Conséquences du contexte haplotypique sur la fonctionnalité des protéines : application à la mucoviscidose

Tania Cuppens

UNIVERSITE BIOLOGIE BRETAGNE SANTE LOIRE

UBO Université de Bretagne Occidentale

THESE DE DOCTORAT DE

L'UNIVERSITE
DE BRETAGNE OCCIDENTALE
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : Génétique, Génomique, Bioinformatique

Par

Tania CUPPENS

Conséquences du contexte haplotypique sur la fonction des protéines : application à la mucoviscidose

Thèse présentée et soutenue à Brest, le 7 Mai 2019
Unité de recherche : INSERM U1078, équipe Génétique Moléculaire et Epidémiologie Génétique

Rapporteurs avant soutenance :

Pascale FANEN Professeure - Praticienne hospitalière, Université Paris-Est, U955
Pascal RIHET Professeur, Aix-Marseille Université, UMR1090

Composition du Jury :

Président :
Claude FEREC Professeur - Praticien hospitalier, Université Bretagne Occidentale, U1078

Examinateurs :
Isabelle CALLEBAUT Directrice de recherche, Sorbonne Université, UMR 7590

Pascale FANEN Professeure - Praticienne hospitalière, Université Paris-Est, U955

Jean MULLER Maître de conférences - Praticien hospitalier, Université de Strasbourg, UMR112

Pascal RIHET Professeur, Aix-Marseille Université, UMR1090

Dir. de thèse : Emmanuelle GENIN Directrice de recherche, Université Bretagne Occidentale, U1078
Co-dir. de thèse : Pascal TROUVE Ingénieur de recherche, Université Bretagne Occidentale, U1078

REMERCIEMENTS

Je tiens en premier lieu à remercier très sincèrement les professeurs Pascale Fanen et Pascal Rihet d'avoir accepté de rapporter ce travail. Je remercie également le Dr. Isabelle Callebaut, le Dr. Jean Muller et le Pr. Claude Férec d'avoir accepté de faire partie du jury de ma soutenance.

Je remercie tout particulièrement Emmanuelle Genin et Pascal Trouvé, mes directeurs de thèse, pour m'avoir permis de faire cette thèse. Je vous remercie de m'avoir accordé votre temps afin de m'accompagner, me conseiller et me soutenir tout au long de ma thèse. J'ai appris beaucoup à votre contact et j'espère que je ne vous ai pas déçus respectivement dans la partie bio et bioinfo de mes travaux de thèse.

Je tiens également à remercier Marie De Tayrac pour toute l'aide qu'elle m'a apportée en début de thèse et pour m'avoir permis de débuter ma thèse au sein de son laboratoire.

Je remercie Nathalie Benz et Elise Rouillé pour m'avoir (re)initié à la biologie moléculaire. Vous m'avez aidée à me former aux techniques de laboratoire qui m'ont permises de réaliser toute la partie analyse moléculaire de ce mémoire. Je tiens également à vous remercier pour votre soutien et pour tous les moments que l'on a passé ensemble. Nath, je te souhaite plein de courage dans ta vie professionnelle et plein de bonheur dans ta vie personnelle avec Leo l'asticot et madame Val. Elise, je te souhaite le meilleur pour la suite avec pleins de gâteaux licornes (bon courage pour la fin de thèse).

Je remercie également Julie et Agathe qui m'ont aidé à obtenir les données nécessaires à ma fin de thèse. Bonne continuation à toutes les deux.

Je remercie aussi Fabien pour ses cascades dangereuses et je remercie également tes parents de t'avoir donné la vie, ce qui nous a permis de rire tous les jours.

Je tiens à remercier toute l'équipe « Bioinfo-GenStats » pour leur soutien, leur écoute et leur conseil. Aude, Gaëlle LF, Gaëlle M, Lourdes, Ozvan et Thomas, j'ai pu compter sur vous pour me remonter le moral en toutes circonstances.

Aude, je te souhaite pleins de bonnes choses dans ta recherche et ton enseignement mais surtout au niveau personnel.

Gaëlle, j'espère et je souhaite que tout aille bien dans ta vie au niveau professionnel et au niveau personnel avec Gus et les petites terreurs.

Gaëlle, je ne peux rien te promettre sur mes prononciations sudistes, mais on pourra toujours se faire un poulèèè à l'occasion.

Lourdes, ton accent espagnol et ta joie de vivre sont un vrai petit soleil surtout lorsque que le temps à Brest se fait pluvieux (ce qui n'arrive presque jamais). Pleins de bonnes choses avec tes monsters.

Ozvan, je te souhaite plein de bonheur pour tes trois années de thèse et pour après bien évidemment. Merci d'avoir été là pendant les moments difficiles surtout en cette fin de thèse.

Thomas, merci pour ta folie au quotidien qui ont égayé le bureau. Plus sérieusement, tu m'a surtout été d'une grande aide de par tes connaissances en informatique et bioinformatique.

Pour finir, je tiens à remercier toute l'unité « Génétique, génomique fonctionnelle et biotechnologies » pour m'avoir accueillie et intégrée dans l'équipe.

Je souhaite remercier infiniment l'ensemble de ma famille. Mes parents, vous m'avez aidé, conseillé et soutenu dans tout ce que j'ai entrepris. Mon frère, même à l'autre bout du monde, tu as toujours été un soutien cher à mes yeux. Chou, tu m'as supporté pendant toutes mes années d'études surtout pendant mes moments de doute. Et pour finir, Bouchon et Babychat mes petits bonheurs du quotidien ♡.

TABLE DES MATIÈRES

LISTES DES ABRÉVIATIONS

508del p.Phe508del.

ABC ATP Binding Cassette.

ADN Acide Désoxyribonucleique.

ADNc Acide Désoxyribonucléique complémentaire.

AMPc Adénosine MonoPhosphate cyclique.

ARN Acide Ribonucléique.

ARNm Acide Ribonucléique messager.

ATP Adénosine Triphosphate.

BAM Binary Alignment/Map format.

CAUV Congenital Absence of the Uterus and Vagina.

CBAVD Congenital Bilateral Absence the Vas Deferens.

CCDS Consensus Coding Sequence.

CFTR Cystic Fibrosis Transmembrane conductance Regulator.

CMV Cytomégalovirus.

ERAD Endoplasmic-reticulum-associated protein degradation.

FLIPR Fluorescent Imaging Plate Reader.

FrEx French Exome project.

FSK Forskoline.

HEK Human Embryonic Kidney.

HMM Hidden Markov Model.

HRP Horseradish Peroxidase.

Hsc70 Heat shock cognate 70.

HTML Hypertext Markup Language.

IBD Identical By Descent.

ICL IntraCellular Loop.

IF Immunofluorescence.

MSD Membrane Spanning Domain.

NBD Nucleotide Binding Domain.

NGS Next Generation Sequencing.

NMD Nonsense-Mediated Decay.

PCR Polymerase Chain Reaction.

PVDF PolyVinyliDene Fluoride.

QC Quality Control.

RE Réticulum Endoplasmique.

RFU RFU : Relative fluorescence units.

SAM Sequence Alignment/Map format.

SNV Single Nucleotide Variants.

TBS Tris-Buffered Saline.

TIR Trypsine Immuno-Réactive.

VCF Variants Call Format.

VEP Variant Effect Predictor.

VMD Visual Molecular Dynamics.

WT Wild-Type.

TABLE DES FIGURES

LISTE DES TABLEAUX

CHAPITRE 1

INTRODUCTION

Sommaire

1.1 L'étude du génome

Le génome correspond à l'ensemble du patrimoine génétique d'un individu. L'ADN ou acide désoxyribonucléique est la molécule qui porte cette information génétique et ce dans chaque cellule de tout organisme vivant. C'est une molécule constituée de deux brins antiparallèles qui forment une double hélice. Chacun de ces brins est un polymère appelé polynucléotide. L'élément de base de l'ADN est appelé nucléotide. L'ADN est donc la succession de quatre nucléotides différents dans un ordre bien précis. Les quatre nucléotides qui le composent sont l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T).

1.1.1 Développement du séquençage

Le séquençage de l'ADN consiste à obtenir la succession exacte des nucléotides d'une partie ou de l'ensemble du génome. Avant la découverte des techniques de séquençage de l'ADN de nouvelle génération, les chercheurs utilisaient principalement trois méthodes pour séquencer les génomes : la méthode de Sanger, la méthode de Maxam et Gilbert ou encore le pyroséquençage [Sanger et al., 1977, Maxam and Gilbert, 1977, Nyrén et al., 1993]. Cependant, ces technologies ne permettaient de séquencer que de courtes séquences du génome de l'ordre de 200 à 800 bases. Pour décoder le premier génome humain, achevé en 2003, il a fallu 13 ans et 3 milliards de dollars au Projet génome humain [International Human Genome Sequencing Consortium, 2001, Hattori, 2005]. Les premiers appareils de séquençage à haut débit permettant de séquencer un génome en entier sont apparus en 2005. Les innovations dans ce domaine ont permis de réduire le coût, ainsi que le temps nécessaire pour séquencer complètement les génomes. Ainsi le séquençage nouvelle génération (NGS, Next Generation Sequencing) se caractérise par l'utilisation d'approches massivement parallèles. Ces techniques permettent aujourd'hui de mettre en œuvre des projets de génomique à grande échelle dans un délai raisonnable et à un coût abordable. Il est désormais possible de séquencer un génome en moins d'une journée pour un coût inférieur à 1000$ [Hayden, 2014].

Il existe au sein de cette succession de nucléotides des variations d'un individu à un autre. On appelle variant toute position du génome qui est différente de la séquence de référence de l'organisme, dans notre cadre, le génome humain. On parle ainsi de SNVs, Single Nucleotide Variants, pour décrire les changement d'un seul nucléotide le long de la séquence d'ADN ; par exemple un A qui devient un C. Il existe également des modifications qui touchent plusieurs bases qu'on appelle des indels et qui peuvent être de deux sortes : des insertions avec l'ajout d'une ou plusieurs bases (A qui devient AC par l'ajout d'un C) ou à l'inverse, une délétion avec la perte d'une ou plusieurs bases (AC qui devient A). D'autres variations plus complexes, appelées variants structuraux, ne seront pas abordées dans le cadre de cette thèse.

Pour nommer les variations d'autres termes sont utilisés que variants. Les termes mutations et polymorphismes sont toujours utilisés mais avec des définitions historiques établis. Un seuil arbitraire de 1% a été fixé pour distinguer les variants communs (polymorphisme) des variants plus rares (mutation) [Karki et al., 2015]. Mais c'était avant que la séquence du génome humain soit décryptée. On a ensuite cherché à étudier la diversité génétique qui existe entre individus mais également à travers leur origine géographique. Il existe en effet des différences de fréquences de variants d'une population géographique à l'autre [The International HapMap Consortium, 2005, Leslie et al., 2015, Karakachoff et al., 2015]. Les variants classés comme

rares peuvent ainsi devenir des polymorphismes ou les polymorphismes devenir des variants rares selon la population analysée. L'étude de la diversité génétique est donc essentielle et nécessaire pour permettre l'identification de variants communs en population et/ou spécifiques d'une région géographique. Cela permet de mettre en évidence des variants plus rares qui peuvent être responsables de maladies [McEvoy et al., 2006, Saint Pierre and Génin, 2014]. Pour ce faire des projets de séquençage sur des individus provenant de différentes populations à travers le monde ont été lancés et ont permis d'identifier des SNVs communs et rares. Des projets, comme l'International HapMap Project et le 1000 Genomes Project, par exemple, ont permis d'obtenir un catalogue détaillé des variants génomiques présents dans les populations européennes, africaines et asiatiques [1000 Genomes Project Consortium et al., 2010, 1000 Genomes Project Consortium et al., 2015].

Le coût et le temps nécessaires au séquençage du génome en entier étant encore élevés, le séquençage restreint à des régions d'intérêt est plus couramment réalisé. Il est en effet aujourd'hui classique de ne séquencer que les parties codantes du génome constituées essentiellement des exons et que l'on appelle exome. Des méthodes de capture sont utilisées pour cibler spécifiquement ces régions exoniques et également d'autres régions d'intérêt sur le génome comme celles contenant des petits ARNs et des zones de régulation. L'exome représente environ 1% du génome soit environ 30Mpb [Xuan et al., 2013]. Même si le fait de séquencer et d'analyser uniquement ces régions ne prend pas en compte les variations dans les régions non codantes, il a été montré que cette stratégie permet de découvrir un grand nombre de variants rares responsables de maladies [Bamshad et al., 2011, Rabbani et al., 2014]. En effet, l'exome contiendrait 85% de ces variants [Ng et al., 2010]. Ces modifications de la séquence primaire de l'ADN peuvent altérer et/ou modifier le fonctionnement d'un gène. Des projets à grande échelle ont également vu le jour comme l'Exome Sequencing Project (ESP) et le French Exome Project (FrEx) [Fu et al., 2013, Genin et al., 2017]. La base de données GnomAD, regroupe et met à disposition les données de variants d'un très grand nombre de projet d'exomes et de génomes [Lek et al., 2016].

De plus en plus de projets voient le jour dans le but de comprendre comment est organisé le génome humain et comment il fonctionne. Les volumes de données à traiter, issus du séquençage, sont en constante augmentation. En effet le séquenceur, Hiseq X de chez Illumina, permettant de réaliser un génome pour moins de 1000$ génère 1.6-1.8 Tb par séquençage en moins de trois jours (http ://www.illumina.com). De plus, les données de séquençage ne sont pas directement exploitables. Ces données se présentent sous la forme de lectures (ou read). Le séquençage se

déroule en 4 étapes principales. Tout d'abord l'ADN est fragmenté en de courts fragments d'environ 250pb. Si l'on ne séquence pas le génome complet, on peut utiliser des librairies de « captures » (ensemble de sondes) produites par des entreprises telles que « agilent ou illumina » (http ://www.genomics.agilent.com). Elles permettent de sélectionner, par hybridation avec l'ADN fragmenté, les régions qui vont être séquencées. Après fixation de l'ADN sur le support de séquençage, les fragments d'ADN sont amplifiés par une technique de pontage à l'aide d'une ADN polymérase qui produit des ADN doubles brins. Ceci permet d'obtenir un très grand nombre de fragments identiques dans une zone appelée « cluster ». Le séquençage de ce cluster produira une lecture. Le séquençage se fait simultanément sur l'ensemble des clusters. Pour chaque cycle on ajoute d'abord les quatre nucléotides bloqués marqués avec un fluorochrome différent. Le nucléotide complémentaire au premier nucléotide de la séquence se fixe et on identifie le fluorochrome du nucléotide ajouté à l'aide d'un laser et d'une caméra microscope. On répète le processus le long de la séquence fixée sur le support (2 x 150pb). Les fluorescences obtenues à chaque nucléotide fixé, de l'ensemble des séquences, sont stockées dans des fichiers binaires en format bcl (base calls). Pour pouvoir analyser cette montagne de données, il est nécessaire de disposer d'outils informatiques performants. Sans le développement de la bioinformatique, le temps gagné au cours des années par l'avancée des méthodes de séquençage serait perdu en temps de traitement de données informatiques.

1.1.2 Traitement bioinformatique des données

1.1.2.1 Pipeline d'analyse

Afin de pouvoir étudier la variabilité génétique qui existe entre les individus, il est nécessaire d'obtenir des données traitées, lisibles par l'homme. Il faut donc passer par une succession de dizaines de programmes informatiques pour obtenir des données analysables. Les données analysables sont généralement stockées dans un fichier au format VCF (variant call format) [Danecek et al., 2011]. Il s'agit d'un tableau regroupant les milliers voire millions de variants observés. Le traitement des données brutes jusqu'à l'obtention d'un tableau regroupant les variants constitue le « pipeline d'analyse ». Il existe plusieurs algorithmes et outils pour construire ce pipeline, chacun présentant des paramètres différents [Mardis, 2008, Altmann et al., 2012] Il faut donc bien s'informer sur leur mode de fonctionnement afin de pouvoir choisir celui qui est le plus adapté aux analyses que l'on souhaite réaliser. L'analyse des données est permise par un ensemble d'étapes depuis le traitement des fichiers bruts (.bcl) jusqu'à l'identification de variants. Ces étapes [Voir Figure 1.1] sont indispensables et permettent le traitement des données générées par le séquenceur. Elles peuvent être regroupées en deux grandes parties, la première étant la plus standardisée et la

seconde visant à interpréter les données obtenues.

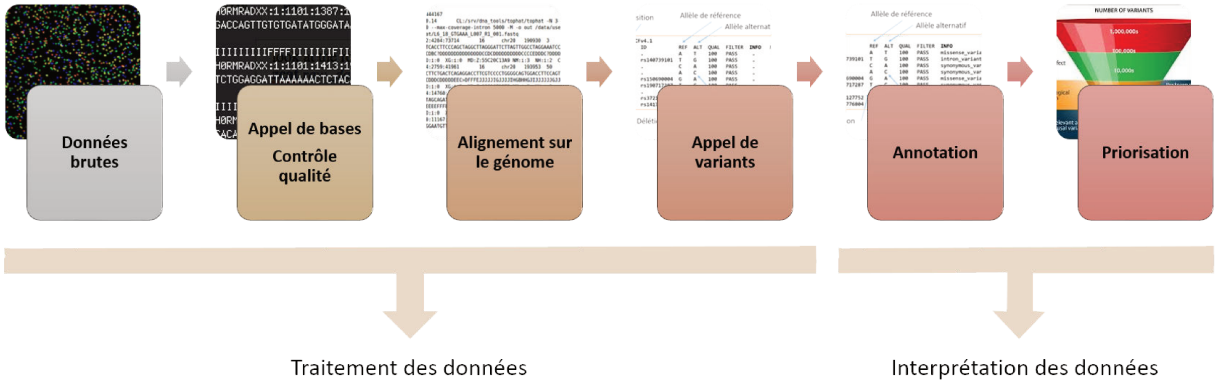


FIGURE 1.1: Pipeline d'analyse. La partie de traitement de données contient les étapes d'appel de bases (ou Base Calling), de contrôle de qualité (QC), d'alignement et l'identification des variants génomiques (Appel de variants ou Variant Calling).

La première partie contient les étapes d'appel de bases (ou Base Calling), de contrôle de qualité, d'alignement enfin l'identification des variants génomiques (ou Variant Calling). Le séquenceur permet d'obtenir la séquence de milliards de fragments d'ADN qui sont appelés lectures (ou reads). Les fichiers issus du séquenceur sont des fichiers binaires en format bcl (base calls). Ces fichiers comprennent les séquences obtenues par analyse d'images ainsi que le niveau de confiance avec lequel chaque base a été identifiée. La première étape est donc l'appel de base qui permet de transformer ces données en format plus lisible et moins volumineux, le format FASTQ (variante du format FASTA permettant d'intégrer une mesure de qualité). Les fichiers fastq contiennent les séquences des lectures avec, pour chaque lecture, son identifiant, sa position dans le séquenceur, la séquence et les scores QPhred de chaque base. Le score QPhred est un indice de qualité de la base séquencée [Ewing et al., 1998].

Il s'en suit des étapes de contrôle de qualité (ou Quality Control généralement abrégé par QC). Ce contrôle qualité permet de vérifier la qualité du séquençage. On peut ainsi mettre en évidence les séquences dupliquées ou surreprésentées, les déviations par rapport à la teneur en GC, la distribution et la qualité des nucléotides par position. Lorsque que l'on réalise un séquençage, l'ADN de départ est fragmenté de manière aléatoire et chaque fragment d'ADN est séquencé une fois sous forme de

cluster comme vu précédemment. On ne s'attend donc pas à avoir un pourcentage très élevé de séquences dupliquées ou surreprésentées qui signifierait qu'une même séquence a été séquencée plusieurs fois (Figure 1.2.A). On analyse également la teneur en nucléotides GC. Dans un génome normal, on s'attend à voir une distribution à peu près normale du contenu en GC dont le pic central correspond au contenu global en GC du génome (Figure 1.2.B). Une distribution de forme inhabituelle pourrait indiquer une contamination ou un autre type de biais (Figure 1.2.B.b). Lorsqu'on analyse la distribution et la qualité des nucléotides par position, on peut observer par exemple que la probabilité d'erreur de séquençage augmente avec la taille des reads (Figure 1.2.C.a). Si l'on observe une diminution de la qualité des reads aux extrémités, des outils bioinformatiques suppriment les bases de mauvaise qualité à l'aide du score QPhred par exemple. Cela permet de traiter que les bases de bonne qualité et d'augmenter la qualité des lectures obervés (Figure 1.2.C.b) [Joshi and Fass, 2011]. On peut par ce contrôle de qualité, identifier des problèmes de préparation des échantillons et de séquençage. Une fois que les données brutes ont été vérifiées, on peut aligner les séquences d'ADN sur le génome de référence (Figure 1.3). Chaque séquence d'ADN est positionnée sur le génome de référence à l'aide de programmes bioinformatiques tel que Burrows-Wheeler Aligner (BWA) [Li, 2013].

Les fichiers de stockage de cet alignement sont les fichiers SAM (Sequence Alignment/Map format). Ils contiennent des informations sur la position à laquelle la lecture s'est alignée sur le génome de référence, la séquence de la lecture, le code CIGAR et la qualité d'alignement. Le code CIGAR indique comment la lecture s'est alignée sur la référence. Par exemple si le code est 11M4D58M cela signifie que la lecture s'est correctement alignée sur 11 bases puis il y a 4 bases délétées et 58 identités (Figure 1.4). Il est ensuite possible d'utiliser des programmes afin d'identifier des variants génomiques.

L'interprétation des données est d'autant plus difficile que la quantité d'information obtenue est importante. Beaucoup de variants sont mis en évidence. Par exemple, dans le génome d'un individu, environ trois millions de positions diffèrent du génome humain de référence. La deuxième partie du pipeline d'analyse a donc pour objectif de trouver le ou les variants répondant à la question biologique posée. Il faut pour cela filtrer ces variants. Pour ce faire, des outils informatiques et des bases de données permettent aujourd'hui d'annoter les fichiers de variants.

