, M., and Zhao, Z. (2013). CNVannotator: a comprehensive annotation server for copy number variation in the human genome. PloS One *8*, e80170.

Zhao, M., Zhi, H., Doust, A.N., Li, W., Wang, Y., Li, H., Jia, G., Wang, Y., Zhang, N., and Diao, X. (2013). Novel genomes and genome constitutions identified by GISH and 5S rDNA and knotted1 genomic sequences in the genus Setaria. BMC Genomics *14*, 244.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S., and Wang, W. (2008). On the origin of new genes in Drosophila. Genome Res. *18*, 1446–1455.

Zody, M.C., Garber, M., Adams, D.J., Sharpe, T., Harrow, J., Lupski, J.R., Nicholson, C., Searle, S.M., Wilming, L., Young, S.K., et al. (2006). DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature *440*, 1045–1049.

Annexe



(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 December 2012 (06.12.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/164401 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/IB2012/001333

(22) International Filing Date:
1 June 2012 (01.06.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/493,010 3 June 2011 (03.06.2011) US

(71) Applicants *(for all designated States except US)*: **GENOMIC VISION** [FR/FR]; 80-84 rue des Meuniers, F-92220 Bagneux (FR). **CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE** [FR/FR]; 3, rue Michel-Ange, F-75794 Paris Cedex 16 (FR). **UNIVERSITE CLAUDE BERNARD DE LYON 1** [FR/FR]; 43 Bd du 11 Novembre 1918, F-69622 Villeurbanne Cedex (FR). **CENTRE DE LUTTE CONTRE LE CANCER LÉON BÉRARD** [FR/FR]; 28 Rue Laennec, F-69008 Lyon (FR).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **MAZOYER, Sylvie** [FR/FR]; 40 ter rue Seignemartin, F-69008 Lyon (FR). **TESSEREAU, Chloé** [FR/FR]; 72 rue Jaboulay, F-69007 Lyon (FR). **CEPPI, Maurizio** [CH/FR]; 2bis Henri Tariel, F-92130 Issy Les Moulineaux (FR). **CHEESEMAN, Kevin** [FR/FR]; 20 avenue Edmond, F-94500 Champigny Sur Marne (FR). **VANNIER, Anne** [FR/FR]; 17 rue Guy de Maupassant, F-76280 Saint-Jouin-Bruneval (FR).

(74) Agent: **GUTMANN, Ernest**; 3, rue Auber, F-75009 Paris (FR).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *with sequence listing part of description (Rule 5.2(a))*

WO 2012/164401 A1

(54) **Title**: ASSESSMENT OF CANCER RISK BASED ON *RNU2* CNV AND INTERPLAY BETWEEN *RNU2* CNV AND *BRCA1*

(57) **Abstract**: Polynucleotides useful for detecting copy number variation of *RNU2* sequences and methods of assessing risk of developing breast or ovarian cancer using molecular combing and/or detection or quantification of *BRCA1* expression.

Claims

Claim 1. An isolated or purified polynucleotide

that binds to an *RNU2* polynucleotide sequence,

that binds to *RNU2 CNV* (copy number variation), or

that binds to a sequence flanking an *RNU2* CNV; or

an isolated or purified polynucleotide that is useful

as a primer for the amplification of an *RNU2 CNV* polynucleotide sequence;

as a primer for the amplification of a sequence lying between BRCA1 and an *RNU2 CNV* sequence; or

as a primer for the amplification of a sequence flanking an *RNU2 CNV* polynucleotide sequence.

Claim 2. The isolated or purified polynucleotide of claim 1 that is selected from the group consisting of L1 (nt 20-542)(SEQ ID NO: 27), L2 (nt 731-1230)(SEQ ID NO: 28), L3 (nt 1738-2027)(SEQ ID NO: 29), L4 (nt 3048-3481)(SEQ ID NO: 30), L5 (nt 3859-5817)(SEQ ID NO: 31), R1 (nt 1-485)(SEQ ID NO: 32), R2 (nt 1288-1787)(SEQ ID NO: 33), R3 (nt 2075-4237)(SEQ ID NO: 34), R4 (nt 4641-5022)(SEQ ID NO: 35), R5 (nt 5391-5970)(SEQ ID NO: 36), R6 (nt 6702-7590)(SEQ ID NO: 37), C1 (SEQ ID NO: 60), C2 (SEQ ID NO: 61), C3 (SEQ ID NO: 62) and C4 (SEQ ID NO: 63); or

a polynucleotide that hybridizes under stringent conditions with said isolated or purified polynucleotide or its full complement;

wherein stringent conditions comprise washing in 0.1 x SSC and 0.1% SDS at a temperature of 68°C.