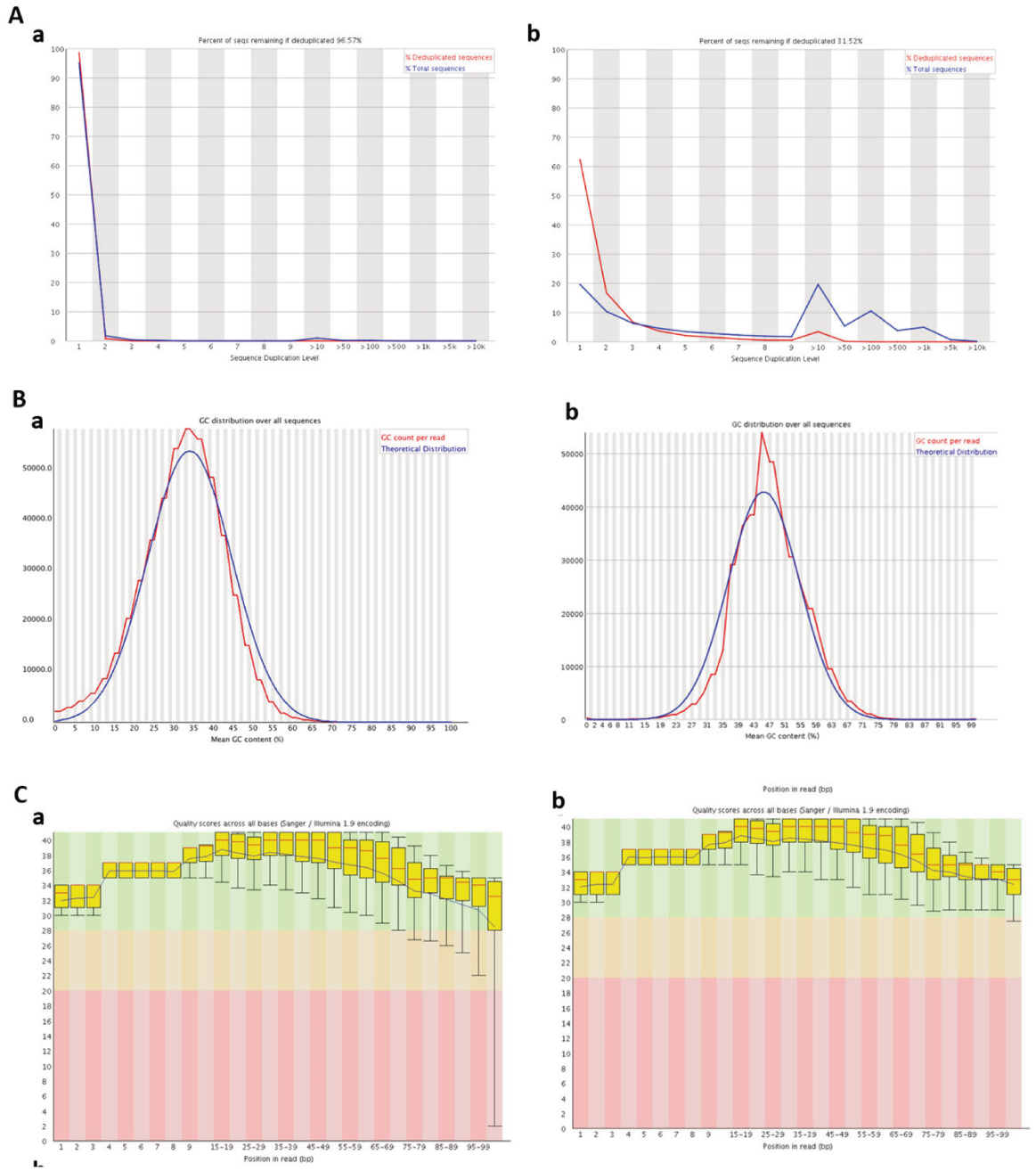


FIGURE 1.2: Exemple de résultats obtenus par l'outil FastQC pour le contrôle de qualité. A. Diagramme de la proportion de lectures provenant de séquences qui se produisent un nombre différent de fois. a. Presque toutes les séquences sont observées qu'une seule fois. Seulement 3% environ des lectures seraient perdues si on enlève les duplicatas (remaining 96,57%) b. Ici seul moins du tiers des séquences seraient conservées (remaining 31,52%). B. Diagramme du pourcentage de GC par séquence comparé à une distribution normale modélisée du contenu GC sans ("a") ou avec ("b") contamination. C. Diagramme de la qualité par séquence de base avant ("a") et après ("b") l'ajustement de la qualité. Pour chacune des positions de base ("axe des X"), les scores de qualité QPHRED sont tracés ("axe des Y"), les scores les plus élevés représentant les meilleurs appels de base. Les couleurs de fond vert, orange et rouge représentent respectivement les appels de base de bonne, raisonnable et de mauvaise qualité.

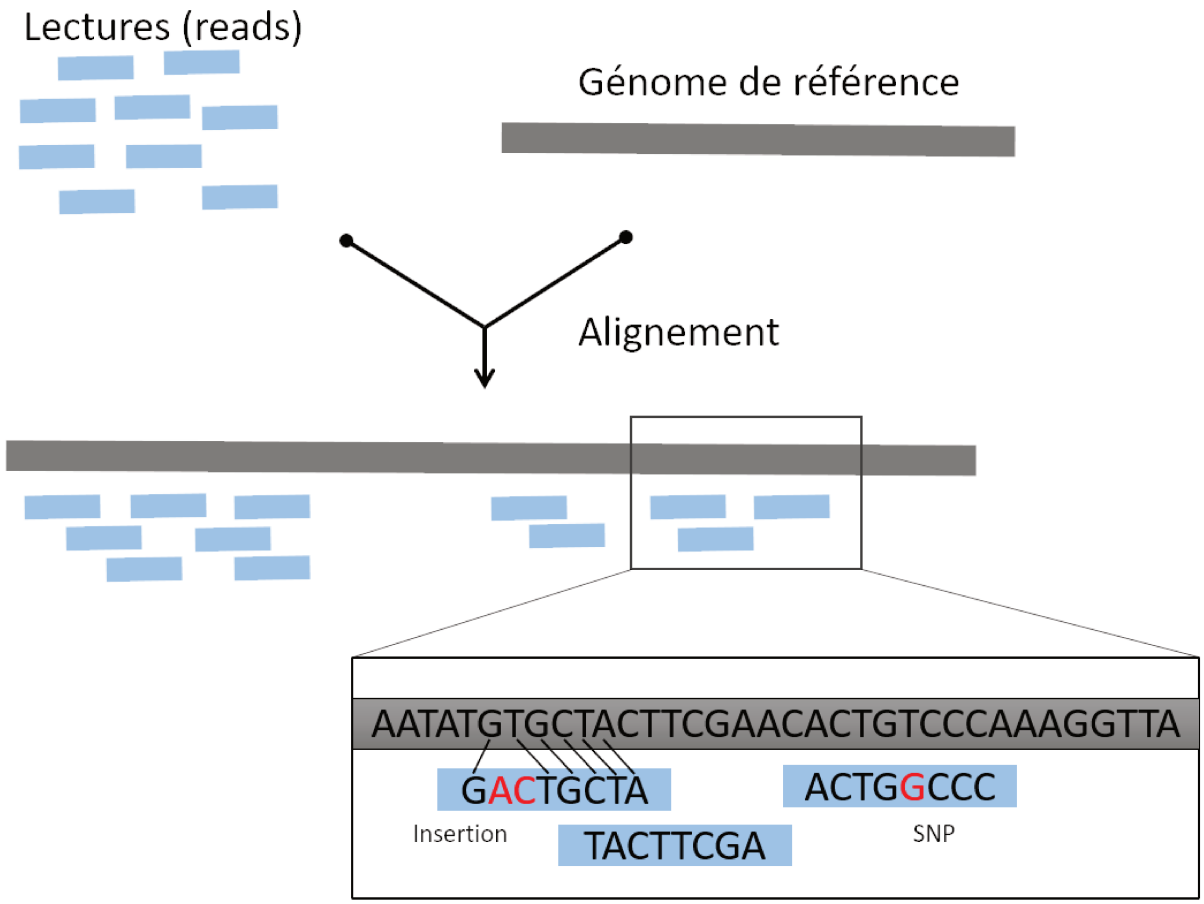


FIGURE 1.3: Illustration de l'alignement sur le génome. Les lectures (ou reads) correspondent aux séquences d'ADN obtenues par séquençage. Chaque séquence d'ADN est positionnée sur le génome de référence. L'identification des variants génomiques se fait par un autre programme qui permet d'avoir une liste des variants issues de cet alignement. Sur l'exemple, le read 1 a une insertion de AC par rapport à la séquence de référence, le read 2 est une substitution d'un T en G et le read 3 est parfaitement aligné sur le génome.

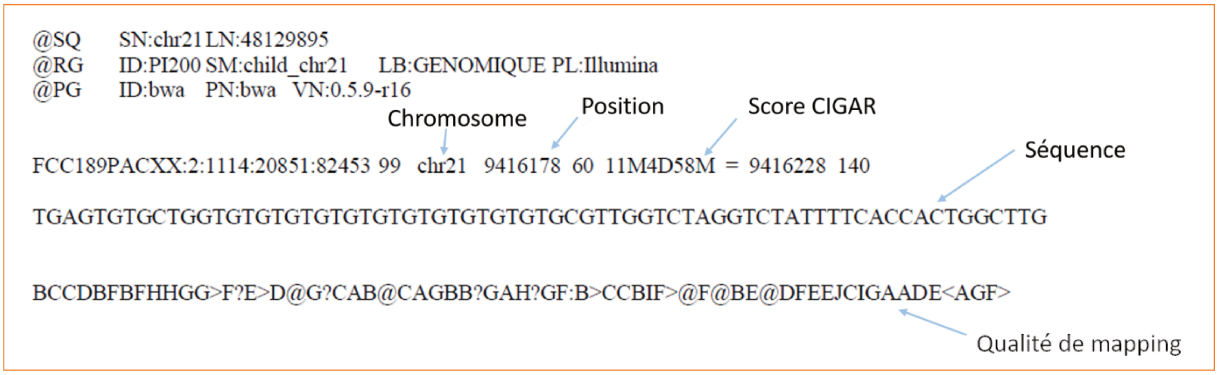


FIGURE 1.4: Représentation d'une lecture alignée sur le génome au format SAM. Cette lecture est alignée sur le génome au niveau du chromosome 21 position 9416178.

1.1.2.2 Le fichier VCF (Variant Call Format)

Le fichier VCF est un fichier standard pour stocker les variants de séquence de gènes [Danecek et al., 2011].Dans ces fichiers, on trouve à chaque ligne un variant avec le chromosome (CHROM), la position génomique (POS), l'allèle de référence (REF), l'allèle alternatif (ALT). Nous avons donc soit des SNVs qui touchent une seule position, soit les indels, suppression ou insertion d'un ou plusieurs nucléotides. Ce fichier peut contenir les informations pour un ou plusieurs individus et le génotype de chaque individu aux positions génomiques où se situent les différents variants (Figure 1.5). Prenons l'exemple du dernier variant de la figure 1.5. Il est localisé sur le chromosome 1, à la position 69590 et il possède un identifiant dbSNP rs141776804 (base de données répertoriant les variants) [Kitts and Sherry, 2002]. L'allèle de référence, présent sur le génome de référence, est un T et l'allèle alternatif, donc le variant, est une insertion d'un C (T>TC). Dans la colonne « INFO », on retrouve une information fournie par un annotateur « intron_variant » qui signifie que ce variant est localisé dans une région intronique. Les deux dernières colonnes représentent le génotype de deux individus. Un 0 représente l'allèle de référence et le 1 l'allèle alternatif. Le premier individu (Sample1) a une génotype 0|0 ce qui signifie qu'il ne possède pas ce variant et n'a donc pas d'insertion d'un C dans son génome à cette position. Pour cette position, on dit qu'il est homozygote pour l'allèle de référence. En revanche, le deuxième individu (Sample2) est hétérozygote avec le génotype 1|0 pour ce variant c'est-à-dire qu'il porte ce variant sur l'un de ses deux chromosomes (Figure 1.5).

1.1.2.3 Filtres des données

Les critères de filtre les plus couramment utilisés sont (1) la fréquence du variant en population, (2) la localisation et l'impact du variant dans le génome, (3) le gène dans

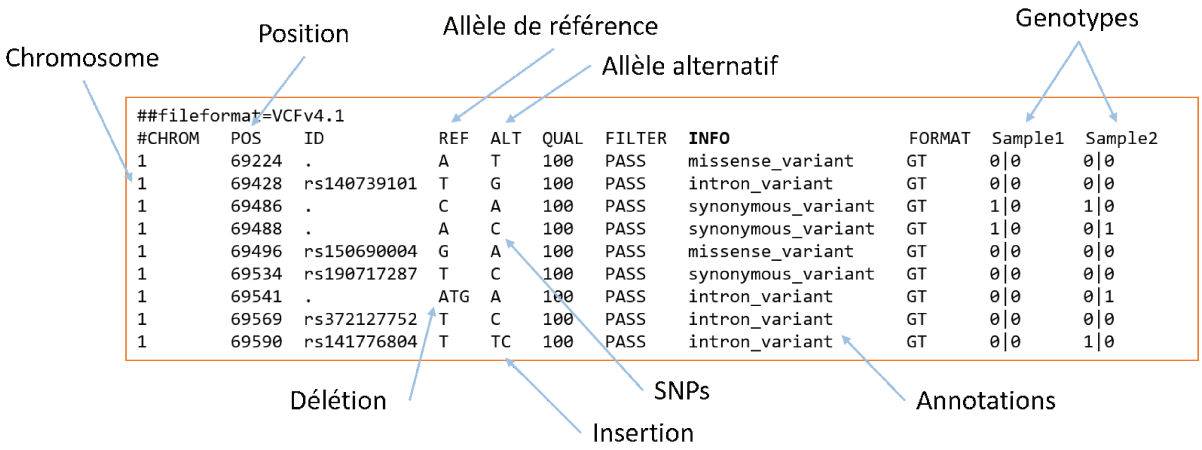


FIGURE 1.5: Illustration d'un fichier au format VCF (Variant Call Format).

lequel il se trouve (notamment s'il s'agit d'un gène déjà connu pour être impliqué dans la pathologie soupçonnée) et (4) l'effet prédit pour le variant par des outils de prédiction d'effet tels que SIFT ou PolyPhen [Kumar et al., 2009, Adzhubei et al., 2013]. Pour obtenir ces différentes informations, des outils d'annotation sont disponibles, les plus connus et les plus utilisés sont SnpEff [Cingolani et al., 2012], Variant Effect Predictor (VEP) [McLaren et al., 2016] et Annotation of Genetic Variants (Annovar) [Wang et al., 2010]. Chaque outil récupère les informations dans des bases de données et, à l'aide d'algorithmes, annote chaque variant du fichier VCF un à un. L'inconvénient de l'utilisation de ces filtres est qu'ils peuvent être soit trop stricts avec un risque de faux négatifs et donc de perdre les variants d'intérêt, soit au contraire trop permissifs, conduisant à trop de faux positifs qu'il ne sera pas possible de valider par des études fonctionnelles. Il est donc nécessaire de trouver le bon équilibre entre sensibilité et spécificité. Pour ce faire, il faut bien définir la question biologique posée afin de pouvoir appliquer des filtres et ainsi réduire, mais raisonnablement, le nombre de variants à analyser.

1.1.3 De l'analyse du génome à la protéine

Pour obtenir une protéine, un gène est d'abord transcrit en acide ribonucléique (ARN). Avant de devenir un ARN messager dit mature (ARNm), l'ARN pré-messager est synthétisé puis épissé dans le noyau. Pendant l'épissage, les introns sont retirés et les exons sont joints. D'un même ARN pré-messager il peut résulter plusieurs ARNm en faisant varier sa composition en exons. Les exons peuvent alors être retenus ou ignorés et/ou des introns peuvent être conservés. Ce phénomène est appelé épissage alternatif [Black, 2003, Lee and Rio, 2015]. L'ARNm est ensuite traduit en protéine. Lors de ce processus, l'information écrite avec quatre lettres - A, C, G, T de l'ADN,

puis A, C, G, U de l'ARN est décodée vers un alphabet à vingt lettres : les vingt acides aminés composant les protéines. Cette opération est réalisée par le ribosome en lisant la séquence de l'ARN par groupes de trois bases, chaque triplet ou codon indiquant un acide aminé de la séquence de la protéine en cours de synthèse. La correspondance entre codons et acides aminés est définie par le code génétique. Il s'en suit une étape de modification post-traductionnelle et de repliement de la protéine. Toute une machinerie est mise en œuvre afin qu'à terme la protéine puisse jouer son rôle au sein de la cellule (Figure 1.6).

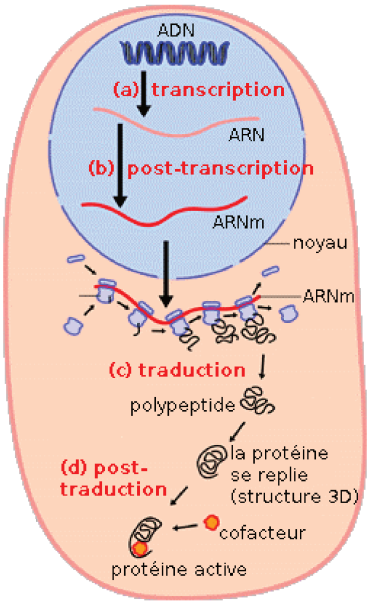


FIGURE 1.6: Biosynthèse des protéines. (a) Un gène est d'abord transcrit en acide ribonucléique (ARN). (b) Avant de devenir un ARN messager dit mature (ARNm), l'ARN pré-messager est synthétisé puis épissé dans le noyau. (c) L'ARNm est ensuite traduit en protéine. (d) La protéine passe par des étapes de modifications post-traductionnelle qui vont permettre son repliement.

Les modifications de la séquence primaire de l'ADN peuvent donc changer la séquence primaire de la protéine codée par le gène muté et donc altérer ou modifier son fonctionnement normal. Les variants par substitution, les SNVs, peuvent avoir différents impacts sur la séquence protéique. On distingue les variants faux-sens, qui remplacent un acide aminé par un autre et les variants non-sens (Figure 1.7). Les variants faux-sens peuvent aussi bien impacter (1) l'ATG initiateur, correspondant à une méthionine, qui initie le début de la traduction, (2) qu'un codon au sein de la séquence qui va être traduite ou (3) le codon stop, correspondant dans le code génétique au signal pour le ribosome de l'arrêt de la traduction. La modification de l'ATG initiateur entraîne soit une absence de production de la protéine soit une protéine différente par l'initiation de la traduction par un deuxième ATG plus loin dans la séquence. Dans le cas d'un variant touchant le codon stop, la protéine ainsi produite sera élonguée. Les variants non-sens correspondent aux variants qui remplacent un

acide aminé par un codon stop. L'apparition d'un variant non-sens peut entraîner la synthèse d'une protéine tronquée.

Les variants synonymes, quant à eux, ne modifient pas la séquence protéique, du fait de la redondance du code génétique (Figure 1.7). Ces variants sont néanmoins importants à prendre en compte car ils peuvent modifier des sites de fixation d'éléments régulateurs ou d'épissage mais également changer la structure secondaire et la stabilité de l'ARN [Chamary et al., 2006, Shabalina et al., 2013]. Ces modifications ont donc un effet sur l'expression de la protéine bien que sa séquence primaire ne soit pas changée.

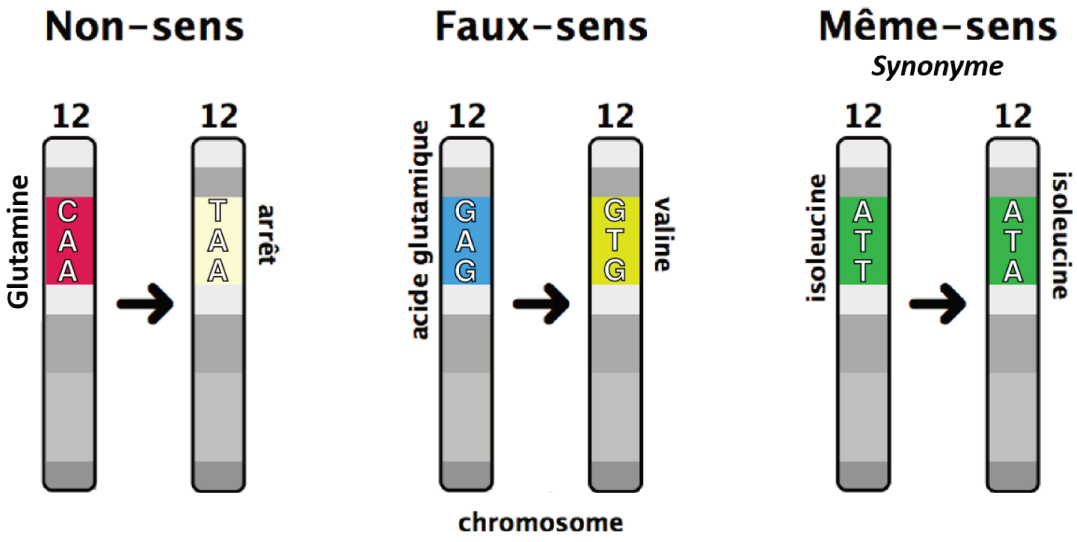


FIGURE 1.7: Trois groupes de variants par substitution dans les régions condantes.

Les insertions et les délétions sont des variants décalants. On distingue les insertions et délétions simples qui sont appelées indels et celles qui changent le cadre de lecture et qui sont appelées en anglais « Frameshift ». En effet, une addition ou une suppression de nucléotides non multiples de 3 provoquera un changement de cadre de lecture du code génétique (Figure 1.8). Au moment de la traduction, cela générera le plus souvent une protéine tronquée par l'apparition d'un codon stop prématuré et le remplacement de la plupart des acides aminés entre le variant et le codon stop.

Ces variants, qui modifient la séquence primaire de la protéine, peuvent avoir des effets sur la fonction des protéines à différents niveaux. La protéine présentant, par exemple, un codon stop prématuré peut être prise en charge en amont de sa traduction par le nonsense-mediated decay (NMD) ou dégradation des ARNm possédant un codon stop prématuré. Ce mécanisme permet d'éviter la traduction d'une protéine anormale qui pourrait avoir un rôle différent, potentiellement délétère, dans la cellule. La position d'apparition de ce codon stop prématuré au sein de l'ARNm détermine sa prise en charge ou non par ce mécanisme de protection [Chang et al., 2007, Palacios,

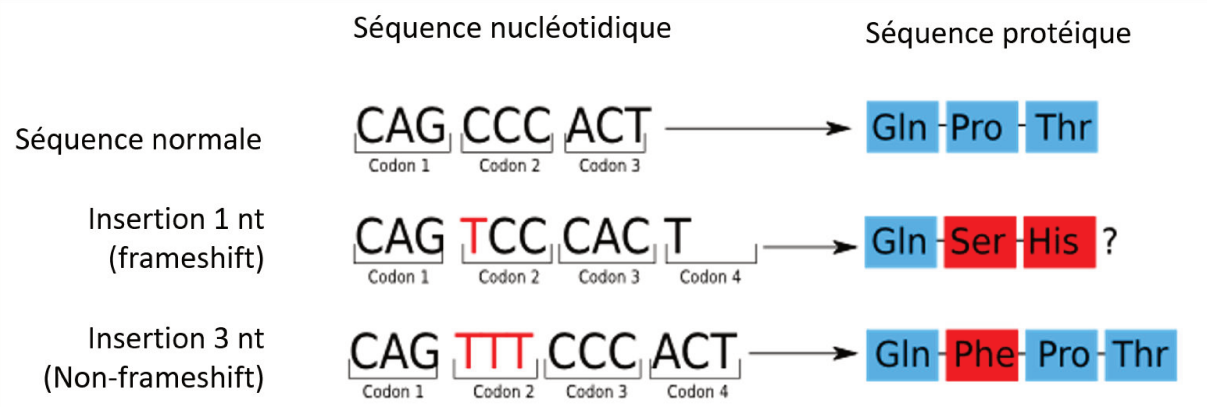


FIGURE 1.8: Représentation d'une insertion modifiant ou non le cadre de lecture. Dans le premier cas, l'insertion T va entraîner le changement de la plupart des acides aminés suivant l'insertion jusqu'à l'apparition d'un codon stop. Dans l'autre cas, c'est une insertion de 3 T, multiple de trois, entraînant l'insertion d'un acide aminé (Phénylalanine) mais cela n'impactera pas le reste de la séquence protéique.

2013, Coban-Akdemir et al., 2018]. En résumé, lors de l'apparition d'un codon stop prématuré soit la quantité de protéine fonctionnelle est réduite (car le transcrit est dégradé par le processus NMD), soit une protéine tronquée est produite qui peut être délétère (car le transcrit a échappé au NMD) [Inoue et al., 2004]. Il existe également des variants qui peuvent affecter l'épissage. De tels variants entraînent des erreurs lors du processus d'épissage et peuvent conduire à la rétention d'une partie ou de la totalité d'un intron. La modification de la séquence nucléotidique de l'ARNm, avant traduction, entraîne un décalage du cadre de lecture [Anna and Monika, 2018]. Ceci engendre des ARNm aberrants, qui lors de la traduction produisent des protéines plus longues ou plus courtes. De telles protéines ne sont alors pas fonctionnelles ou ont une fonction différente qui peut déstabiliser la cellule. Nous verrons d'autres exemples d'effets de ces variants lors de la description des variants affectant le gène CFTR responsable de la mucoviscidose.

1.2 Les haplotypes

1.2.1 Notion d'haplotype

Malgré les progrès rapides du séquençage, la plupart des études sur la génomique humaine ont accordé peu d'attention à la nature diploïde du génome. Celui-ci possède généralement deux copies de chaque chromosome, l'une héritée de la mère et l'autre du père. Quand la base diffère de la même façon sur chacun des chromosomes, l'individu est homozygote muté. A l'échelle d'un gène, on appelle haplotype la succession des positions polymorphes (ou allèles) le long de la séquence

nucléotidique sur chacun des chromosomes. Lorsque qu'il a plus d'une position qui diffère de la séquence de référence nous avons alors deux possibilités de localisation de ces variants sur les chromosomes. (1) Les configurations *cis* avec deux variants ou plus sur le même chromosome laissent un exemplaire du gène, celui situé sur l'autre chromosome, non perturbé, tandis que (2) les configurations *trans* avec des variants sur les deux chromosomes peuvent affecter les deux exemplaires du gène (Figure 1.9).

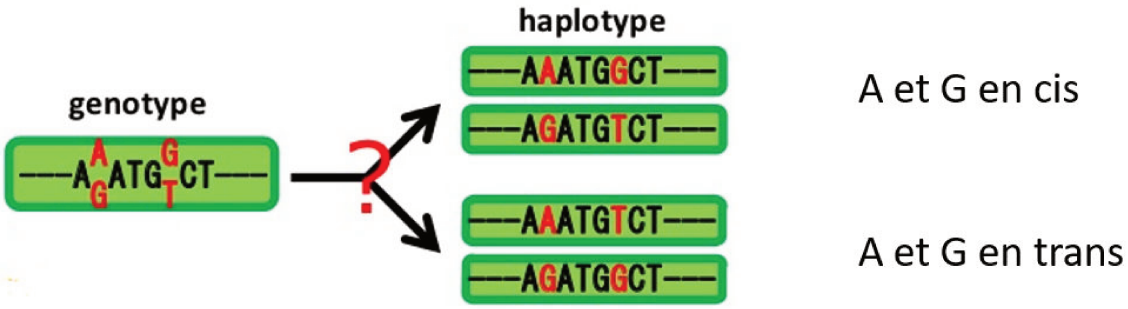


FIGURE 1.9: Définition d'un haplotype. Sur cette figure, nous pouvons voir un génotype avec deux variants hétérozygotes. Dans ce cas, nous avons deux possibilités : A et G sont sur le même chromosome et ils sont dits en *cis* ou A et G sont sur un chromosome différent et ils sont dit en *trans*.

1.2.2 Reconstruction des haplotypes

Pour reconstruire les haplotypes et attribuer les variants sur l'un ou l'autre des haplotypes, on parle d'haplotypage ou de phasage. Deux types de méthodes existent : des méthodes de phasage moléculaire et des méthodes de phasage statistique.

1.2.2.1 Le phasage moléculaire

Le phasage moléculaire permet d'obtenir la séquence exacte de chaque brin d'ADN. Assez fastidieux, il est réalisé sur des petites régions inférieures à 20kb. Il consiste en la réalisation d'amplifications en chaîne par polymérase (PCR), de clonage et de séquençage Sanger [Paul and Apgar, 2005, Chen and Schrijver, 2011]. Grâce aux développements technologiques comme la PCR numérique et le séquençage, on pourra bientôt utiliser les informations de phasage directement en sortie des automates. La PCR numérique permet de s'affranchir des étapes coûteuses en temps, comme le clonage et d'être moins limité par la taille des amplicons. Avec cette technique, on peut donc d'obtenir rapidement l'information sur la phase de séquences d'ADN plus longues [Regan et al., 2015]. Il est également possible d'utiliser des techniques de séquençage long-reads qui produisent des lectures plus longues

que celles obtenues à partir des séquenceurs classiques et conservent ainsi les informations sur les haplotypes [Rhoads and Au, 2015].

1.2.2.2 Le phasage statistique

Le phasage statistique se sert quant à lui des données sur les combinaisons de génotypes observées dans l'échantillon séquencé ou dans des échantillons de référence pour inférer la phase la plus probable. Ainsi, si on reprend l'exemple de la Figure 1.9, le génotype de l'individu est hétérozygote A, G à la première position et hétérozygote G, T à la deuxième. Nous cherchons donc à savoir si les haplotypes de l'individu sont AG, GT ou AT, GG. Si dans le reste de l'échantillon dont fait partie cet individu ou dans un panel de référence, l'haplotype GG n'a jamais été observé mais si on a observé les haplotypes AG et GT, on donnera une forte probabilité à la première combinaison AG, GT et une probabilité nulle ou quasi-nulle à la seconde combinaison.

Différentes méthodes statistiques ont été développées pour réaliser ce phasage et les méthodes les plus efficaces pour réaliser le phasage de grandes régions génomiques voire de chromosomes entiers se basent sur des modèles de chaînes de Markov cachées (HMM) comme les outils bioinformatiques SHAPEIT2, Beagle, Eagle2 ou HAPI-UR [Browning and Browning, 2007, Delaneau et al., 2012, Williams et al., 2012, Loh et al., 2016].

Certains outils utilisent, lorsqu'elles sont disponibles, les informations provenant des apparentés afin de prendre en compte les régions identiques par descendance (IBD) (Figure 1.10). Dans le cas simple où le père de l'individu est homozygote AG, AG et la mère homozygote GT, GT, il sera facile de conclure que l'individu est AG, GT car il aura acquis un chromosome de son père et l'autre de sa mère. Il est important de souligner que la conclusion n'est pas aussi simple lorsque les parents ne sont pas homozygotes pour les deux allèles. Lors de la méiose, division cellulaire permettant d'obtenir une cellule haploïde (gamète) à partir d'une cellule diploïde, il survient des évènements de recombinaisons génétiques (Figure 1.11). Des chromosomes homologues peuvent ainsi échanger des portions de leurs ADN. Il en résulte des chromosomes portant une séquence d'ADN différente de celle du parent. Par exemple, un individu AG, CC peut transmettre l'haplotype AC ou CG après une recombinaison homologue entre les deux allèles. Ces évènements doivent être intégrés à l'algorithme de programmation lors de la conception de ces outils statistiques.

Certains outils se servent des « reads ». Lorsque deux variants sont présents sur le même « read », on peut en effet en déduire qu'ils sont en *cis*. Pour que deux variants soient sur le même « read », il faut qu'ils soient situés à une distance assez courte ($\sim$250 bases pour le séquenceur Illumina). L'outil de phasage statistique SHAPEIT2

permet à la fois d'utiliser des panels de référence, de se servir de l'information des apparentés et des « reads » [Delaneau et al., 2012, Delaneau et al., 2013, O'Connell et al., 2014]. C'est actuellement l'outil de phasage le plus utilisé et le plus efficace en termes de fiabilité et de rapidité [Choi et al., 2018]. Il a d'ailleurs été utilisé pour phaser les données du projet 1000 Genomes [Delaneau et al., 2014].

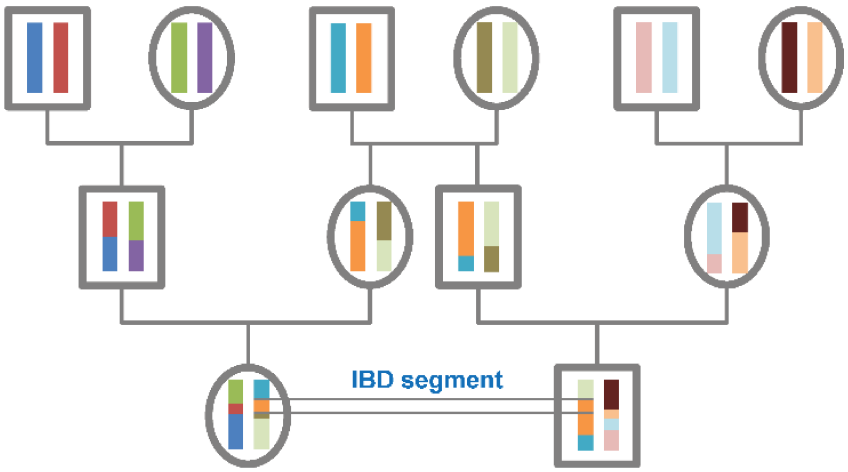


FIGURE 1.10: Représentation de segments identiques par descendances.Chaque carré (homme) et chaque cercle (femme) représentent un individu avec deux chromosomes homologues sous forme de barres. Chaque rangée correspond à une génération. Dans la rangée du bas, les petits enfants ont hérité du segment représenté en orange qui correspond à une partie du chromosome de leur grand-père. Ce segment de chromosome est donc identique par descendance pour ces deux individus.

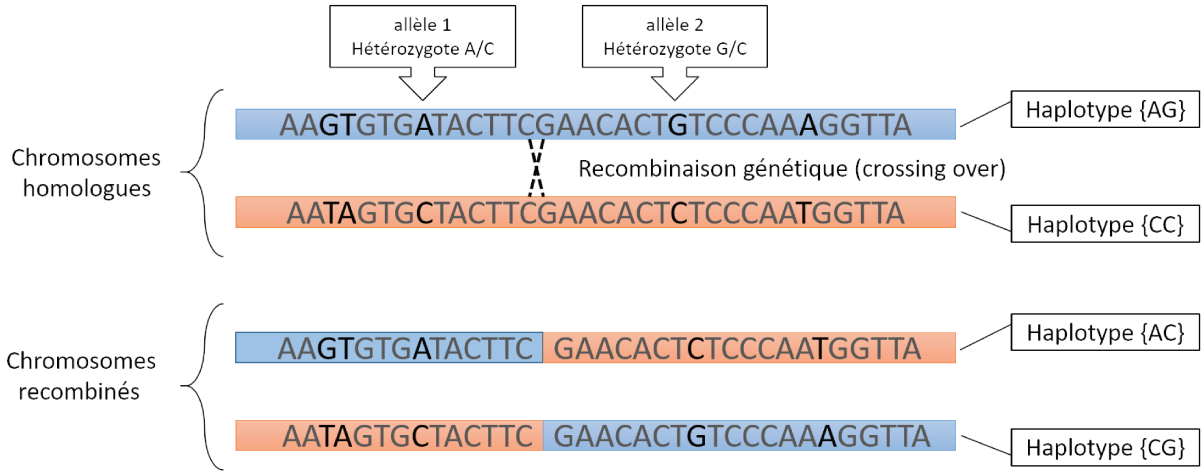


FIGURE 1.11: Évènement de recombinaison génétique entre deux chromosomes homologues. Lors de la méiose des recombinaisons génétiques peuvent se produire. L'individu ayant un génotype AG, CC peut transmettre à ses descendants l'haplotype AC ou CG après une recombinaison homologue entre les deux allèles.

1.2.3 Place des haplotypes et impact des variants

Lorsqu'on analyse des données issues du séquençage du génome ou lorsqu'on réalise des tests fonctionnels in vitro, le contexte haplotypique est rarement pris en compte. Cela pose des problèmes lors de la recherche de variants causaux ou pour l'interprétation des résultats des tests in vitro.

1.2.3.1 Analyse de données de séquençage

Ainsi, comme nous l'avons vu dans la description du pipeline d'analyse (1.1.2.1), le traitement des données de séquençage aboutit à une liste de variants dans un fichier VCF. Chaque ligne de ce fichier correspond à un variant et les annotations et les filtres qui sont réalisés considèrent le variant isolement de son contexte haplotypique. En d'autres termes, on ne tient pas compte des autres variants avec lesquels le variant à la ligne n du fichier est associé en *cis* chez l'individu *i*.

Si l'on revient sur les annotateurs et l'utilisation de filtres sur les variants, la manière dont sont appliqués les filtres peut avoir des conséquences importantes sur l'interprétation des données (1.1.2.3). Cela peut par exemple entraîner l'exclusion de variants répondant à la question biologique. Ainsi, le filtre de fréquence permet de réduire le nombre de variants en ne gardant pour la suite des analyses que les variants dont la fréquence dans les bases de données publiques comme 1000 Genomes ou GnomAD est inférieure à un certain seuil. Ce seuil est souvent déterminé avec la prévalence de la maladie étudiée.

Par exemple, pour une maladie dominante à pénétrance forte dont la prévalence en population est de 1 pour mille, on ne retient que les variants ayant une fréquence inférieure à 0.1%. Ainsi, on ne s'intéresse pas à un variant ayant une fréquence de 1% en population générale qui pourrait conduire à la maladie lorsqu'il est associé en *cis* avec un second variant du gène ayant lui une fréquence de 10%. Ces deux variants sont exclus dès les premières étapes de l'analyse et leur implication dans la maladie n'est donc jamais étudiée.

De même, les algorithmes de prédiction d'effets des variants, utilisés pour exclure des variants neutres, basent leur prédiction uniquement sur une information locale sans prise en compte des autres variants situés en *cis* chez l'individu et qui pourraient modifier l'effet du variant. Il existe bien pourtant quelques exemples d'implication dans des pathologies des variants individuellement neutres qui, lorsqu'ils sont présents sur le même haplotype (en *cis*), peuvent jouer un rôle dans la maladie et conduire à des erreurs de diagnostic. Nous avons l'exemple d'un variant dans le gène *EGFR*, devenu une cible thérapeutique importante pour le traitement des adénocarcinomes pulmonaires. Ce variant a mis à mal les tests de variants dans le cadre du diagnostic.

Ce variant empêche l'amplification de l'ADN portant les deux mutations et induit un test faussement négatif. Cette erreur de diagnostic peut avoir des conséquences lourdes quant à la pris en charge des patients [Santamaría et al., 2013]. Deux exemples de combinaisons de variants modifiant le phénotype de patients atteints de la mucoviscidose et d'hémochromatose, sont décrits plus loin dans le mémoire [Clain et al., 2001, Uguen et al., 2017].

1.2.3.2 Etude *in vitro* de l'impact des variants génomiques

Lors de l'étude de l'impact d'un variant génomique, on réalise le plus souvent une transfection, transitoire ou stable, de la séquence mutée du gène, ou le plus souvent d'un transcrit du gène, par le biais de plasmide ou de rétrovirus (Figure 1.12). La transfection est une technique de biologie moléculaire qui consiste à introduire un ADN étranger dans une cellule eucaryote cultivée in vitro. Dans le cas d'un transfection transitoire, on peut utiliser des plasmides. La séquence d'ADN complémentaire (ADNc) permettant l'obtention de la protéine d'intérêt est introduite dans un plasmide. La cellule est traitée avec un transfectant qui va permettre au plasmide de rentrer dans la cellule. Le plasmide rejoint le noyau de la cellule. Grâce au promoteur présent sur le plasmide, la transcription, puis la traduction en protéine vont pouvoir se faire. Dans ce cas, la transfection est dite transitoire car l'ADN étranger ne s'intègre pas au génome de la cellule hôte. Les cellules sont analysées 24 à 72 heures après transfection, correspondant au temps où l'expression de la protéine nouvellement synthétisée est la plus importante. Dans le cas d'une transfection stable l'ADNc s'intégre au génome de la cellule hôte. Ainsi la cellule synthétise ensuite elle-même l'ADN introduit. Cet ADN va être par la suite transcrit puis traduit en protéine (Figure 1.12).

La séquence mutée du gène, introduit dans la cellule, est obtenue en partant de la séquence de référence humaine sur laquelle on effectue une mutagenèse dirigée [Raraigh et al., 2018]. C'est-à-dire que l'on induit un variant précis au sein de l'ADN. Lors de ces analyses, on ne se soucie généralement pas des variants qui pourraient être associés au variant d'intérêt. Dans le cas où d'autres variants sont associés en *cis*, les résultats de ces études ne sont donc pas toujours le reflet de l'effet du variant dans son contexte physiologique.

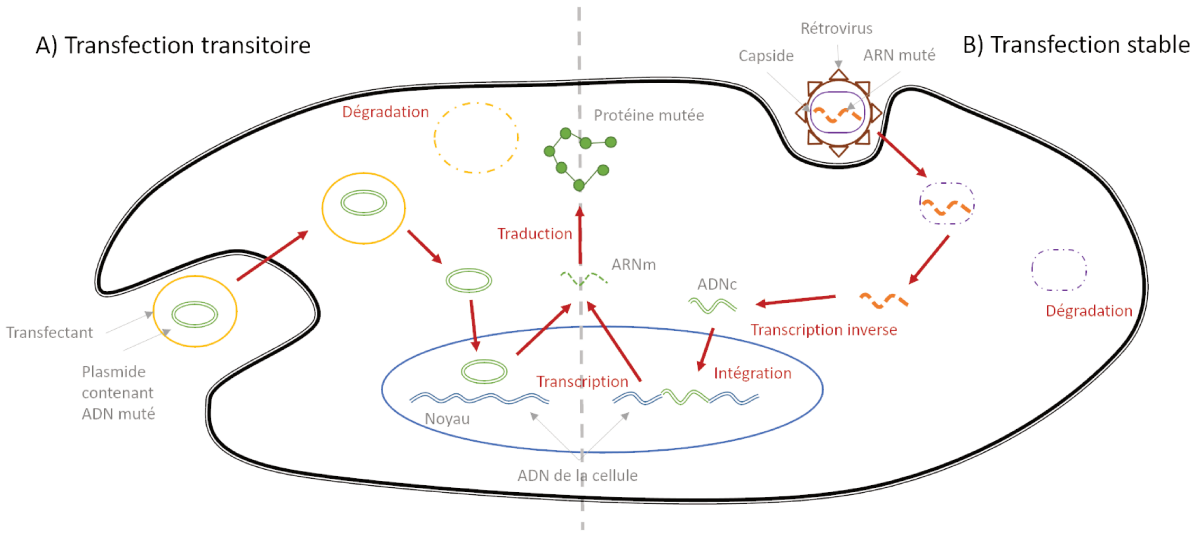


FIGURE 1.12: Transfection transitoire et stable. A. Principe de la transfection transitoire à l'aide d'un plasmide. B. Principe de la transfection stable à l'aide d'un rétrovirus.

1.3 La mucoviscidose

La mucoviscidose est la maladie autosomique récessive létale la plus fréquente chez les Caucasiens. En France, elle a une incidence de 1 naissance sur 4500 avec une incidence plus élevée dans le nord-ouest avec 1 naissance sur 2500 [Girodon-Boulandet and Costa, 2005]. L'étude du génome et le séquençage ont permis une meilleure compréhension de cette maladie et surtout son origine.

1.3.1 La maladie

1.3.1.1 Généralité

Au Moyen Age, on pensait que les enfants atteint de mucoviscidose étaient maudits et condamnés à mourir [Yu and Sharma, 2019]. Ils étaient distingués des autres enfants par la particularité d'avoir la « peau salée », « le goût salé d'un baiser sur le front ». L'espérance de vie, à l'époque, était d'environ 6 mois et les enfants mourraient le plus souvent d'infections pulmonaires [Davis, 2006]. Avant 1938, les enfants souffrants de mucoviscidose étaient diagnostiqués comme atteints d'une forme de maladie cœliaque, de par les troubles digestifs [Andersen, 1938].

C'est la pédiatre américaine Dorothy Hansine Andersen qui considéra la maladie comme étant une entité à part entière, sous le nom de Fibrose kystique du pancréas. Ce sont les lésions histologiques spécifiques du pancréas, découvertes lors des autopsies des enfants décédés, qui lui ont donné son nom. Ce n'est que plus tard, en

1943 que la maladie a vu son nom évoluer en mucoviscidose pour « mucus visqueux » car décrite comme une sécrétion généralisée de mucus visqueux ne se limitant pas au pancréas [Farber et al., 1943].

En 1989, le gène à l'origine de la maladie a été découvert, il s'agit du gène CFTR (Cystic Fibrosis Transmembrane conductance Regulator) qui code pour la protéine du même nom [Riordan et al., 1989, Rommens et al., 1989, Kerem et al., 1989]. Des études ont alors été conduites pour comprendre le rôle joué par cette protéine membranaire et son lien avec la maladie. On sait aujourd'hui que la protéine CFTR joue un rôle clé dans l'homéostasie ionique et hydrique en régulant notamment les ions chlorures et bicarbonates dans de nombreuses cellules épithéliales. Un défaut de la protéine induit une absence ou dysfonction de la protéine CFTR à la membrane. Or, ce canal, lorsqu'il est actif, induit une inhibition des canaux ENaC. Ces canaux ENaC sont responsables de l'influx d'ions sodium. La levée de l'inhibition des canaux ENaC provoque l'entrée d'ions sodium dans la cellule, ainsi qu'une entrée de l'eau qui l'accompagne. Cette entrée d'eau est donc responsable de la déshydratation du mucus [Riordan et al., 1989, Cant et al., 2014]. Une meilleure compréhension et définition de l'impact du déficit partiel ou total de la protéine a permis un meilleur diagnostic et le développement de médicaments.

1.3.1.2 Manifestations cliniques

Les symptômes les plus fréquents chez les personnes atteintes de mucoviscidose sont des atteintes pulmonaires et digestives. L'altération de la fonction de la protéine CFTR entraîne une modification des sécrétions dans plusieurs types cellulaires de l'organisme. Dans les voies respiratoires, le mucus à la surface des cellules devient plus épais. Ce mucus permet au système respiratoire d'évacuer les germes et les agents infectieux. Lorsqu'il est plus épais, le mucus s'écoule plus difficilement, les bronches peuvent s'encombrer et s'infecter provoquant toux et expectorations. Les voies respiratoires deviennent ainsi un terrain propice aux infections et notamment à la colonisation par la bactérie Pseudomonas Aeruginosa [Davies, 2002]. L'atteinte pulmonaire est la première cause de morbidité et de mortalité chez les patients atteints de mucoviscidose [Martin et al., 2016]. En 2017, aux Etats Unis, 380 patients ont succombé à la mucoviscidose dont 63% pour des raisons respiratoires et cardiorespiratoires [Cystic Fibrosis Foundation, 2017].

Au niveau du tube digestif, l'obstruction des canaux pancréatiques par le mucus épaissi empêche l'arrivée des enzymes pancréatiques dans l'intestin. Ces enzymes sont nécessaires à la digestion des aliments et leur diminution entraîne une réduction de l'absorption des aliments et à terme à des troubles nutritionnels [Gibson-Corley

et al., 2016].Le trypsinogène est une proenzyme synthétisée dans le pancréas qui, lors de son passage dans l'intestin est activée et devient la trypsine. Dans le cas de la mucoviscidose, le faible volume sécrétoire induit une accumulation des trypsinogènes. Cette accumulation couplée à un pH acide induit leur autoactivation. Elle passe alors partiellement dans la circulation sanguine. La quantité anormale de trypsine dans le pancréas initie les premiers stades de l'autodigestion du pancréas par les enzymes pancréatiques qui engendre une pancréatite et peut déclencher une fibrose du pancréas [Ooi and Durie, 2012].

On observe également des atteintes génitales pouvant conduire à une infertilité, surtout chez les hommes. En effet plus de 95% des hommes atteints de mucoviscidose sont stériles en raison d'une azoospermie provoquée par une altération du canal déférent. Dans la plupart des cas, il s'agit d'une absence totale congénitale et donc bilatérale des canaux déférents (CBAVD) qui bloque le transport des spermatozoïdes vers le tractus génital externe [Radpour et al., 2008, Ong et al., 1993]. L'infertilité chez les femmes est moins fréquente mais elles sont généralement hypofertiles. L'infertilité est le plus souvent due à la malnutrition qui peut causer un retard de la puberté, un manque d'ovulation ou à une augmentation de l'épaississement du mucus cervical. On trouve également des cas d'absence congénitale de l'utérus et du vagin (CAUV) chez environ 8% des femmes atteintes de mucoviscidose, deux fois plus que dans la population générale [Timmreck et al., 2003, Radpour et al., 2008].

1.3.1.3 Diagnostic

Avant 1959, le diagnostic de la mucoviscidose se faisait surtout à partir des symptômes évocateurs. C'est à la suite de la canicule de 1948 à New-York, entrainant un état de prostation des malades, que les anomalies électrolytiques ont été découvertes et décrites [Kessler and Andersen, 1951]. L'augmentation du chlore et du sodium dans la sueur explique ainsi la spécificité qu'ont les malades à avoir la « peau salée ». C'est ainsi que l'idée est venue de recueillir et d'analyser la sueur. Cela était difficile et n'a pu être réellement mis en place qu'en 1959 avec l'apparition des premiers tests de la sueur [Gibson and Cooke, 1959]. Ces tests reposent sur la détermination de la concentration en chlorure dans la sueur et sont encore utilisés dans le diagnostic de la mucoviscidose [Farrell et al., 2017].