Claim 3. The isolated or purified polynucleotide of claim 1 that is selected from the group consisting of SEQ ID NOS: 1-25 and 26.

Claim 4. The isolated or purified polynucleotide of claim 1 that is selected from the group consisting of SEQ ID NOS: 1-25 and 26, and 44-51 and 52-59.

Claim 5. The isolated or purified polynucleotide of claim 1 that is selected from the group consisting of L1Fq (SEQ ID NO: 38), L1Rq (SEQ ID NO: 39) and Taqman L1 (SEQ ID NO: 42).

Claim 6. A kit for detecting the genetic predisposition of developing a breast or an ovarian cancer comprising:

primers for amplification of DNA corresponding to *RNU2* CNV region, probes specific for *RNU2* CNV, and/or optionally primers and/or probes specific for BRCA1 gene expression.

Claim 7. A kit according to claim 6 wherein the primers are selected from the group consisting of SEQ ID NOS: 1-25 and 26 and 52-59; or

are selected from the group consisting of L1Fq (SEQ ID NO: 38), L1Rq (SEQ ID NO: 39) and Taqman L1 (SEQ ID NO: 42) and/or the probes are selected from the group consisting of L1 (nt 20-542)(SEQ ID NO: 27), L2 (nt 731-1230)(SEQ ID NO: 28), L3 (nt 1738-2027)(SEQ ID NO: 29), L4 (nt 3048-3481)(SEQ ID NO: 30), L5 (nt 3859-5817)(SEQ ID NO: 31), R1 (nt 1-485)(SEQ ID NO: 32), R2 (nt 1288-1787)(SEQ ID NO: 33), R3 (nt 2075-4237)(SEQ ID NO: 34), R4 (nt 4641-5022)(SEQ ID NO: 35), R5 (nt 5391-5970)(SEQ ID NO: 36) R6 (nt 6702-7590)(SEQ ID NO: 37), C1 (SEQ ID NO: 60), C2 (SEQ ID NO: 61), C3 (SEQ ID NO: 62) and C4 (SEQ ID NO: 63); or a polynucleotide that hybridizes under stringent conditions with said isolated or purified polynucleotide or its full complement, wherein stringent conditions comprise washing in 0.1 x SSC and 0.1% SDS at a temperature of 68°C.

Claim 8. A method of detecting the number of copies of an *RNU2 s*equence in a sample containing an *RNU2* copy number variant (CNV) comprising:

contacting the sample with one or more probes that identify an *RNU2* CNV sequence of interest, and

determining the number of sequences based on the pattern of probe binding to the sequence of interest or on the quantity of probe bound to the sample.

Claim 9. A method according to claim 8 wherein the sample is subjected to molecular combing prior to contacting the sample with one or more probes that identify and *RNU2* CNV sequence of interest, and

determining the number of sequences based on the pattern of probe binding to the combed sequence of interest.

Claim 10. The method of claim 8 or 9, wherein determining the number of *RNU2* sequences comprises determining (a) the position of the probes, (b) the distance between probes, or (c) the size of the probes.

Claim 11. The method of any of claims 8 to 10, wherein at least one of said probes is selected from the group consisting of L1 (nt 20-542)(SEQ ID NO: 27), L2 (nt 731-1230)(SEQ ID NO: 28), L3 (nt 1738-2027)(SEQ ID NO: 29), L4 (nt 3048-3481)(SEQ ID NO: 30), L5 (nt 3859-5817)(SEQ ID NO: 31), R1 (nt 1-485)(SEQ ID NO: 32), R2 (nt 1288-1787)(SEQ ID NO: 33), R3 (nt 2075-4237)(SEQ ID NO: 34), R4 (nt 4641-5022)(SEQ ID NO: 35), R5 (nt 5391-5970)(SEQ ID NO: 36) R6 (nt 6702-7590)(SEQ ID NO: 37), C1 (SEQ ID NO: 60), C2 (SEQ ID NO: 61), C3 (SEQ ID NO: 62) and C4 (SEQ ID NO: 63); or a polynucleotide that hybridizes under stringent conditions with said isolated or purified polynucleotide or its full complement, wherein stringent conditions comprise washing in 0.1 x SSC and 0.1% SDS at a temperature of 68°C.