En France, depuis 2002, le dépistage néonatal de la mucoviscidose par un dosage de la trypsine immuno-réactive (TIR) est devenue systématique. Quelques jours après la naissance et avec le consentement des parents, une goutte de sang est prélevée au nouveau-né pour déterminer la quantité de trypsine dans la circulation sanguine. Si le dosage est positif, on recherche par génétique moléculaire les variants du gène *CFTR*

les plus fréquemment observés chez les personnes atteintes de mucoviscidose. Cela représente une trentaine de variants. La mucoviscidose étant une maladie récessive, deux variants, un sur chaque chromosome, doivent être présents pour développer la maladie. Dans le cas où l'on retrouve sur chaque chromosome un ou plusieurs des variants testés alors le diagnostic est posé. Lorsqu'aucun ou 1 un seul des 30 variants est retrouvé alors le génotype est incomplet. Une exploration exhaustive du gène *CFTR* à la recherche d'un autre variant, moins fréquent, expliquant la maladie est alors réalisée (Figure 1.13) [Centre de Référence Mucoviscidose de Lyon, 2017]. Une absence de variant est identifiée dans 2% des cas. Dans tous les cas un test à la sueur est réalisé afin de conforter les résultats.

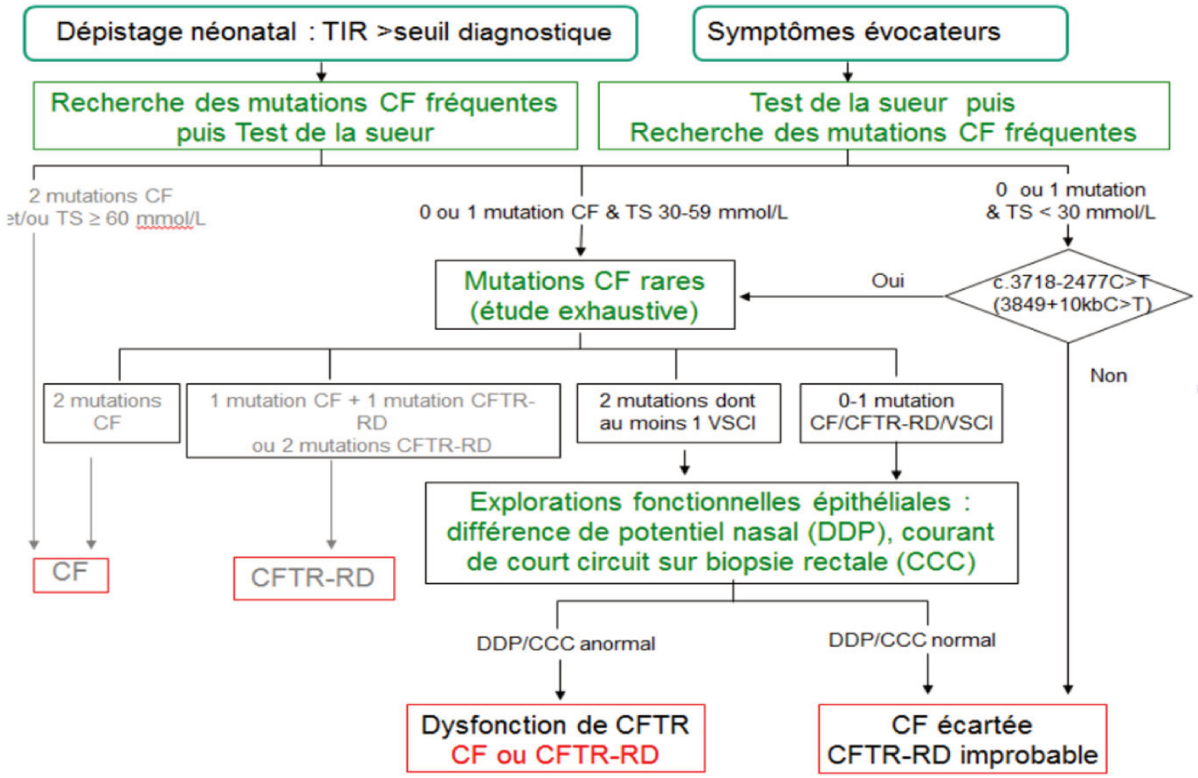


FIGURE 1.13: Arbre décisionnel de dépistage et diagnostic de la mucoviscidose. (Centre de Référence Mucoviscidose de Lyon).

1.3.2 CFTR : du gène à la protéine

Le gène *CFTR* est composé de 27 exons et est localisé sur le bras long du chromosome 7. Il code pour la protéine du même nom. La mucoviscidose est due à des variants sur ce gène qui entraîne un défaut ou une diminution de la quantité et/ou de la fonction de la protéine. Cette protéine de 1480 acides aminés est un membre de la famille des transporteurs à ATP (Adénosine TriPhosphate) Binding Cassette

(ABC). Elle a également la propriété de fonctionner comme un canal chlorure (Cl-) Adénosine MonoPhosphate cyclique (AMPc) - dépendant [Anderson et al., 1991]. Elle est exprimée à la membrane apicale des cellules et contribue à maintenir l'hydratation du mucus et l'homéostasie électrolytique à travers les membranes [Rich et al., 1990].

1.3.2.1 Voie de biosynthèse de la protéine CFTR

Le gène *CFTR* est tout d'abord transcrit en ARNm dans le noyau de la cellule, l'ARNm qui passe ensuite à travers les pores nucléaires. À la suite de l'interaction ARNm/ribosome, l'ARNm est traduit en protéine dans le cytoplasme ou dans le réticulum endoplasmique. Le polypeptide ainsi formé correspond à une forme non glycosylée de 130kDa. La protéine CFTR passe ensuite par une voie de maturation complexe au sein de la bicouche lipidique du réticulum endoplasmique (RE) impliquant des chaperons moléculaires et des co-chaperons [Rich et al., 1990]. Les protéines chaperons et co-chaperons aident au repliement et à la maturation correcte des protéines. Elles interviennent également dans la dégradation des protéines mal repliées [Mayer and Bukau, 2005, Kriegenburg et al., 2012, Hanzén et al., 2016].

La protéine CFTR de type sauvage (Wt-CFTR) est partiellement glycosylée par l'ajout d'oligosaccharides, conduisant à une forme immature d'environ 145 KDa [Gregory et al., 1990]. Les différents domaines s'assemblent par des interactions inter-domaines. La maturation et la glycosylation se poursuivent à travers l'appareil de Golgi, à l'origine de la forme mature complexe d'environ 170 kDa [Helenius and Aebi, 2001, Patrick et al., 2011]. La protéine mature rejoint ensuite la membrane apicale des cellules où elle joue son rôle de canal chlorure. Les protéines mal repliées sont reconnues par le système de surveillance et de dégradation du RE (Endoplasmic-reticulum-associated protein degradation (ERAD)). Elles sont ainsi ubiquitinées et dégradées par le protéasome [Okiyoneda et al., 2010, Rogan et al., 2011].

1.3.2.2 Structure et régulation de la protéine CFTR

La protéine CFTR comme les autres membres de la famille des ABC comprend deux domaines transmembranaires appelés membrane spanning domain (MSD), chacun composé de six hélices, qui délimitent un pore transmembranaire. Elle possède également deux domaines intracellulaires de liaison aux nucléotides (Nucleotide Binding Domain : NBD) capables de fixer l'ATP et interagissant avec les quatre boucles intracellulaires (ICL1 à ICL4) des domaines membranaires (Figure 1.14). La protéine CFTR présente la particularité de posséder une région régulatrice (R), qui la distingue des autres transporteurs ABC.

L'ouverture du canal est ainsi induite par deux évènements. Tout d'abord, le domaine régulateur est phosphorylé par la protéine kinase (PKA) régulée par l'AMPc.

La phosphorylation du domaine R permet ensuite aux domaines NBD de fixer l'ATP entraînant des changements de conformation et l'ouverture du canal. L'hydrolyse de l'ATP, médiée par le site de liaison de l'ATP canonique (impliquant les motifs Walker A et Walker B du NBD2), favorise le retour à la conformation de repos fermé (Figure 1.15) [Ostedgaard et al., 2000, Basso et al., 2003, Mornon et al., 2009, Mornon et al., 2015, Fay et al., 2018]. L'étude des structures 3D de l'assemblage multidomaines de CFTR dans différentes conformations, tout d'abord par voie de modélisation [Serohijos et al., 2008, Mornon et al., 2008, Mornon et al., 2015], ensuite grâce à la cryo-microscopie électronique [Zhang and Chen, 2016, Liu et al., 2017, Zhang et al., 2017, Fay et al., 2018, Zhang et al., 2018] a permis de mieux comprendre les bases moléculaires de l'activité de canal ionique ainsi que l'impact de mutations (Figure 1.16).

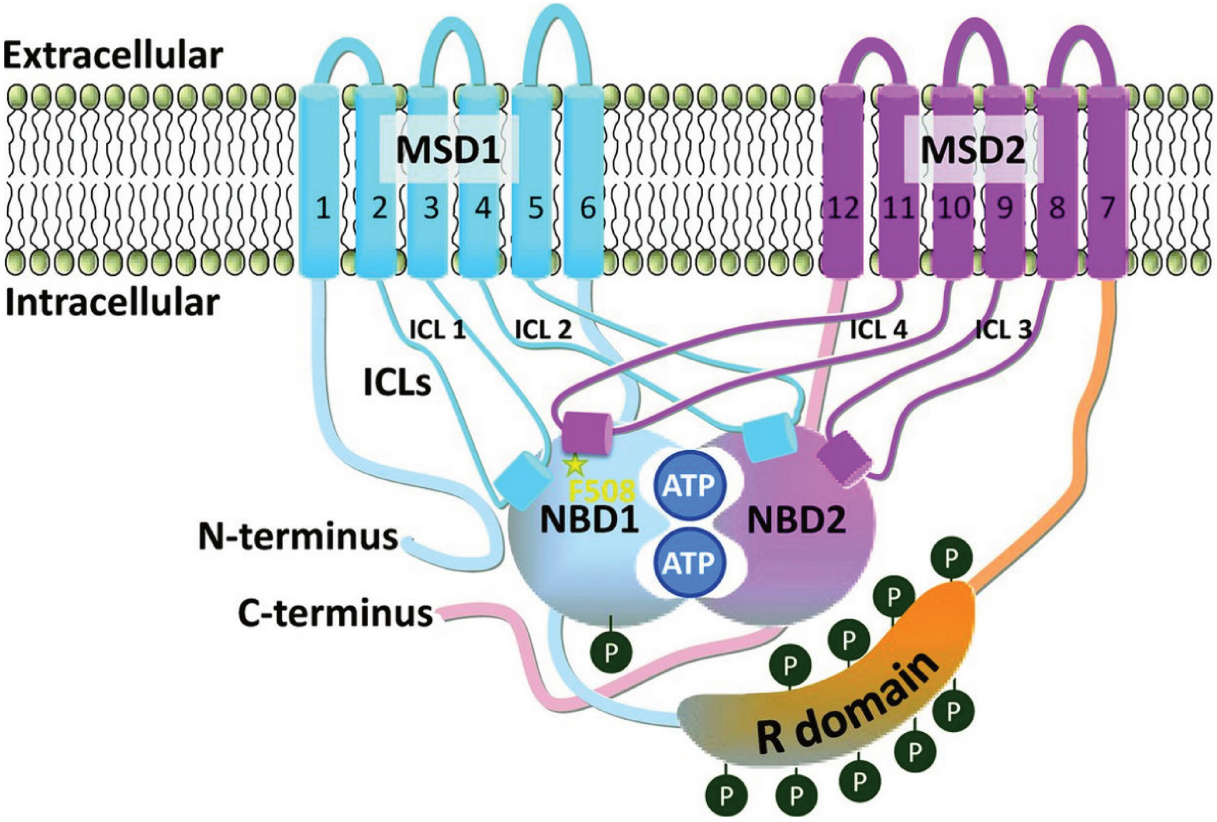


FIGURE 1.14: Schéma représentatif de la protéine CFTR montrant l'organisation de ses domaines. La protéine CFTR possède deux domaines transmembranaires (MSD1 et MSD2), deux domaines de liaisons aux nucléotides (NBD1 et NBD2) et le domaine de Régulation (R). Les phosphorylations permettant la régulation du canal sont indiquées en vert. Les MSD entrent en contact avec les NBD par l'intermédiaire des ICLs. Ces zones de contact sont représentées sous forme de cylindres. La Phénylalanine en position 508 est représentée en jaune (d'après Chiaw et al.).

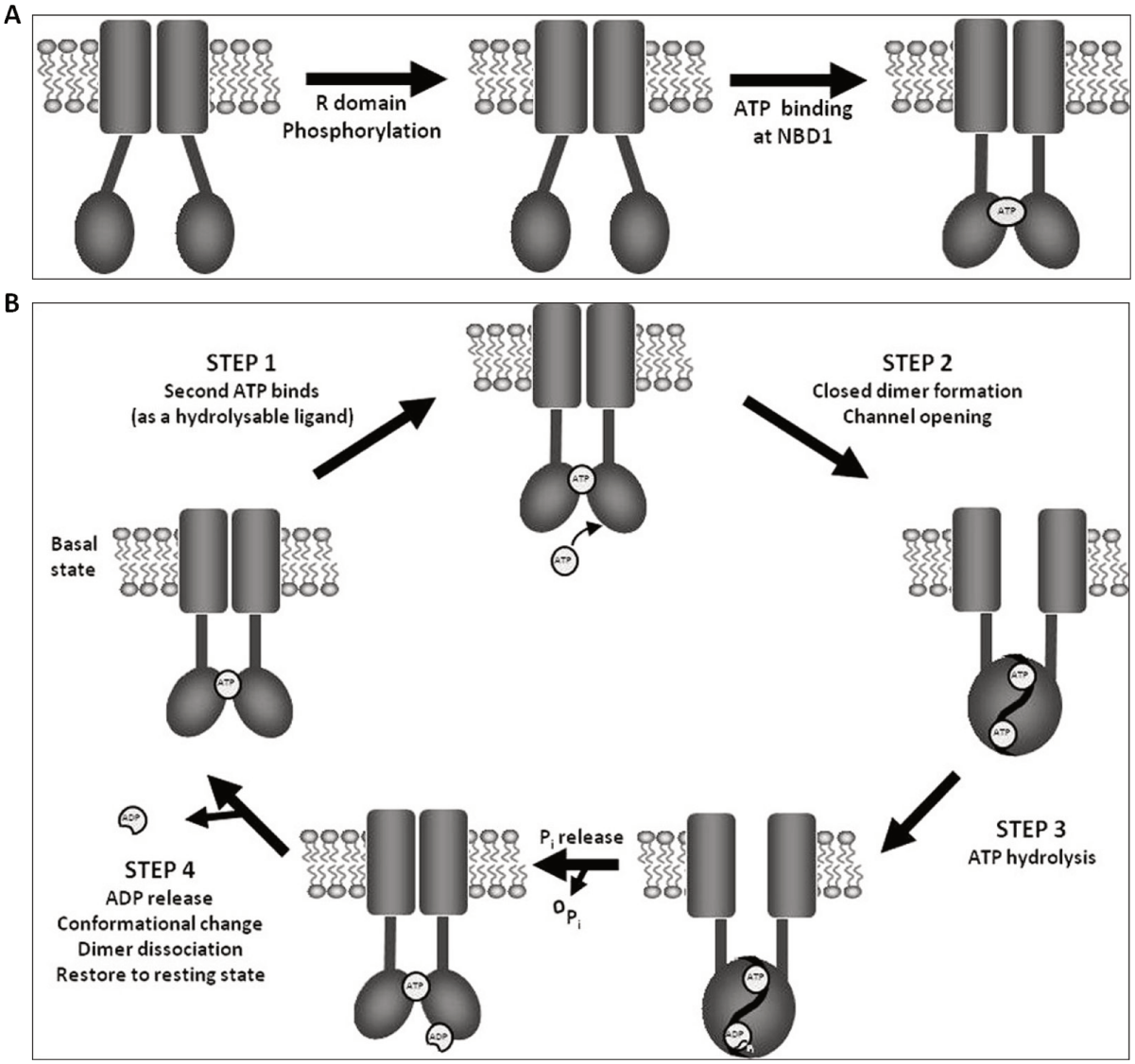


FIGURE 1.15: Régulation de l'activation de la protéine CFTR. A. La phosphorylation du domaine R permet la liaison de l'ATP sur le site NBD1. B. Suite à la liaison de la première molécule d'ATP, une deuxième molécule d'ATP se lie au site catalytique du NBD2 (Etape 1). Les NBD forme un dimère qui entraîne des changements de conformation des TMDs et permet l'ouverture du canal (Etape 2). L'hydrolyse de l'ATP et la dissociation du dimère déclenchent la fermeture du canal (Étape 3). La libération du phosphate (Pi) et adénosine di-phosphate (ADP) restaure le canal à sa conformation basale (Etape 4) (d'après Gout, 2012).

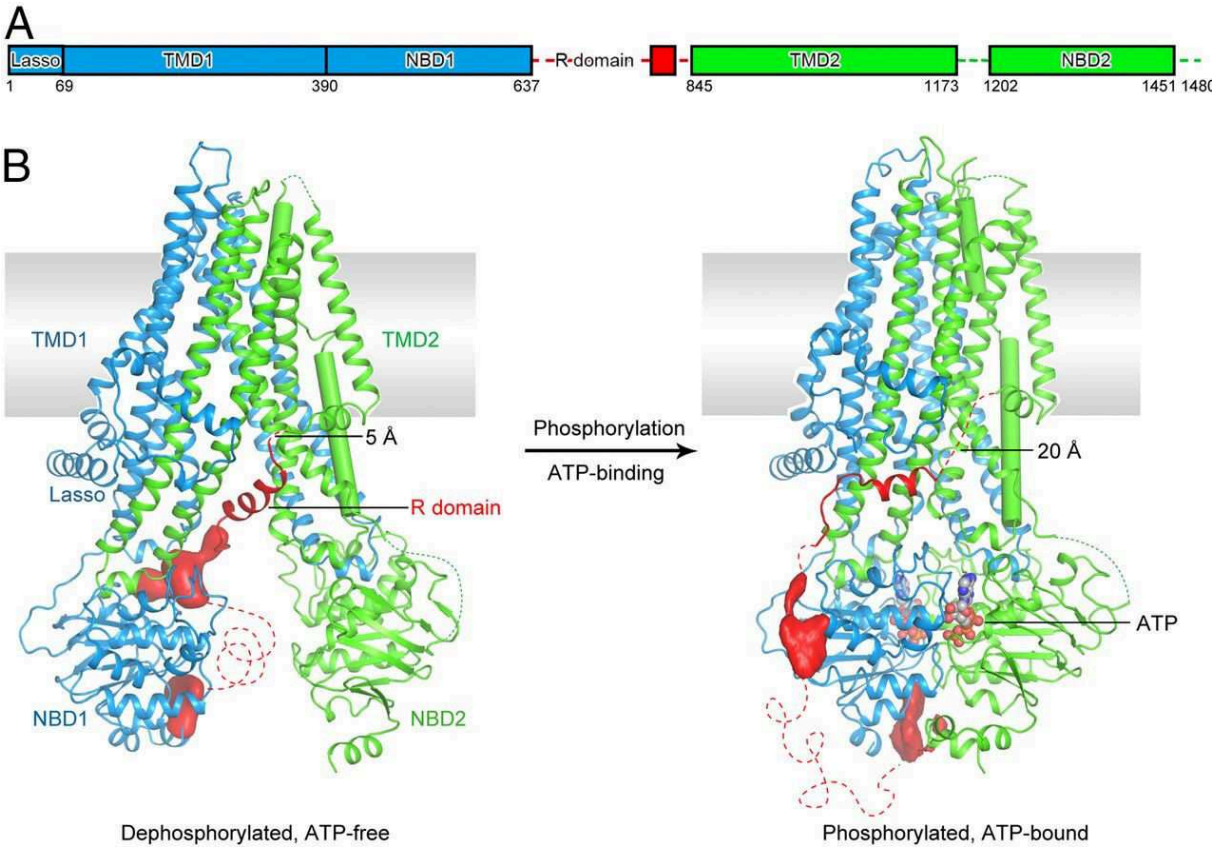


FIGURE 1.16: Deux états conformationnels du CFTR humain. (A) Schéma montrant les domaines de la protéine CFTR. Les chiffres représentent la gamme de résidus visibles sur la carte cryo-EM, et non les limites exactes des différents domaines. (B) Conformation de la structure déphosphorylée, sans ATP (à gauche ; PDB 5UAK) et de la structure phosphorylée, liée ATP (à droite). Les régions non résolues dans la structure sont affichées sous forme de lignes pointillées (d'après Zhang, 2018).

1.3.3 Impact des variants et thérapeutiques

On répertorie actuellement plus de 2000 variants sur le gène *CFTR*. Parmi eux, le variant le plus fréquent est la délétion d'une phénylalanine en position 508 (p.Phe508del) qui est présente chez 85% des malades dont 45% sous forme homozygote [Cystic Fibrosis Foundation, 2017]. Comme nous l'avons vu précédemment les variants peuvent avoir un impact sur la protéine à différents niveaux. Cela peut subvenir avant même sa traduction.

1.3.3.1 Les classes de variants

Ainsi on a classé les variants retrouvés dans le gène *CFTR* dans 6 classes [Welsh and Smith, 1993, Zielenski and Tsui, 1995]. Ce système de classification regroupe les variants en fonction des problèmes qu'ils entraînent sur la protéine CFTR. L'intérêt d'avoir ce type de classification est de pouvoir adopter la même stratégie thérapeutique pour chaque classe. Il existe en effet aujourd'hui plusieurs traitements utilisés chez les patients dans le but de corriger les défauts du CFTR.

- Les variants de classe I sont des variants qui entrainent l'absence ou la modification de la synthèse de la protéine. Ce sont principalement des variants non-sens, des délétions/insertions et des variants modifiant l'épissage. Il en résulte soit une protéine tronquée et instable soit, en amont de la traduction, un ARNm qui est dégradé par le processus de dégradation NMD. Une absence de protéine à la membrane est alors constatée.
- Les variants de classe II sont des variants qui modifient le repliement normal de la protéine et altèrent son trafic intracellulaire. Ces protéines étant mal formées, elles sont retenues par le système de surveillance du RE puis rapidement dégradées. Seule une petite quantité de protéine partiellement fonctionnelle échappe au système de vérification du RE et est transportée vers la membrane. On observe donc une très forte diminution de la quantité de protéine à la membrane. Le variant p.Phe508del est un variant de classe II [Serohijos et al., 2008, Lukacs and Verkman, 2012].
- Les variants de classe III et IV sont des variants qui impactent le canal CFTR lorsqu'il est à la membrane plasmique. Les variants de classe III modifient l'ouverture du canal et les variants de classe IV modifient sa conductance. Le deuxième variant le plus fréquent en population fait partie de la classe III. C'est un variant faux sens qui change une glycine en une asparagine en position 551 (p.G551D). Elle est présente sur au moins un allèle chez 4-5% des patients [Ramsey et al., 2011]. La protéine mutée p.G551D est présente à la membrane

mais son activité est très réduite [Bompadre et al., 2007, Bompadre et al., 2008].

- Les variants de classe V affectent la transcription du *CFTR*. Ce sont souvent des variants qui affectent l'épissage. Les variants de classe VI réduisent la stabilité de la protéine à la surface des cellules.

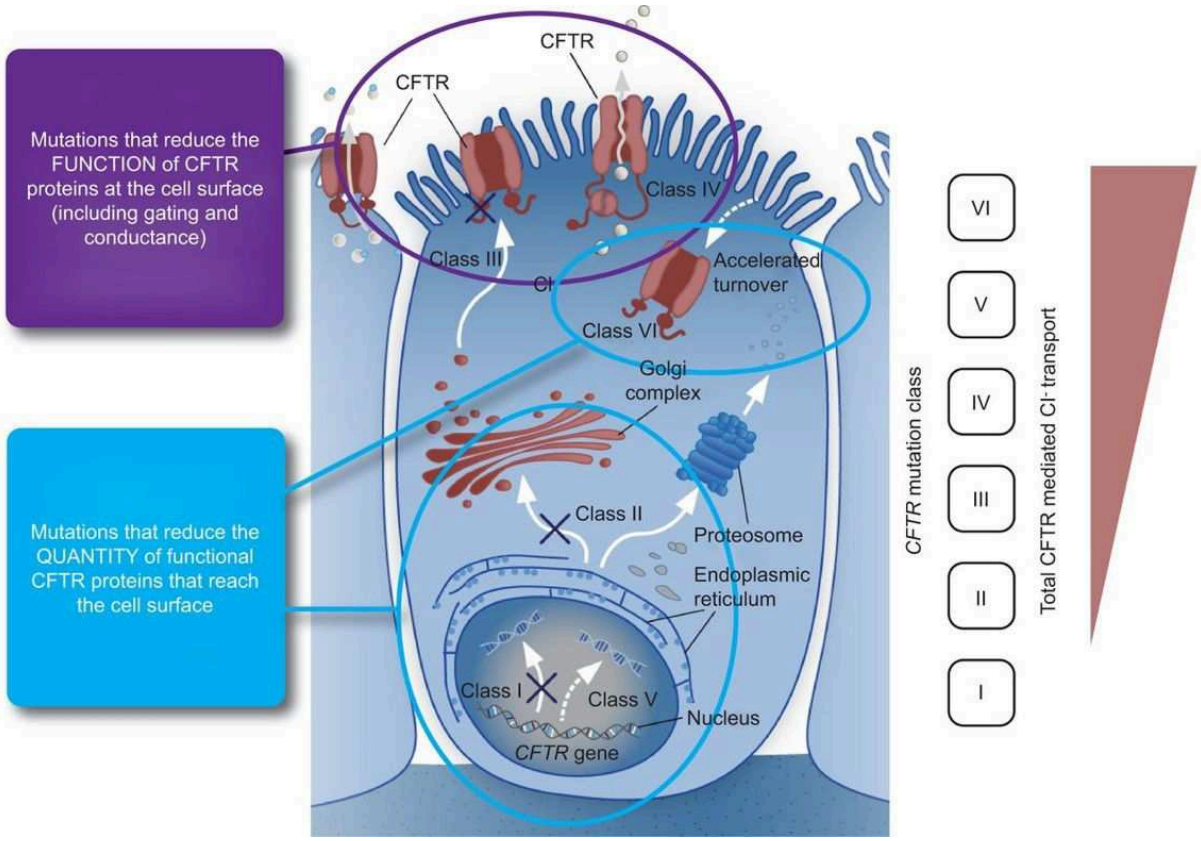


FIGURE 1.17: Biosynthèse de la protéine CFTR sauvage et classes de variant (d'après Derichs, 2013).

Les variants de classe I, II, V et VI entraînent donc une absence ou une réduction de la quantité de protéine CFTR au niveau de la membrane cellulaire, alors que les variants de classe III et IV modifient la fonction ou l'activité du CFTR au niveau de la membrane cellulaire. Ces différents variants engendrent des dysfonctionnements dont la sévérité est variable. Les variants de classe I, II et III donnent des formes relativement sévères de mucoviscidose [Rogan et al., 2011, Fraser-Pitt and O'Neil, 2015] alors que les variants de classe IV, V et VI sont associés à des phénotypes plus modérés.

On observe également des différences de sévérité entre deux patients porteurs des mêmes variants (Mekus et al., 2000 ; Lucarelli et al., 2012). On peut expliquer ces différences phénotypiques de plusieurs manières. La première est l'environnement et la qualité de vie des patients [Collaco et al., 2016]. La deuxième est le contexte génétique, avec la présence possible de gènes modificateurs qui vont avoir un impact

sur l'expression des variants mais également la possibilité que des allèles situés en *cis* puissent jouer un rôle (contexte haplotypique) [Kerem et al., 1990, Hubert et al., 1996, Genin et al., 2008].

Il existe en effet quelques exemples d'implication de combinaisons de variants qui ont des conséquence sur la protéine CFTR lorsqu'ils sont présents sur le même haplotype (en *cis*) [Fanen et al., 1999, Rohlfs et al., 2002, de Prada Merino et al., 2010, Lucarelli et al., 2010, El-Seedy et al., 2012, Diana et al., 2016]. Ainsi, des études in vitro ont montré que deux variants R347H et D979A ont des effets modérés lorsqu'ils sont seuls mais un effet délétère lorsqu'ils sont présents en *cis* (Figure 1.18) [Clain et al., 2001]. Ces deux variants agiraient donc de concert pour donner un phénotype sévère. Cet exemple illustre bien l'importance de la prise en compte du contexte haplotypique dans l'étude de l'effet des variants du gène *CFTR* chez les patients.

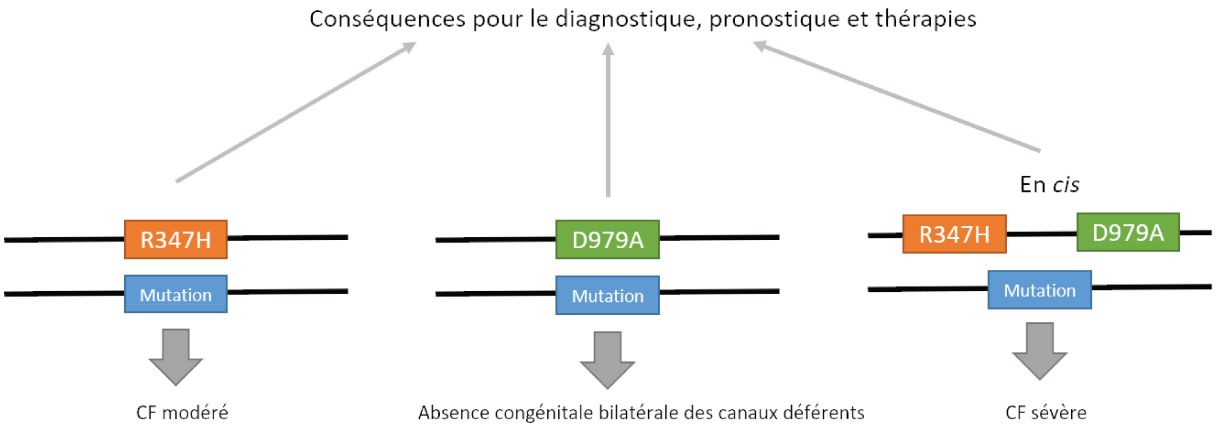


FIGURE 1.18: Exemple de deux variants dans le gène CFTR.

1.3.3.2 Impact de la p.Phe508del sur la protéine CFTR

La délétion de la phénylalanine en position 508 de la protéine CFTR résulte de la délétion de trois nucléotides (c.1521_1523delCTT) dans l'exon 10 qui ne modifie pas le cadre de lecture. La position 508 se trouve dans le domaine cytoplasmique NBD1. La p.Phe508del est un variant qui entraîne un défaut de repliement de la protéine mais n'empêche pas la liaison à l'ATP [Riordan, 2005]. Une mauvaise conformation de la protéine a pour conséquence une glycosylation partielle de la protéine qui est caractérisée par un poids moléculaire de 145 kDa. Lors de sa synthèse et son insertion dans la membrane du RE, l'assemblage des différents domaines du CFTR ne s'effectue pas correctement. En effet, la perturbation ou suppression des sites de contact de la Phénylalanine en position 508 provoque des défauts de repliement [Lewis et al., 2005, Du et al., 2005]. Etonnamment l'altération de la conformation NBD1 est

modérée mais entraine un défaut de stabilité thermique. L'assemblage inter-domaine est lui aussi affecté et il existe une instabilité de l'interaction existant entre NBD1 et la boucle intracellulaire ICL4 du domaine MSD2 (Figure 1.19) [Serohijos et al., 2008, Mornon et al., 2008, Mendoza et al., 2012].

Deux étapes clés du repliement de la protéine CFTR, sous sa forme native, sont donc modifiées en absence de la phenylalanine en position 508 et induit une protéine CFTR bien moins fonctionnelle. En effet des expériences ont montré que malgré un bon repliement du NBD1, la déstabilisation de l'interaction NBD1/ICL4 ne permet pas d'obtenir une structure native fonctionnelle. [Mendoza et al., 2012].

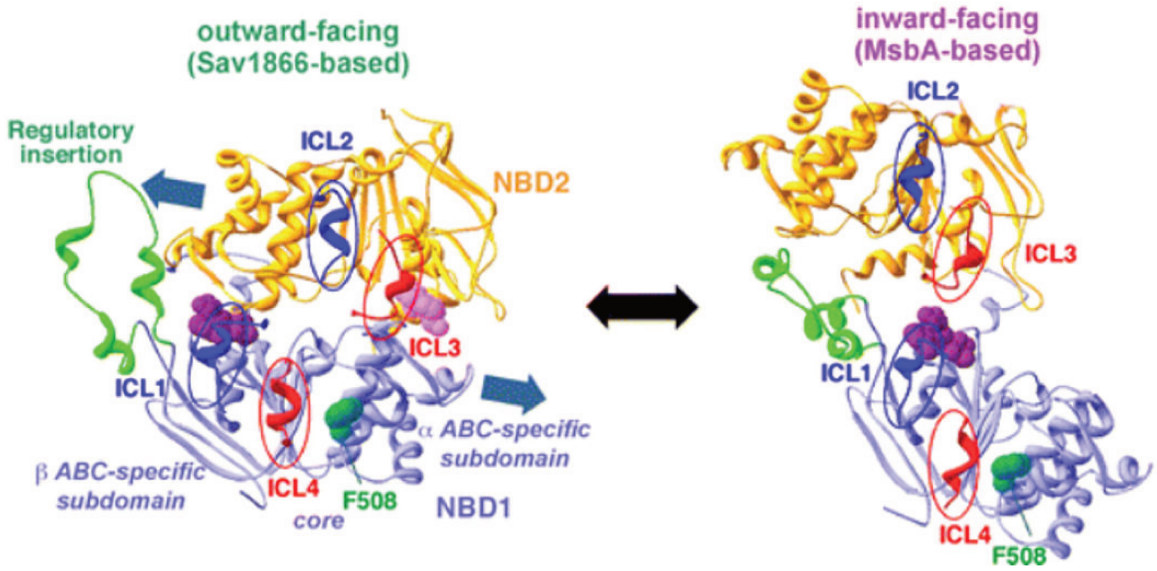


FIGURE 1.19: Assemblage des domaine NBD1 et NBD2 canal ouvert (outward-facing) et canal fermé (inward-facing) vu de la membrane. L'hétérodimère NBD1/NBD2 est représenté respectivement en bleu et en jaune, ainsi que les hélices interagissant avec les boucles intracellulaires des MSDs (ICL1-ICL4). L'ATP se liant au domaine NBD1 est représenté en violet et celui du domaine NBD2 en rose. La phénylalanine en position 508 est représentée en vert sous forme sphérique (d'après Mornon et al., 2009).

L'altération globale de l'assemblage des différents domaines en présence du variant p.Phe508del conduit à un mauvais repliement de la protéine CFTR qui est alors reconnue par les protéines chaperons. La protéine chaperon Hsc70 (Heat shock cognate 70) a même une interaction plus forte avec la protéine mutée qu'avec la protéine non-mutée (Meacham, 2002 ; Scott-Ward and Amaral, 2009). Cette protéine chaperon serait d'ailleurs en partie responsable de la dégradation du CFTR par le système ERAD [Chanoux and Rubenstein, 2012]. Certaines protéines parviennent cependant à échapper à la dégradation et à atteindre la membrane plasmique. Elle présente cependant une probabilité d'ouverture (Po) et un temps de présence dans la membrane fortement réduits [Dalemans et al., 1991, Haws et al., 1996, Mendoza et al., 2012].

1.3.3.3 Thérapeutiques

Les traitements de la mucoviscidose qu'ils soient symptomatiques ou curatifs ont significativement amélioré la qualité de vie des malades et augmenté considérablement leur espérance de vie. Alors que les enfants décédaient le plus souvent avant l'âge de 5 ans en 1965, l'espérance de vie est aujourd'hui d'environ 44 ans aux Etats-Unis [Cystic Fibrosis Foundation, 2017]. Les tout premiers traitements de la mucoviscidose ont été les antibiotiques comme la pénicilline pour les infections pulmonaires et les enzymes pancréatiques pour les problèmes digestifs [Shwachman, 1960, Norman, 1964, Matthews et al., 1964]. Pour l'atteinte pulmonaire, on diffusait du propylène glycol sous forme de brume sous une tente sous laquelle on faisait dormir les patients afin de fluidifier les sécrétions muqueuses. On réalisait également de la physiothérapie respiratoire [Matthews et al., 1964]. Ces traitements sont encore utilisés mis à part la diffusion sous la tente qui imposait aux patients de dormir « mouillés jusqu'aux os ». Ces traitements permettent de réduire les symptômes et non d'en corriger les causes. Après la découverte du gène *CFTR*, de nombreux travaux ont été réalisés pour mieux comprendre le fonctionnement du gène et de la protéine CFTR et de ses partenaires et de trouver des cibles thérapeutiques (Figure 1.20) [Castellani and Assael, 2017].

Il existe aujourd'hui plusieurs traitements utilisés chez les patients dans le but de corriger les défauts du CFTR (Table 1.1). Trois traitements mis au point par Vertex Pharmaceuticals ont passé les essais cliniques et sont maintenant prescrits aux patients : ivacaftor (Kalydeco®), lumacaftor/ivacaftor (Orkambi®), et tezacaftor/ivacaftor (Symdeko®). Kalydeco® a obtenu l'autorisation de mise sur le marché pour les enfants de 2 ans et plus, pesant moins de 25 kg, porteurs de variants de classe III du gène *CFTR* [Haute Autorité de Santé, 2012]. L'Orkambi® est prescrit aux patients âgés de 12 ans et plus homozygotes pour le variant p.Phe508del du gène *CFTR* [Haute Autorité de Santé, 2019]. Symdeko® a quant à lui obtenu les autorisations aux Etats Unis et au Canada mais pas encore en France pour être prescrit aux patients âgés de 12 ans et plus ayant les deux copies p.Phe508del ou aux patients présentant une p.Phe508del avec certains variants.

D'autres molécules ont également été testées comme la molécule PTC124 testée en essai clinique chez les patients atteints de mucoviscidose sous le nom de Translarna®. Il s'agit d'un médicament prescrit chez les patients atteints de dystrophie musculaire de Duchenne ayant un variant dans le gène de la dystrophine entraînant un codon stop. La molécule agit sur le ribosome pour qu'il lise le codon stop et ne produise pas de protéine tronquée, ce phénomène est appelé translecture du codon stop (Figure 1.20). Les effets chez les patients atteints de dystrophie musculaire de Duchenne n'ont

pour le moment pas été très concluants mais la molécule reste prescrite [McDonald et al., 2017]. En revanche l'essai clinique pour les patients atteints de mucoviscidose ayant un variant stop s'est arrêté faute d'effet visible d'amélioration chez les patients [Kerem et al., 2014].

Dans la suite du mémoire on utilisera les codes médicaments VX-770 pour l'ivacaftor et VX-809 pour lumacaftor. Le VX-770 est un potentiateur c'est-à-dire qu'il agit sur l'ouverture du canal à la membrane. Il est utilisé pour les variants de classe III et IV. Le VX-809 est un correcteur, il permet au CFTR mal replié lors de sa biogenèse de pouvoir atteindre la membrane plasmique en corrigeant son défaut de maturation et de transport (Figure 1.20). Bien que le VX-809 soit un correcteur prometteur, il ne corrige principalement que l'interface NBD1-MSD1/2. Il a un donc un avantage clinique limité de part la nécessité de la stabilisation du domaine NBD1 pour un repliement et un assemblage correcte de la protéine [Okiyoneda et al., 2013, He et al., 2015]. On observe aussi des différences de réponses aux traitements d'un patient à l'autre ayant pourtant le même génotype. La phase II de l'essai clinique pour le médicaments Orkambi, par exemple, a montré une grande hétérogénéité des réponses des patients homozygotes pour la p.Phe508del [Boyle et al., 2014]. Les patients, sélectionnés pour cet essai clinique, ont été soumis à la recherche génétique d'un panel de 33 variants du gène *CFTR*. La recherche de variants retrouvés chez les patients n'est donc pas exhaustive. On ne peut donc pas éliminer la possibilité que des variants présents en *cis* avec la p.Phe508del puissent modifier la réponse de ces patients au traitement.

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Défaut	Absence de protéine	Pas de trafic	Pas de fonction	Moins de fonction	Moins de protéine	Défaut de stabilité
Mutation	Gly542X	Phe508del Asp979Ala	Gly551Asp Arg347His	Arg117His	Ala455Glu	Gln1412X
Approche thérapeutique	Permettre la synthèse de la protéine	Corriger le repliement de la protéine	Restaurer la conductance du canal	Restaurer la conductance du canal	Correction de l'épissage	Améliorer la stabilité
Cible moléculaire	Ribosome	NBD1 et ses intéractions	Ouverture du pore	Conductance		
Traitement en essaie clinique ou déjà prescrit	Translarna®	Orkambi® Symdeko®	Kalydeco®			

TABLE 1.1: Classes de variants, impacts et thérapeutiques

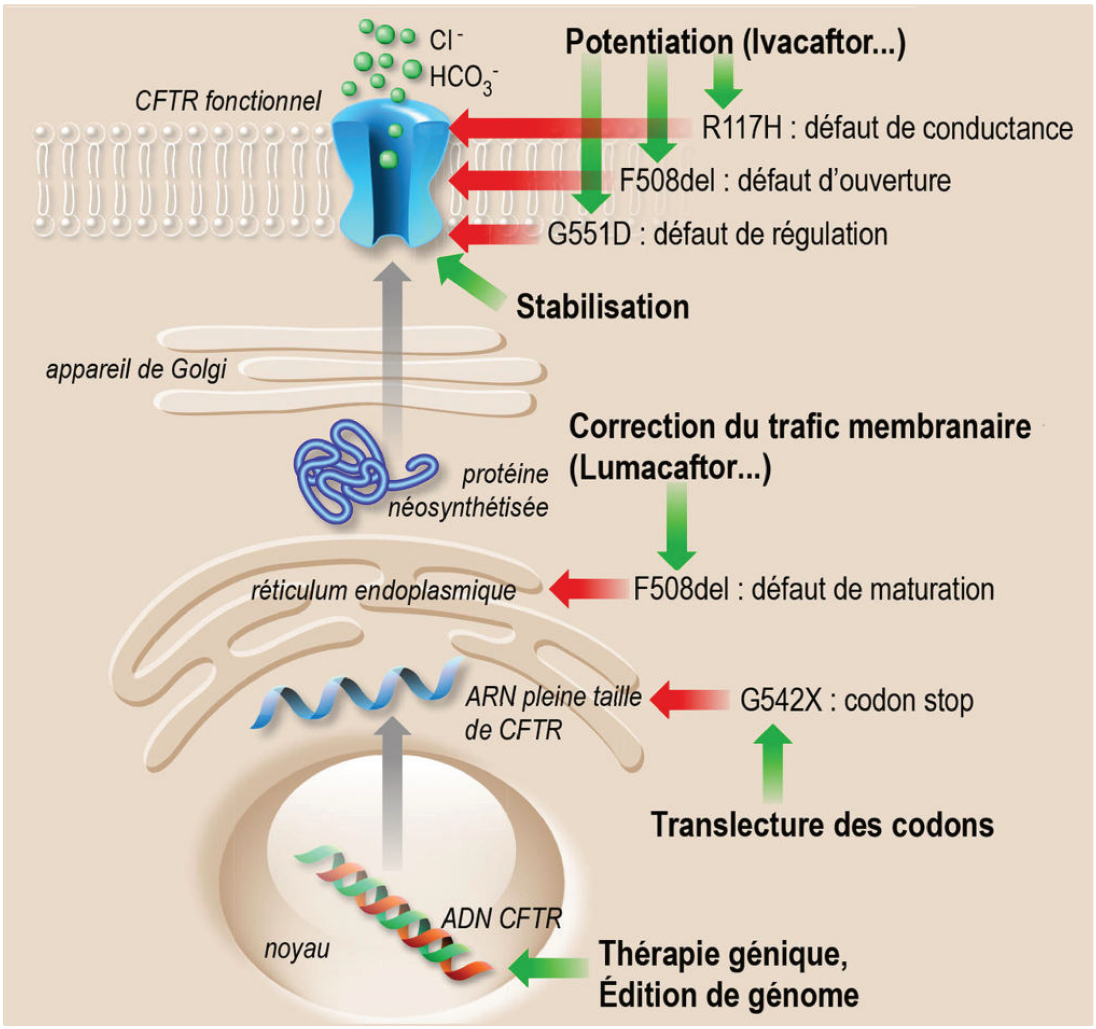


FIGURE 1.20: Les différents modulateurs de la protéine CFTR selon le type de variant dont elle est l'objet. (Adapté de F. Aguila)

1.4 Objectif de la thèse

A l'interface entre la bioinformatique et la biologie moléculaire, mon sujet de thèse vise à interpréter les conséquences de la variabilité du génome sur la fonction des protéines. Comme nous l'avons vu, le séquençage du génome d'un individu révèle des millions de variants génétiques par rapport à la séquence de référence humaine (1.1.2.1) [1000 Genomes Project Consortium et al., 2015, Fu et al., 2013]. La quantité de variants étant trop importante pour être traitée efficacement, l'application de filtres bioinformatiques permet de sélectionner le(s) variant(s) d'intérêt(s) au regard du problème étudié [Richards et al., 2015, MacArthur et al., 2014].

Pour filtrer les variants, on utilise des informations fournies par des annotateurs qui vont prédire l'impact des variants sur la protéine. Ces prédictions sont faites pour chaque variant pris isolément et sans tenir compte des autres variants associés en *cis* dans le même gène chez un individu donné. Pourtant, comme nous l'avons vu,

ce contexte haplotypique peut se révéler très important pour prédire l'expression phénotypique avec la possibilité que des variants neutres quand pris isolément deviennent fortement délétères lorsqu'ils sont associés en *cis* (1.3.3.1).

Pour étudier ces combinaisons de variants, j'ai développé mon projet de thèse autour de deux objectifs (1) la conception d'un programme bioinformatique pour visualiser les combinaisons de variants présents dans un gène et leur impact sur la séquence protéique et (2) l'étude de l'impact fonctionnel sur la protéine de combinaisons de variants du gène *CFTR* trouvées en population.

CHAPITRE 2

DÉVELOPPEMENT D'UN OUTIL BIOINFORMATIQUE

Sommaire

Lorsqu'on analyse des données de séquençage, on considère trop souvent les variants présents dans un gène un à un sans s'intéresser aux combinaisons de variants présentes chez l'individu. De telles combinaisons pourraient cependant expliquer en partie certaines maladies et des variants, qui n'ont pas d'impact, peuvent lorsqu'ils sont réunis chez un individu devenir délétères. Pour étudier cette hypothèse, j'ai développé un outil bioinformatique de visualisation des variations de la séquence protéique induites par les variations génétiques présentes chez l'individu. Cet outil baptisé GEMPROT permet la visualisation de ces combinaisons de variants qui peuvent toucher un même domaine fonctionnel de la protéine.

2.1 Matériels et méthodes

2.1.1 Panels de données

Le projet 1000 Genomes a été mis en place dans le but de réaliser un catalogue de variants communs dans la population générale. Le projet a débuté en 2008 et a pris fin en 2015. Il a abouti à la mise à disposition sur des bases de données publiques des données de séquences de 2504 individus provenant de 26 populations à travers le monde. La base de données recense 84,7 millions SNVs et 3,6 millions d'insertions/délétions courtes [1000 Genomes Project Consortium et al., 2015].

Parmi les 26 populations du projet 1000 Genomes figurent plusieurs populations européennes mais aucune donnée française. En 2013, notre laboratoire, en collaboration avec plusieurs équipes françaises, a donc décidé de répondre à un appel d'offre de France Génomique pour mettre en place un panel de référence d'exomes français dans le cadre du « French Exome (FrEx) project » [Genin et al., 2017]. Les études GWAS avaient en effet montré que, même au sein de l'Europe, il existait des différences de fréquences alléliques importantes selon les zones géographiques. Ces différences seraient vraisemblablement encore plus importantes pour des variants rares, apparus récemment en population et qui n'avaient pas eu le temps de se diffuser sur de grandes échelles géographiques [Heath et al., 2008, Novembre et al., 2008]. Il apparaissait donc nécessaire pour étudier les données de séquençage de patients d'origine française de pouvoir disposer de données sur la variabilité génétique de la population française. Ainsi, un séquençage des exomes a été réalisé sur 573 individus originaires de 6 régions françaises (Nord Pas de Calais, Normandie, Bretagne (Finistère), Pays de la Loire, Aquitaine, Bourgogne). Le séquençage a été réalisé au Centre National de Recherche en Génomique Humaine et financé par France Génomique (`http://lysine.univ-brest.fr/FrExAC/`).