Claim 12. The method of any of claims 8 to 11, wherein the sample contains several DNA molecules with different numbers of copies of an *RNU2* sequence and wherein the number of copies of an *RNU2* sequence is determined independently for each DNA molecule.

Claim 13. A method of detecting the number of copies of one or several *RNU2* sequences in a sample containing an *RNU2* copy number variant (CNV) comprising:

contacting a DNA sample suspected to contain an *RNU2* CNV with primers under conditions suitable for amplification of all or part of the *RNU2* sequences;

amplifying all or part of the *RNU2* sequences;

determining the number of sequences based on the characteristic of the bound primers or of the amplified products.

Claim 14. The method of claim 13, wherein at least one of said primers are selected from the group consisting of SEQ ID NOS: 1-25 and 26 and 52-59; or

are selected from the group consisting of L1Fq (SEQ ID NO: 38), L1Rq (SEQ ID NO: 39) and Taqman L1 (SEQ ID NO: 42).

Claim 15. A method for detecting a cancer or assessing the risk of developing cancer or detecting a predisposition to cancer comprising:

determining the length or number of copies of *RNU2* sequences in sample and correlating the said length or copy number with a risk or predisposition to cancer and optionally

correlating the said length or copy number with expression of a BRCA1 gene or a gene of interest within 500 kb of said *RNU2* sequences, associated with said *RNU2* sequences on a DNA molecule and optionally

determining a risk or predisposition to cancer when the length or number of copies of said *RNU2* sequences reduces the expression of BRCA1 or a gene of interest.

Claim 16. The method of claim 15, wherein said cancer is ovarian cancer or breast cancer.

Claim 17. The method of claim 15, wherein a risk or predisposition to cancer is positively correlated with the length or number of copies of said *RNU2* sequences.

Claim 18. The method of any of claims 15 to 17 wherein the number of copies of *RNU2* sequences in sample is detected using a probe as defined in claim 10 or 11.

Claim 19. The method of claim 15, wherein expression of a BRCA1 gene is determined by detecting mRNA transcribed from said gene.

Claim 20. The method of claim 15, wherein expression of a BRCA1 gene is determined by detecting the presence of a polypeptide expressed by the BRCA1 gene.

Claim 21. The method of claim 15, wherein the presence of said polypeptide is detected by one or more antibodies that bind to a normal or to a mutated BRCA1 polypeptide.

Claim 22. A method using of molecular combing to detect the presence or absence of *RNU2* sequences or the length or number of copies of *RNU2* sequences in a DNA single or a double stranded DNA molecule possibly containing BRCA1 gene.

Claim 23. A method using molecular combing to detect the presence or absence of genetic abnormalities at an *RNU2* locus associated with BRCA1, wherein an *RNU2* abnormality is defined as a structure of *RNU2* sequences found at a higher frequency in a subject having a lower level of BRCA1 expression than the mean level of BRCA1 expression of control subjects.

Claim 24. A method using molecular combing to detect the predisposition of developing ovarian or breast cancer by identification of *BRCA1* and *RNU2* genes or the number of copies of *RNU2* sequences in a sample.

Claim 25. A method for detecting a cancer or assessing the risk of developing cancer or detecting a predisposition to cancer according to claim 15, wherein the determined length or number of copies of an *RNU2* sequence is compared either with values obtained in normal subjects and in cancer-affected subjects, or with a threshold value previously established as being a minimum value characteristic of a cancer or an increased risk of cancer, or a predisposition to cancer.