2.1.2 Reconstruction des haplotypes

Le projet 1000 Genomes met à disposition des données déjà phasées. Pour construire des haplotypes de haute qualité intégrant plusieurs types de variants, les chercheurs ont mis en place une approche en plusieurs étapes. Ils ont tout d'abord reconstruit les haplotypes des données de puces de SNVs. Lorsqu'elles étaient disponibles, ils ont utilisé les données des parents pour améliorer le phasage. Les sites variants provenant du séquençage ont été identifiés à l'aide d'une combinaison d'outils bioinformatiques pour définir un ensemble de variants bi-alléliques de confiance, comprenant à la fois des SNVs et des indels. Pour les autres variants, ils ont été phasés en exploitant les informations de déséquilibre de liaison. Par autres variants on entend

(1) les SNVs multi-alléliques, c'est-à-dire les sites où dans la population plusieurs allèles alternatifs sont présents, (2) les indels et des variants complexes par exemple les très grandes délétions ou insertions. Les haplotypes bi-alléliques et multi-alléliques ont été fusionnés en une seule représentation d'haplotype. Ils ont conjointement utilisé les outils Beagle et SHAPEIT2 afin d'obtenir un phasage complet de l'ensemble des données disponibles [Browning and Browning, 2007, Delaneau et al., 2012].

Pour les données de FrEx, le pipeline Alzheimer Disease Exome Sequencing-France (ADES-FR) a été appliqué aux données de séquençage d'exomes [Bellenguez et al., 2017]. Seuls les variants bi-alléliques, comprenant à la fois des SNP et des indels, avec une haute qualité, ont été retenus pour le phasage. Le phasage de ces variants a été réalisé avec SHAPEIT2. C'est l'outil de phasage le plus utilisé et le plus efficace en termes de fiabilité et de rapidité. Il se base sur la théorie de coalescence. Cette théorie est basée sur l'idée que l'on possède tous un ancêtre commun et que l'on peut suivre l'évolution d'un gène au sein d'une population. Les outils utilisant cette théorie recherche donc les régions IBD. SHAPEIT2 utilise une approche d'échantillonnage de Gibbs dans laquelle chaque haplotype individuel a été optimisé en fonction des estimations des haplotypes des autres échantillons. En plus des autres outils de phasage SHAPEIT à optimisé de temps de calcul par l'utilisation d'arbres binaires pour représenter les ensembles d'haplotypes possibles pour chaque individu. Ainsi il n'explore pas l'intégralité des haplotypes possible, représentant 2(n-1) avec n le nombre de SNV, mais seulement les plus plausibles. SHAPEIT2 préconise d'utiliser un panel de référence lorsque l'échantillon comprend moins de 100 individus. Le phasage avec un panel de référence nécessite d'avoir dans son fichier VCF uniquement les positions présentes dans le panel de référence et donc d'éliminer tous les variants qui sont dans les données qu'on analyse mais pas dans le panel de référence ce qui en pratique conduit à l'élimination des variants les plus rares. Le panel de FrEX étant composé de 573 individus, j'ai choisi de réaliser le phasage sans données externes pour ne pas perdre les variants non présents dans le panel de référence. J'ai également utilisé les informations présentes sur les lectures de séquençage (Fichier BAM) [Delaneau et al., 2013].

2.1.3 Outils et bases de données utilisés par GEMPROT

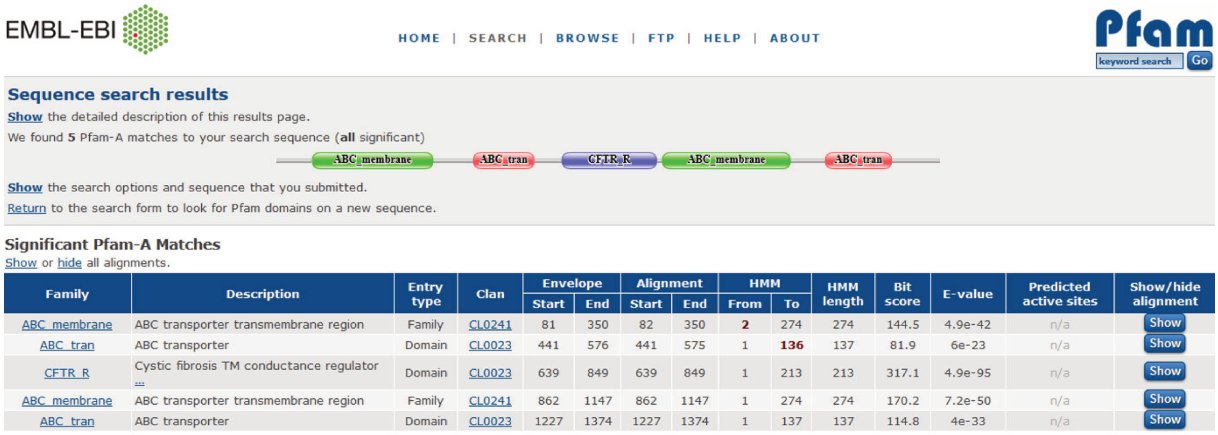
Le pipeline de GEMPROT est composé d'une suite d'outils bioinformatiques et se sert de plusieurs bases de données pour permettre à l'utilisateur une meilleure appréhension des combinaisons de variants.

- Le projet Consensus Coding Sequence (CCDS) est un projet qui a regroupé et validé les séquences codantes des gènes de l'homme et de la souris pour les différents transcrits. Ce projet met à disposition un fichier avec, pour chaque

transcrit, la position génomique de début et de fin des exons. Chaque transcrit de gène annoté est associé à un identifiant CCDS. L'ensemble des transcrits répertoriés commence avec un ATG initiateur et un codon stop qui permet la terminaison de la traduction [Pruitt et al., 2009].

- VCFtools est un programme qui permet de lire, d'analyser et de manipuler des fichiers au format VCF. Le programme est divisé en deux modules. Le premier est utilisé pour effectuer diverses opérations sur les fichiers VCF. On peut par exemple sélectionner un variant en fonction de sa position ou de son génotype (hétérozygote ou homozygote), fusionner et comparer des fichiers VCF. Le second module est un exécutable principalement utilisé pour analyser les données, permettant à l'utilisateur d'estimer les fréquences alléliques, les niveaux de déséquilibre de liaison et diverses métriques de contrôle de la qualité [Danecek et al., 2011].
- SAMtools est un programme qui permet de lire, d'analyser et de manipuler des fichiers au format SAM ou leur version binaire BAM. Il est capable de convertir d'autres formats d'alignement, de trier et de fusionner les alignements, d'enlever les doublons de PCR, et d'afficher des alignements sous forme de texte (conversion BAM>SAM). SAMtools propose également des options qui permettent d'indexer les fichiers de séquences au format FASTA ou d'extraire une sous-séquence de la séquence de référence indexée [Li and Durbin, 2009].
- Pfam est une base de données et un outil en ligne qui répertorie une vaste collection de familles de protéines, chacune représentée par de multiples alignements de séquences et des modèles de Markov cachés [Finn et al., 2016]. Ainsi lorsque l'on présente une séquence nucléotidique ou protéique, l'outil va faire une recherche de la séquence afin d'identifier des domaines répertoriés dans la base de données Pfam (Figure 2.1). En effet, les protéines sont généralement composées d'une ou de plusieurs domaines fonctionnels, communément appelées domaines. Ces domaines présentent une signature nucléotidique propre, c'est-à-dire une suite de nucléotide qui varie assez peu en fonction de la protéine qui la porte. L'identification des domaines présents dans les protéines peut donc fournir des informations sur leur fonction [Finn et al., 2016, El-Gebali et al., 2019].
- ClinVar est une base de données qui fournit les interprétations cliniques qui relient un variant à une ou plusieurs maladies. Elle fait le lien entre les variations présentes dans le génome et la santé humaine. Les interprétations sont basées sur les résultats obtenus par des analyses fonctionnelles ou publiés dans la littérature (Figure 2.2). ClinVar agrège les résultats obtenus pour une combinaison variation/phénotype et indique lorsque les interprétations cliniques

sont contradictoires [Landrum et al., 2016, Landrum et al., 2018].



EMBL-EBI HOME | SEARCH | BROWSE | FTP | HELP | ABOUT Pfam keyword search Go

Sequence search results

Show the detailed description of this results page.

We found 5 Pfam-A matches to your search sequence (all significant)

ABC_membrane ABC_tran CFTR_R ABC_membrane ABC_tran

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

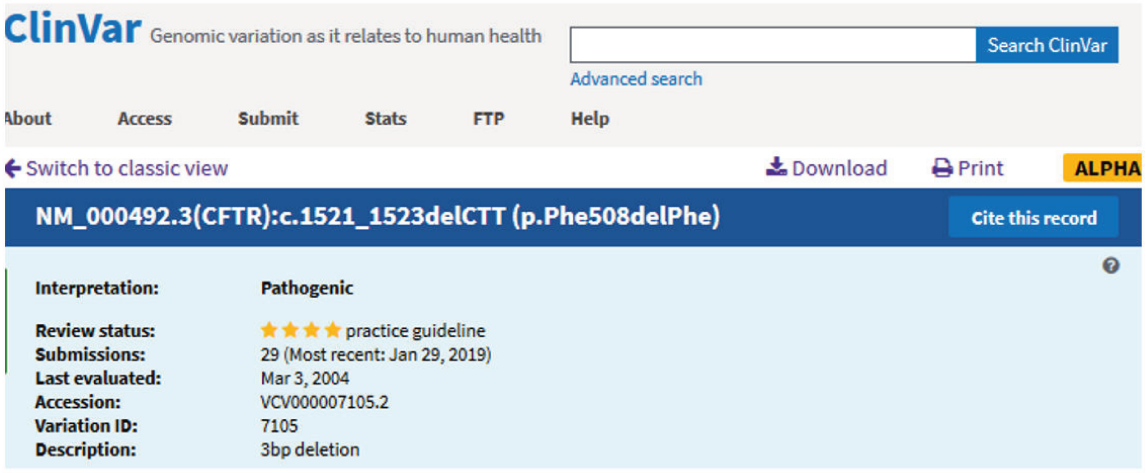
Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
ABC_membrane	ABC transporter transmembrane region	Family	CL0241	81	350	82	350	2	274	274	144.5	4.9e-42	n/a	Show
ABC_tran	ABC transporter	Domain	CL0023	441	576	441	575	1	136	137	81.9	6e-23	n/a	Show
CFTR_R	Cystic fibrosis TM conductance regulator ...	Domain	CL0023	639	849	639	849	1	213	213	317.1	4.9e-95	n/a	Show
ABC_membrane	ABC transporter transmembrane region	Family	CL0241	862	1147	862	1147	1	274	274	170.2	7.2e-50	n/a	Show
ABC_tran	ABC transporter	Domain	CL0023	1227	1374	1227	1374	1	137	137	114.8	4e-33	n/a	Show

FIGURE 2.1: Résultat Pfam avec la séquence de la protéine CFTR. A partir de la séquence de la protéine CFTR, la base donnée Pfam a identifié cinq domaines fonctionnels. Quatre domaines correspondent à des domaines spécifiques de la famille des ABC dont deux domaines transmembranaires. Le dernier domaine est le domaine régulateur spécifique du CFTR.

A



B

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
Pathogenic (Jul 30, 2018)	criteria provided, single submitter (GeneDX Variant Classification(06012015)) Method: clinical testing	Not Provided	GeneDx Accession: SCV000329246.6 Submitted: (Jan 29, 2019)	Evidence details Comment: The c.1521_1523delCTT pathogenic variant in the (more...)
Pathogenic (Mar 30, 2016)	criteria provided, single submitter (ACMG Guidelines, 2015) Method: clinical testing	None	Knight Diagnostic Laboratories,Oregon Health and Sciences University Study: CSER-NextGen Accession: SCV000538017.1 Submitted: (Aug 10, 2016)	Evidence details Comment: The c.1521_1523delCTT (p.Phe508del), also known (more...)
Pathogenic (Apr 15, 2015)	criteria provided, single submitter (Counsyl Autosomal and X-linked Recessive Disease Classification criteria (2015)) Method: clinical testing	None	Counsyl Accession: SCV000678175.1 Submitted: (Jun 22, 2017)	Evidence details Publications PubMed (1)

FIGURE 2.2: Information de la base de donnée clinvar via l'interface graphique pour le variant p.Phe508del. A. Information relative au variant. L'interprétation clinique de la délétion est "pathogenic". B. Quelques références qui ont conduit à l'interprétation clinique.

2.2 GEMPROT : GEnetic Mutation to PRotein Translation

2.2.1 Proposition et spécification de la méthode d'analyse GEMPROT

La méthode proposée permet de visualiser l'impact des variants génétiques présents sur chacun des deux haplotypes d'un individu, sur la séquence en acides aminés de la protéine. Elle permet également d'annoter les variants et de les positionner au sein des domaines fonctionnels connus de la protéine. Elle utilise un fichier VCF phasé : un fichier contenant l'ensemble des variants portés par l'individu sur l'un et l'autre de ses haplotypes. GEMPROT fonctionne sous deux modes différents, un mode individuel qui permet de visualiser la séquence protéique de chaque individu et un mode population qui donne les différentes combinaisons d'haplotypes observées et leur fréquence.

2.2.2 Développement et implémentation de GEMPROT

2.2.2.1 Algorithme

GEMPROT nécessite (1) un fichier VCF phasé, contenant tous les variants portés par un (ou plusieurs individus) sur ses deux haplotypes, (2) une liste d'individus avec le nom de leur population d'appartenance ou de groupe comme cas/témoin et (3) le nom du gène d'intérêt. (Figure 2.3.A).

La première étape de l'outil permet une lecture plus rapide du fichier VCF. J'ai réduit la liste des variants à ceux présents dans l'intervalle « début-fin » du transcrit du gène à l'aide de VCFtools. L'étape suivante permet d'extraire les séquences de nucléotides. Tout d'abord, les positions des exons du transcrit du gène sont extraites de la base de données CCDS [Pruitt et al., 2009]. Les séquences nucléotidiques des transcrits sont extraites du génome de référence à l'aide de l'outil SAMtools [Li and Durbin, 2009]. La séquence de référence est alors dupliquée pour représenter les deux haplotypes et chaque haplotype est modifié pour tenir compte de la ou des mutations présentes (Figure 2.3.B).

La deuxième étape est la traduction. Les deux haplotypes sont traduits à l'aide d'un script Perl qui lit l'intégralité de la séquence nucléotidique en tenant compte de toutes les mutations. Ces variations sont placées sur les différents domaines fonctionnels connus de la protéine qui sont reconstruits à l'aide de la base de données Pfam [Finn et al., 2016]. Une option de GEMPROT permet d'introduire ses propres informations sur la position des domaines sans faire appel à Pfam. En effet, comme Pfam annote les domaines en fonction d'une signature, il est possible que les bornes des domaines

ne soient pas exactes par rapport à ce qui est connu de la protéine. GEMPROT annote les acides aminés modifiés avec leurs propriétés physicochimiques et les informations présentes dans la base de données Clinvar (Figure 2.3.C) [Landrum et al., 2016, Landrum et al., 2018].

2.2.2.2 Les sorties du programme

Lorsque le fichier VCF en entrée contient des données sur plusieurs individus, GEMPROT répertorie chaque individu et ses deux haplotypes. Il donne également un résumé par haplotype avec leurs fréquences. En cliquant sur l'identifiant d'un individu, l'utilisateur peut obtenir les résultats pour l'individu concerné (Figure 2.4).

GEMPROT fournit un rapport HTML avec un résumé des mutations présentes pour chaque individu et une visualisation des variations de séquence de protéines au sein des domaines fonctionnels connus. Sur cette page, on dispose de l'information sur le nom de l'individu, le nom du gène avec son identifiant CCDS. On donne également les informations sur le gène et la protéine non mutée (nombre d'acides aminés, nombre d'exons et liste de ses domaines). Il est possible de cliquer sur le nom d'un domaine pour accéder au site Web de Pfam et de la même manière, on peut obtenir les informations issues de Clinvar si la position variable est présente dans cette base de données. Pour chaque haplotype, la liste des mutations génétiques et leur impact sur la séquence protéique est indiquée. Les mutations faux-sens sont annotées avec les propriétés physicochimiques des acides aminés de référence et des acides aminés mutés. En bas de page on retrouve la visualisation des variations le long des deux séquences protéiques codées dans les domaines fonctionnels connus de la protéine (Figure 2.5).

À l'échelle d'une population, les résultats sont donnés sous la forme d'un tableau qui permet la comparaison des distributions d'haplotypes entre les différents sous-groupes d'individus de la population. Dans l'exemple Figure 2.6, nous pouvons visualiser le résumé des résultats obtenus pour le gène *GJB2* avec le mode de population des cinq sous-populations du projet 1000 Genomes.

2.2.2.3 Mise à disposition de l'outil

GEMPROT est disponible sur github en tant qu'application en ligne de commande (Figure 2.7). De cette manière il est possible de travailler sur de gros jeux de données, sur plusieurs individus et sur plusieurs gènes.

Nous avons également développé une version Web de GEMPROT qui permet d'exécuter le programme sur de petits ensembles de données dans un environnement plus convivial (`http://lysine.univ-brest.fr/GEMPROT/`) (Figure 2.8).

2.2.3 Article : GEMPROT : visualization of the impact on the protein of the genetic variants found on each haplotype

Bioinformatics, 2018, 1–3
doi: 10.1093/bioinformatics/bty993
Advance Access Publication Date: 3 December 2018
Applications Note

Genome analysis

GEMPROT: visualization of the impact on the protein of the genetic variants found on each haplotype

Tania Cuppens*, Thomas E. Ludwig, Pascal Trouvé and Emmanuelle Genin*

UMR1078 'Génétique, Génomique Fonctionnelle et Biotechnologies', INSERM, Univ Brest, EFS, IBSAM, CHU, Brest F-29200, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 30, 2018; revised on November 9, 2018; editorial decision on November 26, 2018; accepted on November 30, 2018

Abstract

Summary: When analyzing sequence data, genetic variants are considered one by one, taking no account of whether or not they are found in the same individual. However, variant combinations might be key players in some diseases as variants that are neutral on their own can become deleterious when associated together. GEMPROT is a new analysis tool that allows, from a phased vcf file, to visualize the consequences of the genetic variants on the protein. At the level of an individual, the program shows the variants on each of the two protein sequences and the Pfam functional protein domains. When data on several individuals are available, GEMPROT lists the haplotypes found in the sample and can compare the haplotype distributions between different sub-groups of individuals. By offering a global visualization of the gene with the genetic variants present, GEMPROT makes it possible to better understand the impact of combinations of genetic variants on the protein sequence.

Availability and implementation: GEMPROT is freely available at https://github.com/TaniaCuppens/GEMPROT. An on-line version is also available at http://med-laennec.univ-brest.fr/GEMPROT/.

Contact: tania.cuppens@inserm.fr or emmanuelle.genin@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

By providing a complete view of all the genetic variants found in the genome or its coding part, the exome, of individuals, next-generation sequencing has opened up some new possibilities in the study of both rare and common diseases (Majewski and Rosenblatt, 2012; Saint Pierre and Génin, 2014). In a typical individual, whole-exome and whole-genome sequencing reveal, respectively, hundreds of thousands to a few millions of genetic variants compared to the human reference sequence (1000 Genomes Project Consortium *et al.*, 2015; Lek *et al.*, 2016). It is, therefore, necessary to filter these variants by following standards and guidelines for prioritizing candidate genetic variants. The commonly used filter criteria are (i) population frequencies, (ii) localization of variants in the genome and (iii) effect prediction from tools (Mahmood *et al.*, 2017). These filters are used on each variant individually without taking into account the haplotypic context in which each variant is found. This could have serious implications as it can lead to the exclusion of true causative variants. There are some examples of the involvement of combinations of individually neutral variants that, when present on the same haplotype (in cis), can play a role in disease and lead to diagnosis errors (Iossa *et al.*, 2010; Uguen *et al.*, 2017). Considering the different genetic variants present in cis with a known disease mutation might also help interpret phenotypic variability and evidence modifier genes (Génin *et al.*, 2008; Monari *et al.*, 1994). In order to

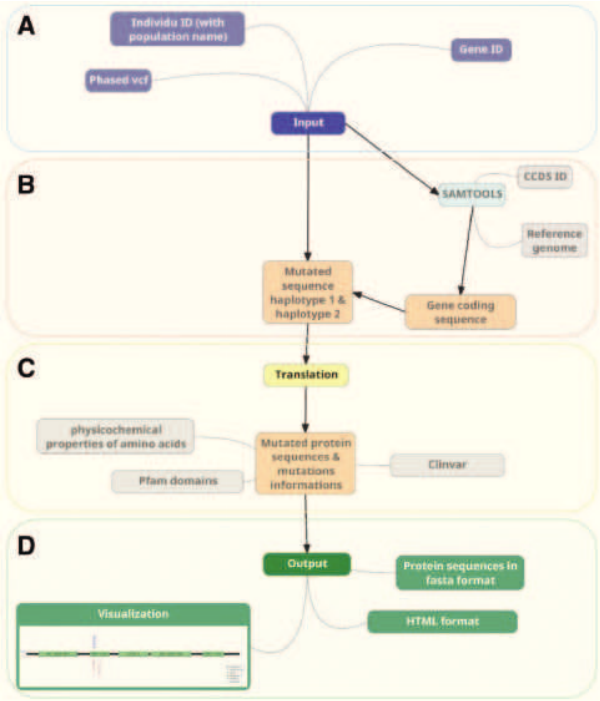


Fig. 1. Workflow of GEMPROT. (**A**) Input file. Three input files are necessary for analysis: Gene ID, phased vcf and sample file with sample ID (and the name of the population for each sample). (**B**) Nucleotide sequence extraction. The CCDS file lists the positions of each exon of the transcript and SAMtools extracts the sequences of the reference genome. (**C**) Translation. The sequences are then translated and modified amino acids are associated with their physicochemical properties, their Pfam domains and Clinvar information. (**D**) Output. An HTLM output is provided as well as a text file

evidence variant combinations that could play a role in disease susceptibility or phenotype variability, the first step is to provide tools to visualize them. This requires a change in the way variants are recorded to consider them globally at the scale of the individual and not at each genomic position one by one.

We propose GEMPROT a new analysis tool to visualize, for each individual, the impact on the encoded protein amino-acids of the genetic variants he/she carries in cis in the coding gene sequence. The modified amino acids are associated with their physicochemical properties and the functional domain in which they are located to provide a better interpretation of the expected changes at the individual level than possible when reading vcf-files line by line.

2 Materials and methods

GEMPROT needs as input a file in the phased VCF format (variant calling format) listing the variants present in one or more individuals and the phased genotypes of each individual. The user also provides the name of his/her gene of interest (Fig. 1).

2.1 Transcript choice and nucleotide sequences extraction

The positions of the coding regions of the reference genome are obtained from NCBI data, based on the Consensus Coding Sequence Project (Pruitt *et al.*, 2009). The nucleotide sequence is then extracted from the reference genome using SAMtools (Li *et al.*, 2009). This step enables to retrieve the nucleotide sequences of the different transcripts from the gene of interest. The idea is to work only on variants that have a potential impact on the protein sequence. The reference sequence of the gene is then duplicated to represent the two haplotypes and modified at the positions where variants were found on each haplotype of each individual.

2.2 Translation

The two haplotypes of each individual are translated using a dedicated *perl* script. It allows the identification of variants that modify the amino acid sequence and shows the impact of all the variants present in the nucleotide sequence globally. This will avoid misinterpreting the effect of two variants within the same codon as two amino-acid changes.

2.3 Visualization

At the level of an individual, the program allows the visualization of the variations on each of the two encoded protein sequences. The physicochemical properties of the modified amino acids and the functional domains of the protein obtained from Pfam (Finn *et al.*, 2016) are shown (Supplementary Fig. S2). When data on several individuals are available, GEMPROT gives the haplotypes found in the sample and their respective frequency (Supplementary Fig. S1). GEMPROT also has a mode intended for larger populations to show the haplotypes distribution across sub-groups of individuals (Supplementary Fig. S3). The output is provided in HTML format. A user-friendly web application is also available and known variants are annotated based on the ClinVar database.

3 Discussion

By offering a global visualization of the putative encoded protein with amino-acid changes highlighted, GEMPROT will allow a better understanding of the impact of combinations of genetic variants on the protein sequence. Assumptions need to be made regarding the transcript but, to our knowledge, this is the first tool to provide such a protein-centric view of the genetic variants found in each individual starting from a phased VCF. With the development of long reads sequencing, we believed phasing will become more reliable and tools such as GEMPROT very useful in the analysis of sequencing data.

Funding

This work was supported by the Regional Council of Brittany and Inserm (Ph.D. fellowship to Tania Cuppens).

Conflict of Interest: none declared.

References

1000 Genomes Project Consortium *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Finn,R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.

Génin,E. *et al.* (2008) Identifying modifier genes of monogenic disease: strategies and difficulties. *Hum. Genet.*, **124**, 357–368.

Iossa,S. *et al.* (2010) R75Q dominant mutation in GJB2 gene silenced by the in cis recessive mutation c.35delG. *Am. J. Med. Genet. Part A*, **152 A**, 2658–2660.

Lek,M. *et al.* (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Mahmood,K. *et al.* (2017) Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum. Genomics*, **11**, 10.

Majewski,J. and Rosenblatt,D.S. (2012) Exome and whole-genome sequencing for gene discovery: the future is now! *Hum. Mutat.*, **33**, 591–592.

Monari,L. *et al.* (1994) Fatal familial insomnia and familial Creutzfeldt-Jakob disease: different prion proteins determined by a DNA polymorphism. *Proc. Natl. Acad. Sci. USA*, **91**, 2839–2842.

Saint Pierre,A. and Génin,E. (2014) How important are rare variants in common disease? *Brief. Funct. Genomics*, **13**, 353–361.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Uguen,K. *et al.* (2017) Diagnostic value of targeted next-generation sequencing in suspected hemochromatosis patients with a single copy of the HFE p.Cys282Tyr causative allele. *Am. J. Hematol.*, **92**, E664–E666.

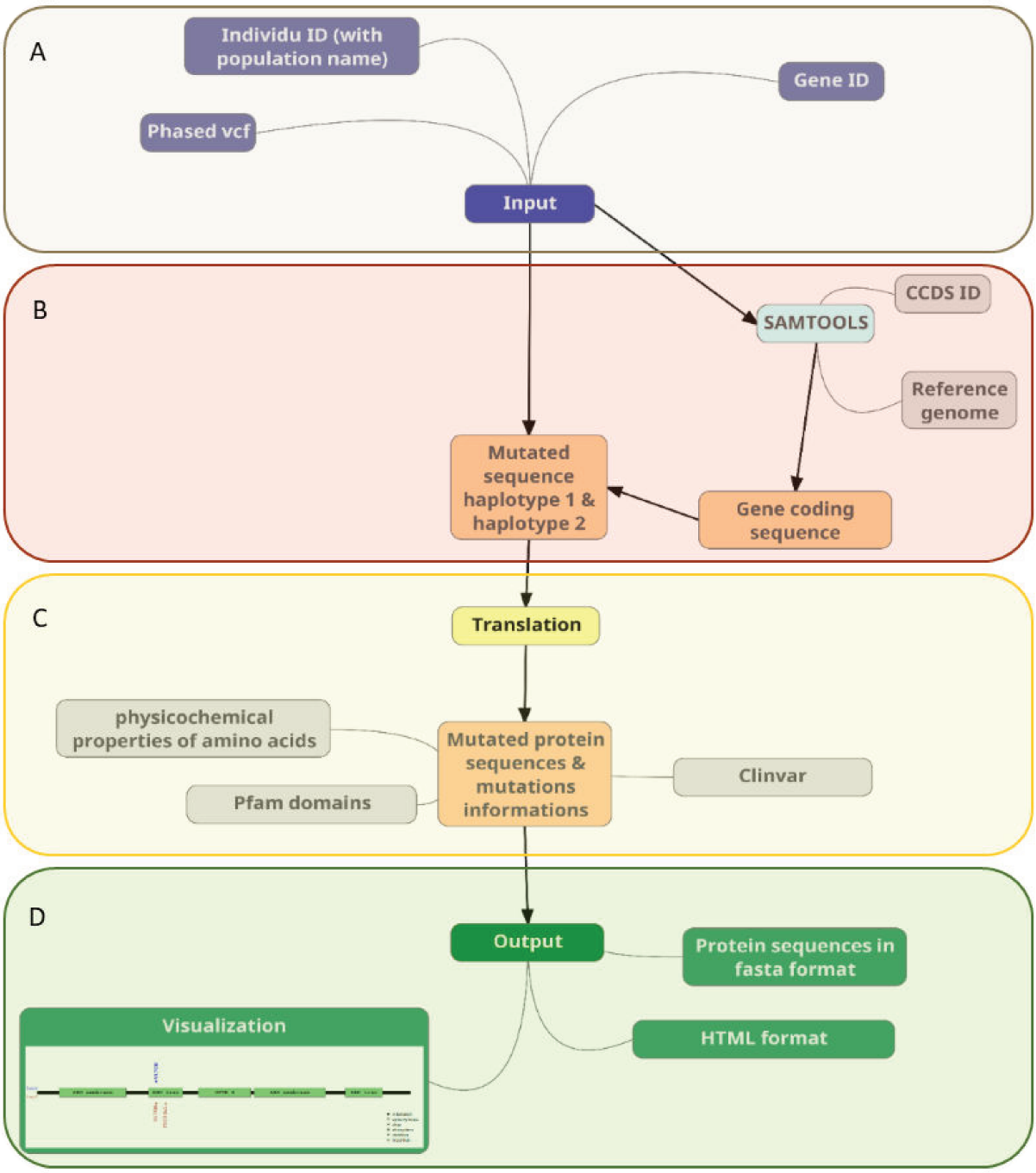


FIGURE 2.3: Pipeline de GEMPROT. (A) Fichiers d'entrée. Trois fichiers sont nécessaires pour l'analyse : un fichier d'identifiants de gènes, un fichier VCF phasé et un fichier d'individus (et le nom de la population pour chaque individu). (B) Extraction des séquences nucléotidiques. Le fichier fourni par le CCDS répertorie les positions de chaque exon de la transcription et SAMtools extrait les séquences du génome de référence. (C) Traduction. Les séquences sont ensuite traduites et les acides aminés modifiés sont annotés par leurs propriétés physico-chimiques et par les informations présentes dans la base de données Clinvar. Les domaines reconnues par Pfam, des séquences traduites, sont stockés pour la visualisation. (D) Sortie. Une sortie HTLM est fournie.

Change Transcript

Results summary

Summary by sample(s)

Sample	Haplotype 1	Haplotype 2
HG00096	no_mutations	no_mutations
HG00097	no_mutations	no_mutations
HG00099	no_mutations	no_mutations
HG00100	no_mutations	no_mutations
HG00101	no_mutations	M:34:T
HG00448	I:203:T	V:27:I,E:114:G
HG00449	V:27:I,E:114:G	no_mutations
HG00451	V:37:I	I:203:T
HG00452	V:37:I	no_mutations

Summary by haplotype(s)

Gene	Haplotype	No. of chromosomes	Frequency	Individuals
GJB2_CCDS9290.1	FS1:79:80,L:81:*	8	0.0016	HG00478 HG00613 HG00704 HG01600 HG02140 NA18549 NA18949 NA19063
GJB2_CCDS9290.1	R:127:H,L:214:L	1	0.0002	HG03991
GJB2_CCDS9290.1	E:120:K	1	0.0002	HG02420
GJB2_CCDS9290.1	P:70:A	1	0.0002	NA19028
GJB2_CCDS9290.1	W:24:*	2	0.000399	HG03767 HG03886
GJB2_CCDS9290.1	A:171:T	2	0.000399	HG00232 HG00252

FIGURE 2.4: Résumé des résultats par le mode individuel. Page de résultat au format HTML qui répertorie : chaque individu avec ses haplotypes ; les différents haplotypes trouvés dans l'échantillon étudié et leurs fréquences. Des liens permettent d'accéder aux résultats pour chaque individu en cliquant sur le nom de l'individu concerné.

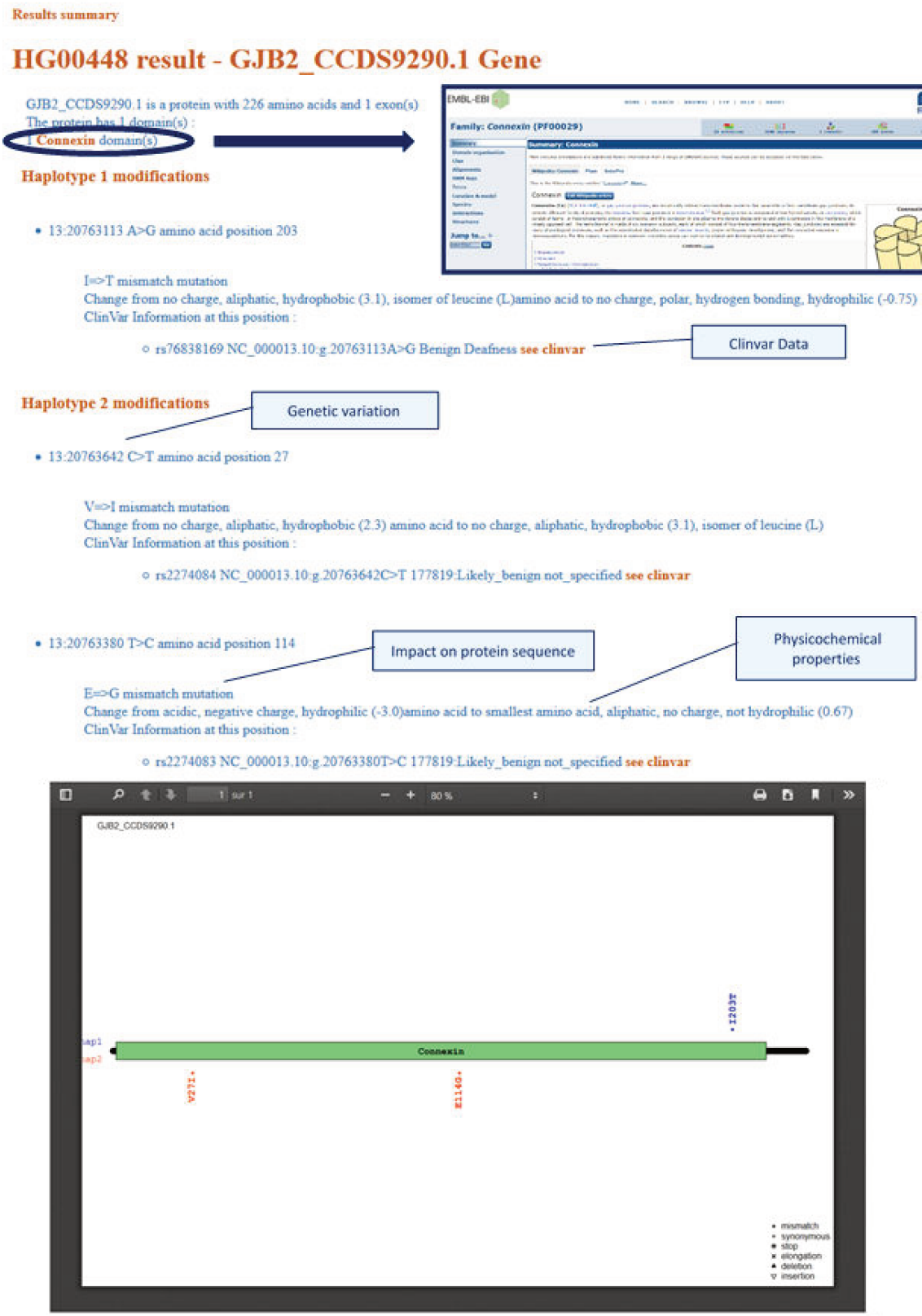


FIGURE 2.5: Résultat pour un individu de 1000G (HG00448). Page de résultat HTML qui répertorie les différents variants trouvés et décrit les deux protéines codées. Un lien vers les bases de données Pfam et Clinvar est également disponible pour les informations de domaines et de variants.

Summary result - GJB2_CCDS9290.1 Gene

GJB2_CCDS9290.1 is a protein with 226 amino acids and 1 exon(s)
The protein has 1 domain(s) :
1 **Connexin** domain(s)

Population Results 2504 sample(s).

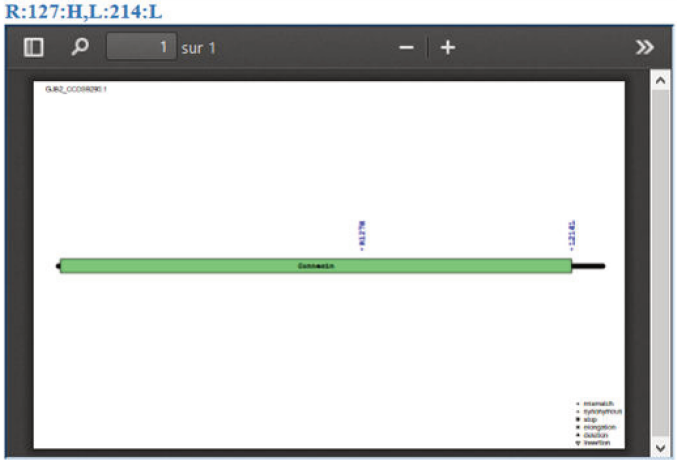
Gene	haplotype	AFR	AMR	EAS	EUR	SAS
GJB2_CCDS9290.1	V:27:I,E:114:G	0	0	148	0	4
GJB2_CCDS9290.1	M:34:L	0	0	1	0	0
GJB2_CCDS9290.1	R:127:H	0	0	1	6	127
GJB2_CCDS9290.1	R:127:H,L:214:L 	0	0	0	0	1
GJB2_CCDS9290.1	V:37:I	1	2	74	0	0
GJB2_CCDS9290.1	A:171:T	0	0	0	2	0
GJB2_CCDS9290.1	E:114:G	0	0	2	0	0
GJB2_CCDS9290.1	V:27:I,T:123:N	0	0	9	0	0

FIGURE 2.6: Résumé des résultats obtenus avec le mode population pour le gène GJB2 sur les cinq sous-populations de 1000 Genomes.

```
2 modes :
      -indiv
   MANDATORY:
              --phased-vcf    : phased vcf
              --output-dir    : specify output directory
              --sample | --sample-file        : sample or sample file (one sample name by line)
              --gene | --gene-file    : official gene symbol or official gene symbol with one ccds in tab delimited
      -pop
   MANDATORY:
              --phased-vcf    : phased vcf
              --output-dir    : specify output directory
              --sample-file   : sample name with his location ; tab delimited file (ex: HG00096       EUR)
              --gene | --gene-file    : official gene symbol or official gene symbol with one ccds in tab delimited


   OPTIONAL:
          --loc-file      : location file
          --fasta : if you want protein fasta file for each haplotype and reference
          --synonymous    : if you want to see synonymous SNP
          --domain : if you know protein domain
          -h      : show this message and quit
```

FIGURE 2.7: Liste des commandes et options de GEMPROT.

Documentation and Download available on GitHub

GEMPROT Web Application

Introduction

GEMPROT (Genetic Mutation to Protein Translation) is a tool to visualize the changes induced on the protein by the different variants found within a gene in an individual. Starting from a phased vcf, it translates the two haplotypes of each individual into the two corresponding protein sequences, allowing to vizualize if variants are in cis or trans. GEMPROT proposes two different modes: the first mode outputs both sequences for each individuals and provides a frequency summary of all haplotypes. The second mode is intended for larger populations and shows difference of genes haplotypes repartition.

Enter your data

Variant file (.vcf / .vcf.gz, max 25MB) Parcourir... Aucun fichier sélectionné.
Gene Name (Symbol)
Include Synonymous Mutation ☑
Run Mode ◉ Individual ○ Population
Population file (only for population mode) Parcourir... Aucun fichier sélectionné.
Ok

FIGURE 2.8: Interface web de la version web de GEMPROT.

2.3 Comparaison

Un outil bioinformatique, Haplosaurus, a été publié en parallèle de notre outil et permet de visualiser les combinaisons de variants sur les haplotypes et les changements d'acides aminés en *cis* sur la protéine [Spooner et al., 2018]. Cet outil a été intégré au projet Ensembl et permet une récupération des haplotypes du projet 1000 Genomes. Tout comme GEMPROT, Haplosaurus part d'un fichier VCF phasé et réalise une lecture et une traduction globale des deux séquences d'un gène et l'application des variants présents dans le fichier VCF. Une des différences entre GEMPROT et Haplosaurus est le réalignement des haplotypes sur la séquence de référence. Les algorithmes d'alignement représentent généralement l'alignement de deux séquences sous forme de matrice. Ainsi le passage d'une case à une autre de la matrice représente le passage d'un nucléotide à un autre lorsque l'on essaie d'aligner deux séquences nucléotidiques. Des scores de passage, d'une case à une autre, sont assignés en fonction du changement : similarité de séquence, substitution ou gap (insertion/délétion).

Haplosaurus utilise l'algorithme de Myers et Miller, implémenté par BioPerl, qui part du principe que le meilleur alignement passe par la case de plus haut score de la colonne du milieu de la matrice [Myers and Miller, 1988, Stajich et al., 2002]. Au début de l'algorithme, la séquence A est divisée en deux, puis il est recherché dans la séquence B le résidu qui permet d'aligner au mieux B avec le premier résidu de A. Et ainsi de suite de part et d'autre du point central, jusqu'à ce que tous les résidus de chaque moitié de la séquence A soient alignés avec la séquence B. Pour GEMPROT, j'ai traduit indépendamment les séquences protéiques de chaque haplotype et je n'ai pas cherché à aligner globalement les deux séquences nucléotidiques mais plutôt à trouver les différences entre les séquences protéiques. La difficulté principale lorsqu'on aligne des séquences sont les insertions et les délétions. Si elles sont multiples de trois alors il n'y a pas de changement du cadre de lecture, un acide aminé sera alors supprimé. En revanche lorsqu'elles ne sont pas multiples de trois, toute la séquence protéique résultant va être modifiée. J'ai donc introduit, lors de la traduction, un symbole dans la séquence protéique lorsque la délétion ou l'insertion était un multiple de 3. De cette façon j'ai pu identifier les différences entre les deux séquences acide aminé par acide aminé en commençant au début de la séquence.

J'ai comparé les résultats obtenus par Haplosaurus et GEMPROT sur les données 1000G du gène *CFTR*. Du fait de l'absence des indels sur l'interface web d'Ensembl, je n'ai pas pu regarder les haplotypes porteurs de la délétion p.Phe508del. J'ai donc pris l'exemple d'une autre combinaison de variants sur le gène *CFTR* retrouvée chez 6 individus de 1000 Genomes. Les sorties de GEMPROT et de Haplosaurus dans

sa version ligne de commande et interface Ensembl sont présentées respectivement Figure 2.9 et 2.10. Nous obtenons bien les mêmes résultats. On peut cependant constater que Haplosaurus ne fournit pas de visualisation graphique de la protéine, ni la position des variations au sein des domaines fonctionnels connus de la protéine. Lorsqu'on exécute Haplosaurus sur un fichier VCF dans sa version en ligne de commande, il fournit un fichier avec l'ensemble des variants et des transcrits présents dans l'intervalle des positions du VCF. Ce système est intéressant quand on cherche les combinaisons de variants présents dans une région et leurs impacts sur la protéine. Malheureusement, lorsqu'on veut analyser un gène précis il est difficile d'obtenir des informations des sorties de l'outil Haplosaurus (Figure 2.9.A). De plus, le gène n'est pas clairement indiqué, seul l'identifiant Ensembl du ou des transcrits, ainsi que l'identifiant de la protéine correspondante, sont donnés.

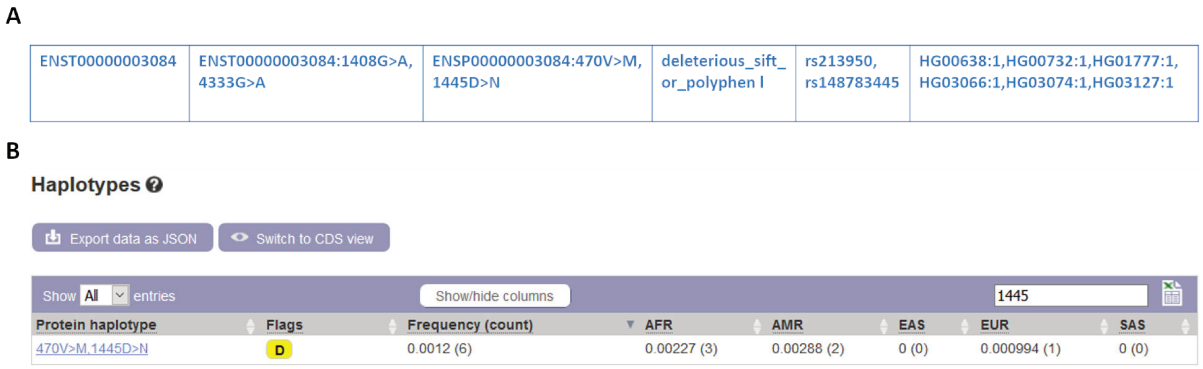


FIGURE 2.9: Sorties de Haplosaurus sur une combinaison de variants avec les deux versions disponibles de Haplosaurus. A. Résultats obtenus dans la version en ligne de commande. B. version disponible sur le serveur Ensembl.

A

Gene	haplotype	AFR	AMR	EAS	EUR	SAS	TOTAL
CFTR_CCDS5773.1	V:470:M,D:1445:N	3	2	0	1	0	6

B

Gene	Haplotype	No. of chromosomes	Frequency	Individuals
CFTR_CCDS5773.1	V:470:M,D:1445:N	6	0.0012	HG00638 HG00732 HG01777 HG03066 HG03074 HG03127

C

Haplotype 2 modifications

- 7:117199533 G>A amino acid position 470

  V=>M mismatch mutation
  Change from no charge, aliphatic, hydrophobic (2.3) amino acid to initiator of proteins, containing Sulfur, no charge, hydrophobic (1.1)
  ClinVar Information at this position :

  - rs213950 NC_000007.13:g.117199533G%3D Benign Hereditary_pancreatitis|not_specified **see clinvar**
  - rs213950 NC_000007.13:g.117199533G>A Benign/Likely_benign Cystic_fibrosis|not_specified|not_provided **see clinvar**

- 7:117307052 G>A amino acid position 1445

  D=>N mismatch mutation
  Change from acidic, negative charge, hydrophilic (-3.0)amino acid to amide of Aspartate (D), polar, hydrogen bonding, no charge, hydrophilic (-2.7)
  ClinVar Information at this position :

  - rs148783445 NC_000007.13:g.117307052G>A Uncertain_significance Cystic_fibrosis|Hereditary_pancreatitis|not_specified|not_provided **see clinvar**

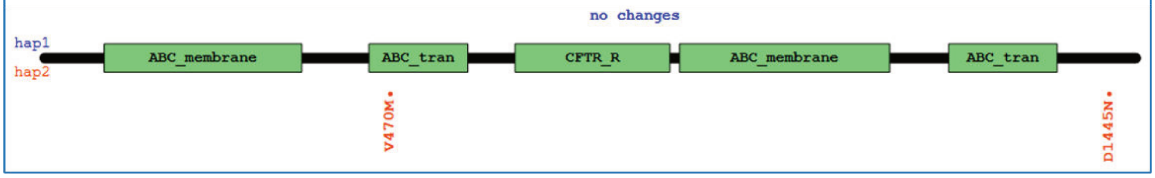


FIGURE 2.10: Sorties de GEMPROT sur une combinaison de variants dans le gène CFTR. A. Table de résultats du mode population. B. Table résumé du mode individuel. C. Résultat du mode individuel pour un individu portant la combinaison de variants.