Europäisches Patentamt
European Patent Office
Office européen des brevets

European Patent Office
Postbus 5818
2280 HV RIJSWIJK
NETHERLANDS
Tel. +31 (0)70 340-2040
Fax +31 (0)70 340-3016

MAZOYER, Sylvie
40 ter rue Seignemartin
F-69008 Lyon
FRANCE

For any questions about this communication:
Tel.:+31 (0)70 340 45 00

Date
27.01.14

Reference	Application No./Patent No.
	12738602.7 - 1403
Applicant/Proprietor	
Genomic Vision, et al	

Notification of the data mentioned in Rule 19(3) EPC

In the above-identified patent application you are designated as inventor/co-inventor.
Pursuant to Rule 19(3) EPC the following data are notified herewith:

DATE OF FILING	: 01.06.12
PRIORITY	: US/03.06.11/ USP201161493010
TITLE	: ASSESSMENT OF CANCER RISK BASED ON RNU2 CNV AND INTERPLAY BETWEEN RNU2 AND BRCA1
DESIGNATED STATES	: AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

Receiving Section

Europäisches Patentamt • European Patent Office • Office européen des brevets

Le macrosatellite *RNU2* est constitué de répétitions en tandem d'une unité de 6,1 kb. Largement étudié pendant les années 1980 et 1990, il est maintenant oublié des études pan-génomiques du fait de son absence du génome de référence.

J'ai dans un premier temps finement caractérisé ce macrosatellite, en réalisant un assemblage *in silico* de la région génomique, en développant un code-barres pour la technique de peignage moléculaire et en analysant les données du projet 1000 Génomes. J'ai ainsi validé la localisation du locus *RNU2* 124 kb en amont de *BRCA1*, et affiné les données de polymorphisme en montrant que le nombre allélique de copies pouvait varier entre 5 et 82 chez 42 individus. J'ai tiré profit de sa localisation au sein d'un large bloc de déséquilibre de liaison pour définir le taux de mutation de ce macrosatellite à l'origine du nombre important d'allèles identifiables au sein de la population générale. Compte tenu de sa proximité avec *BRCA1* et de son fort taux de polymorphisme, j'ai étudié le nombre global de copies du CNV dans 2 cohortes de cas de cancer du sein et témoins associés. J'ai montré que le nombre global de copies est significativement plus élevé chez les cas que chez les témoins.

Ce travail suggère que le nombre de copies du macrosatellite *RNU2* pourrait être impliqué dans la prédisposition génétique au cancer du sein, impliquant ainsi pour la première fois un CNV dans un mécanisme d'inactivation d'un gène de prédisposition au cancer.

The *RNU2* macrosatellite : characterization, evolution and link with breast cancer genetic predisposition

The *RNU2* macrosatellite is composed by tandem repeats of a 6.1 kb-long unit. Extensively studied during the 1980's and the 1990's, this locus is now omitted from genome-wide analysis as a result of its absence from the human reference genome.

Firstly, I finely characterized this macrosatellite by performing an *in silico* assembly of this genomic region, by designing a barcode for the molecular combing technique and by analyzing the 1,000 Genomes data. I thus validated the localization of the *RNU2* locus 124 kb upstream of *BRCA1*, and refined the polymorphism data by showing that the allelic copy number ranged from 5 to 82 in 42 individuals. I took advantage of its localization in a large disequilibrium block to determine the mutation rate of this macrosatellite, responsible for the high number of alleles found in the general population. Considering its close proximity to *BRCA1* and its high level of polymorphism, I studied the global copy number in 2 cohorts of breast cancer cases and controls. I thus showed that the *RNU2* global copy number is significantly higher in breast cancer cases than in controls. My work suggests that the RNU2 macrosatellite copy number could be involved in breast cancer genetic predisposition. This would provide the first example of one inactivating mechanism of a cancer predisposing gene by a macrosatellite.

DISCIPLINE : Génétique, Biologie Moléculaire, Oncologie.

MOTS-CLES : *RNU2*, macrosatellite, CNV, *BRCA1*, Cancer du sein.

INTITULE ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Centre de Recherche en Cancérologie de Lyon, UMR Inserm 1052, CNRS 5286

Centre Léon Bérard, Bâtiment Cheney D, 28 Rue Laennec, 69373 Lyon Cedex 08