2.4 Exemples d'Application de GEMPROT

2.4.1 HFE dans l'hémochromatose

L'hémochromatose liée au gène *HFE* pour « High Fe », est une maladie récessive autosomique de l'adulte touchant le métabolisme du fer. Elle est principalement associée au génotype homozygote p.Cys282Tyr. Un second variant p.His63Asp, peut être retrouvé en *trans* de p.Cys282Tyr chez les patients (on parle d'hétérozygote composite). Les conséquences pathologiques associées à cette hétérozygotie composite ne sont pas clairement établies. Elle pourrait conduire à de légères perturbations biologiques du métabolisme du fer sans toutefois aboutir à la constitution d'une réelle surcharge en fer [Walsh et al., 2006] Un allèle complexe p. [His63Asp ; Glu168Gln] a été trouvé en *trans* avec la mutation p.Cys282Y chez un petit nombre de patients atteints d'hémochromatose [Menardi et al., 2002, Uguen et al., 2017]. Cet allèle complexe a été détecté chez 4 individus brestois en utilisant GEMPROT sur les données FrEx (Annexe I) et confirmé par phasage moléculaire. En utilisant GEMPROT, il pourrait donc être plus facile de repérer cette combinaison de variants chez les patients.

2.4.2 Problèmes liés à l'annotation

Lors du traitement des données par GEMPROT, les deux haplotypes sont traduits dans leur intégralité en tenant compte de toutes les mutations. Ainsi, au lieu de traduire chaque mutation une à une comme le font la plupart des outils d'annotation de séquences, on évite d'interpréter de manière erronée les modifications apportées aux acides aminés.

2.4.2.1 Cas des mutations dans le même codon

Lorsque deux variants affectant le même codon sont situés en *cis*, les modifications des acides aminés fournies par la plupart des annotateurs sont incorrectes. Dans l'exemple de la Figure 2.11, nous pouvons voir que le premier variant modifie le premier nucléotide de C en A et ne modifie pas l'acide aminé (Arg). Le second change la troisième base A en C et ne change pas non plus l'acide aminé (Arg). Ainsi, dans le fichier VCF, on retrouve deux lignes qui correspondent à l'un et l'autre de ces variants et l'annotation de chaque variant est « synonymous_variant ». Cependant, si un individu possède ces deux variants sur le même haplotype, il en résulte un changement d'acide aminé (Ser).

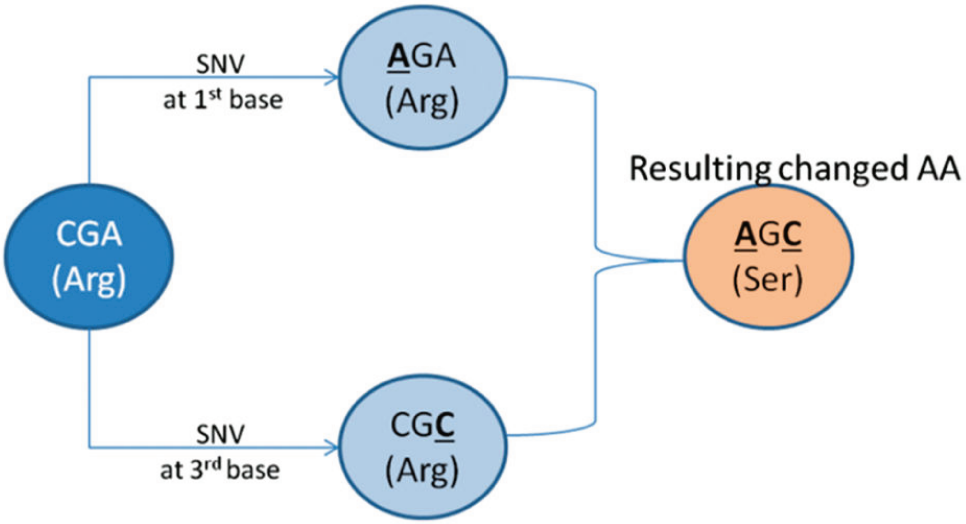


FIGURE 2.11: Illustration de l'impact de la présence de deux variations d'un seul nucléotide dans le même codon génétique sur le changement d'acide aminé (d'après Khan et al., 2018).

On peut prendre un autre exemple à partir du panel de FrEx (Figure 2.12) : il s'agit de deux variants faux-sens sur le même codon. Le premier change une Alanine (A) en Glutamate (E) et le deuxième change l'Alanine en Thréonine (T) lorsqu'on considère ces variants indépendamment. Lorsque ces variants sont en *cis* en revanche, on obtient un changement de l'Alanine (A) en Lysine (K). La Figure 2.12 montre le résultat de l'annotation issue de deux annotateurs, SnpEff et VEP. Les annotateurs ne prennent pas en compte la phase car il serait compliqué d'annoter un fichier VCF de plusieurs individus, certains individus étant porteurs des deux variants et d'autres n'en ayant qu'un seul. GEMPROT lit l'intégralité de la séquence nucléotidique en tenant compte de tous les variants pour traduire la séquence des deux haplotypes. De cette façon nous n'avons pas cette erreur lors de l'annotation du variant. La sortie de GEMPROT est donc deux variants nucléotidiques, mais une seule variation au niveau de la séquence protéique (Figure 2.13).

Autour de cette question de l'annotation des variants en cis dans le même codon, j'ai été amenée à collaborer à la mise en place d'un nouvel outil MACARON (pour Multi-bAse Codon-Associated variant Re-annotatiON). Cet outil permet d'identifier et d'annoter les SNV multiples survenant dans le même codon (Annexe II) (Khan et al., 2018).

A

```
##SnpEffVersion="4.3i (build 2016-12-15 22:33), by Pablo Cingolani"
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: ' Annotation | HGVS.p' ">
#CHROM  POS        ID          REF  ALT  QUAL  FILTER  INFO                                  FORMAT     B00FWXR
14      95921733   rs771184132 G    T    .     PASS    ANN=missense_variant|p.Ala373Glu      GT:DS:GP   0|1:1:0,1,0
14      95921734   rs781208017 C    T    .     PASS    ANN=missense_variant|p.Ala373Thr      GT:DS:GP   0|1:1:0,1,0
```

B

```
##INFO=<ID=CSQ,Description="Consequence annotations from Ensembl VEP. Format: Consequence|Protein_position|Amino_acids">
#CHROM  POS        ID          REF  ALT  QUAL  FILTER  INFO                                  FORMAT     B00FWXR
14      95921733   rs771184132 G    T    .     PASS    CSQ=missense_variant|373|A/E          GT:DS:GP   0|1:1:0,1,0
14      95921734   rs781208017 C    T    .     PASS    CSQ=missense_variant|373|A/T          GT:DS:GP   0|1:1:0,1,0
```

FIGURE 2.12: Lignes de deux variants faux-sens, impactant le même codon retrouvés en *cis* chez un individu du panel FrEx, dans un fichier VCF avec les annotations apportées par deux annotateurs. A. Annotation par SnpEff. B. Annotation par VEP.

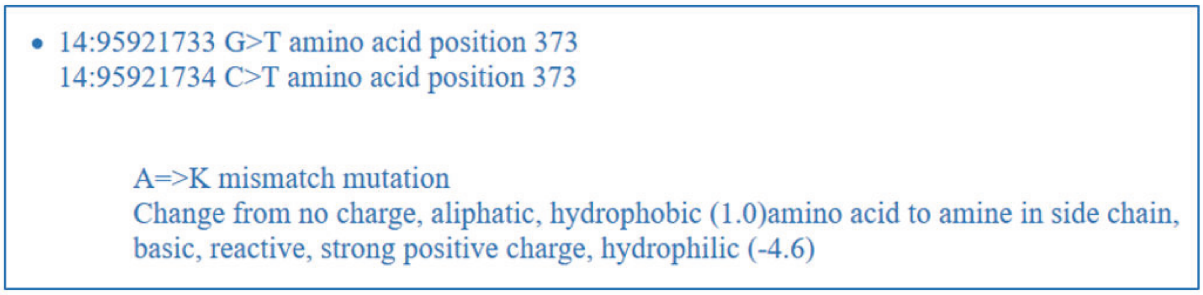


FIGURE 2.13: : Sortie de GEMPROT pour deux variants faux-sens impactant le même codon retrouvés en *cis* chez un individu du panel FrEx.

2.4.2.2 Cas des mutations décalant le cadre de lecture

Dans le cas des mutations « frameshift » qui décalent le cadre de lecture les annotateurs caractérisent le variant par rapport à sa position et non par les conséquences du décalage. Prenons l'exemple de deux variants sélectionnés chez deux individus du panel FrEx.

Le premier variant est une délétion d'un G qui entraîne un décalage du cadre de lecture à partir de la position 583 d'une protéine de 604 acides aminés, la protéine CES5A. Ce variant induit donc des variants non-sens jusqu'à la fin de la traduction de la protéine et entraîne une élongation. Ainsi, le codon à la position 604 n'est plus un codon stop. Pour cet individu, sur le gène codant la protéine CES5A, GEMPROT annote ce variant comme un variant entraînant un décalage du cadre de lecture « frameshift » et identifie la conséquence de ce variant, qui est ici une élongation de la protéine (Figure 2.14). En revanche les annotateurs SnpEff et VEP n'identifient pas le résultat de ce décalage. SnpEff et VEP utilisent la base de données Ensembl afin de caractériser le variant (Annexe III). Dans ce cas, le variant est uniquement annoté comme « frameshift_variant » alors qu'il devrait être également annoté comme « stop_loss » dans la mesure où il induit une perte du codon stop (Figure 2.15). Pour déduire les conséquences du variant génétique sur la séquence protéique, SnpEff

fait appel à la base de données « Human Genome Variation Society » (HGVS) [den Dunnen et al., 2016] et annote le variant p.His583fs. Cela indique qu'à la position 583 de la séquence protéique il existe une Histidine et que le variant entraîne un frameshift (fs). En revanche, VEP se sert de la séquence de référence du génome et du CCDS afin d'identifier les conséquences du variant. Il donne l'annotation correcte, à savoir p.His583ArgfsTer38. Ceci signifie qu'on a bien une Histidine à la position 583, que l'ajout du C dans la séquence modifie celle-ci en Arginine, puis qu'il y a un frameshift et pour finir un codon stop 38 acides aminés plus loin en partant de la position 583 (Figure 2.15). La protéine ayant de 604 acides aminés, on en conclut qu'il y a une élongation. VEP permet d'avoir une information de plus que GEMPROT, à savoir la position du prochain codon stop, mais il est difficile de ressortir directement la notion d'élongation.

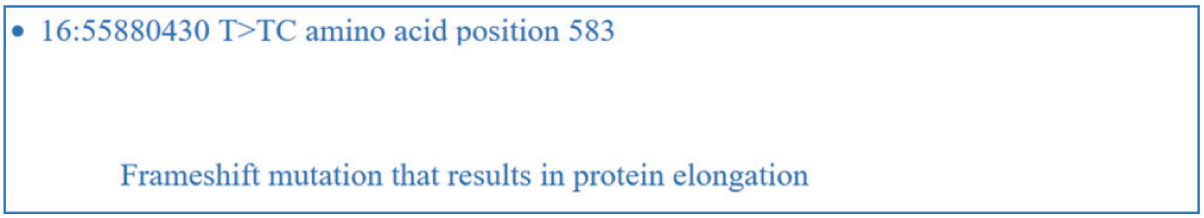


FIGURE 2.14: Sortie de GEMPROT d'un variants entraînant une élongation de la protéine chez un individu de FrEx.

A

```
##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"
##INFO=<ID=ANN,Description="Functional annotations: ' Annotation | HGVS.p ' ">
#CHROM  POS       ID          REF  ALT  FILTER  INFO                              FORMAT    B00G74R
16      55880430  rs750154282 T    TC   PASS    ANN= frameshift_variant|p.His583fs  GT:DS:GP  1|0:1:0,1,0
```

B

```
##INFO=<ID=CSQ,Description="Consequence annotations from Ensembl VEP : Consequence|Protein_position|Amino_acids| HGVS.p" >
#CHROM  POS       ID          REF  ALT  FILTER  INFO                                          FORMAT    B00G74R
16      55880430  rs750154282 T    TC   PASS    CSQ=frameshift_variant|583|H/RX|p.His583ArgfsTer38  GT:DS:GP  0|1:1:0,1,0
```

FIGURE 2.15: Ligne d'un variant entraînant une élongation de la protéine dans un fichier VCF avec les annotations apportées par deux annotateurs. A. Annotation par SnpEff. B. Annotation par VEP.

Le second variant est une insertion d'un C qui intervient dans la formation d'un codon à la position 157 et entraîne un codon stop prématuré à la position 315 sur la protéine ZSCAN32 qui a 697 acides aminés. GEMPROT identifie la position du codon stop prématuré (Figure 2.16). Dans le cas des annotateurs VEP et SnpEff, le variant est également noté « frameshift_variant » mais devrait donc également être annoté « stop_gained ». Comme précédemment, SnpEff reproduit uniquement les informations en nomenclature HGVS et ne signale donc pas l'existence d'un codon stop prématuré. VEP permet d'avoir la bonne information (p.Pro157GlnfsTer159) qui

est le changement d'une proline en glutamine en position 157 suivi d'un « frameshift », puis un codon stop 38 acides aminé après, c'est-à-dire à la position 315 (Figure 2.17).

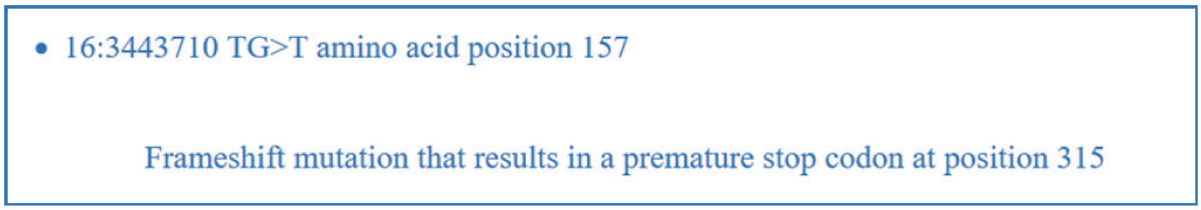


FIGURE 2.16: Sortie de GEMPROT d'un variant entraînant un codon stop prématuré éloigné de la position du variant chez un individu de FrEx.

A

```
##SnpEffVersion="4.3i (build 2016-12-15 22:33), by Pablo Cingolani"
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: ' Annotation | HGVS.p' ">
#CHROM  POS      ID  REF  ALT  FILTER  INFO                                  FORMAT     B00FWXR
16      3443710  .   TG   T    PASS    ANN=frameshift_variant| p.Pro157fs     GT:DS:GP   1|0:1:0,1,0
```

B

```
##INFO=<ID=CSQ,Description="Consequence annotations from Ensembl VEP : Consequence|Protein_position|Amino_acids|HGVS.p ">
#CHROM  POS      ID  REF  ALT  FILTER  INFO                                              FORMAT     B00G5XP
16      3443710  .   TG   T    PASS    CSQ=frameshift_variant|157|P/X| p.Pro157GlnfsTer159  GT:DS:GP   1|0:1:0,1,0
```

FIGURE 2.17: Lignes d'un variant entraînant un codon stop prématuré éloigné de la position du variant dans un fichier VCF avec les annotations apportées par deux annotateurs. A. Annotation par SnpEff. B. Annotation par VEP.

2.4.3 CFTR dans la mucoviscidose

Sur les données de séquençage du gène *CFTR* issues du projet 1000 Genomes, GEMPROT révèle 137 haplotypes différents en considérant les variants modifiant la séquence protéique et 197 en ajoutant les variants synonymes. On note une différence de répartition des variants du gène *CFTR* en fonction des populations. Ainsi, la mutation p.Phe508del est retrouvée chez 20 individus de 1000 Genomes dont 9 en Europe. On peut également noter que cette délétion n'est pas retrouvée dans les populations d'Afrique et d'Asie du Sud. Un seul variant est retrouvé en *cis* avec p.Phe508del : il s'agit du variant p.Val470Met qui est présent sur tous les haplotypes portant la délétion. Deux autres variants montrent une spécificité régionale marquée, il s'agit des variants p.Ile556Val et p.Gln1352His qui sont retrouvés l'un et l'autre chez respectivement 54 et 20 individus différents, tous issus de la population d'Asie de l'Est (EAS : Chine, Japon, Vietnam). Il est à noter que ces variants ont été observés avec des fréquences plus élevées (p.Gln1352His : 6,94%, p.Ile556Val : 3,47%) dans un article répertoriant le spectre génétique de patients chinois présentant une absence congénitale de canaux déférents [Yuan et al., 2019].

Une combinaison de variants, la combinaison p.[Arg74Trp ;Asp1270Asn], montre également une distribution hétérogène dans les populations du panel 1000 Genomes puisqu'on ne la trouve qu'en Afrique où elle est portée par 25 individus. Cette combinaison de deux variants a été décrite chez des patients atteints de mucoviscidose et associée à des défauts de conductance du canal CFTR [Fanen et al., 1999]. Deux des individus de la population africaine du panel 1000 Genomes portent également un troisième variant, p.Val201Met en *cis* avec p.[Arg74Trp ;Asp1270Asn]. Cette combinaison de 3 variants a été plusieurs fois décrite chez des individus ayant une absence congénitale des canaux déférents lorsqu'elle est associée, en *trans*, à un autre variant [Claustres et al., 2004, Brugnon et al., 2008].

Sur les données de séquençage d'exomes du projet FrEx, GEMPROT révèle 41 haplotypes différents et 75 avec les variants synonymes pour le gène *CFTR*. Le panel FrEx étant constitué uniquement d'individus d'origine européenne, nous n'avons pas observé les variants ou combinaisons de variants décrits ci-dessus pour les populations asiatiques ou africaines. Je me suis donc intéressée essentiellement à la délétion p.Phe508del qui est portée par 13 individus du panel FrEx (2.27%), toujours associée en *cis* au variant p.Val470Met. On observe que la délétion p.Phe508del est plus fréquente dans l'échantillon brestois (4.0%), ce qui était attendu car cette délétion est plus fréquente en Bretagne. Dans l'échantillon brestois, un individu porte également en *cis* le variant p.Ile1027Thr déjà décrit comme pouvant être associé à la délétion p.Phe508del chez des patients brestois [Fichou et al., 2008].

Le variant p.Val470Met est toujours associé en *cis* à la délétion p.Phe508del dans toutes les populations que j'ai étudiées mais également dans d'autres populations [Vecchio-Pagán et al., 2016]. Ce variant existe également en population générale non associée à la délétion p.Phe508del. Il est donc très probable que la délétion p.Phe508del soit apparue sur un chromosome portant le variant génétique codant pour une méthionine à la position 470 (470M). Bien que ce variant soit considéré comme l'allèle ancestral, l'allèle de référence sur le génome humain porte une Valine en position 470 (470V) (c.1408G > c.1408A, p.Val470Met) [Kosova et al., 2010]. Ces deux variants ont des fréquences très variables d'une région géographique à l'autre. Les fréquences du variant 470M dans 1000 Genomes et gnomAD sont de 41,8% et 48,65% si on ne tient pas compte des origines géographiques [1000 Genomes Project Consortium et al., 2015, Lek et al., 2016]. Dans les populations africaines cependant, sa fréquence est de 93,5% et 86,95% respectivement, pour l'un et l'autre des panels (Figure 2.18). Ce variant est considéré comme un polymorphisme ne causant pas la mucoviscidose. Mais il est parfois rapporté comme ayant des conséquences sur la fonction du CFTR et/ou étant associé ou prédisposant à des phénotypes particuliers, par exemple à des taux de natalité plus faibles chez les hommes [Cuppens et al., 1998, Lázaro et al., 1999, Wang et al., 2010, Stankovic et al., 2008, Kosova et al.,

2010].

Les plasmides commerciaux se basent sur la séquence de référence qui contient le variant codant pour une valine en position 470 (exemple : Origene et Vigene Biosciences, NP_000483). Si on réalise l'analyse fonctionnelle de l'effet de la délétion 508del après mutagenèse dirigée en utilisant ces plasmides, on associera la délétion au variant 470V, ce qui ne représentera pas son contexte haplotypique naturel.

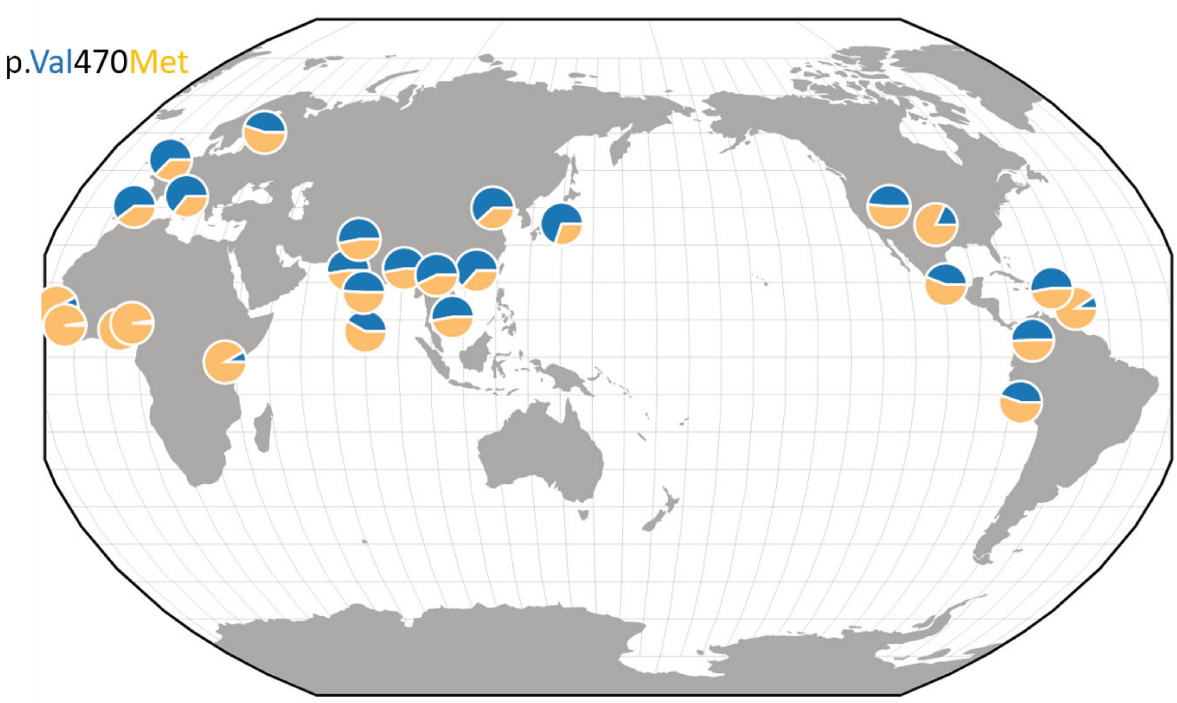


FIGURE 2.18: Répartition géographique du variant p.Val470Met dans le panel 1000 Genomes. (Marcus and Novembre, 2017).

Le variant qui transforme une Isoleucine en thréonine (1027T) n'est retrouvé que sur une partie des haplotypes 508del et cela plus fréquemment en Bretagne [Fichou et al., 2008, Vecchio-Pagán et al., 2016]. Dans la base de données CFTR2, qui recense les données cliniques de patients atteints de mucoviscidose, le variant 1027T est retrouvé chez 36 patients, tous porteurs de la 508del [Castellani and CFTR2 team, 2013]. Elle est également fréquemment rapportée en *cis* avec 508del [Fanen et al., 1992, Ziedalski et al., 2006, Penmatsa et al., 2009, El-Seedy et al., 2012]. Il semblerait donc que ce variant soit apparu sur un chromosome portant la délétion car il n'est pas retrouvé seul chez les patients. Ce variant est donc souvent défini avec un impact incertain de par son association avec la 508del. Son impact sur la fonction du CFTR n'a, à notre connaissance, pas été étudié. En effet, le fait que ce variant soit toujours associé à la délétion 508del a conduit à considérer qu'il n'aurait probablement qu'une contribution à la maladie négligeable face à la 508del qui entraîne déjà une altération importante de la fonction. Dans la littérature, le variant est rapporté de manière variable

comme un polymorphisme ou une mutation délétère [Suarez et al., 2012, Castellani and CFTR2 team, 2013, Landrum et al., 2016]. Ce variant permet donc de distinguer deux types d'haplotypes porteurs de la 508del et pourrait avoir un effet modificateur sur la protéine CFTR-508del.

CHAPITRE 3

CONSÉQUENCE DU CONTEXTE HAPLOTYPIQUE SUR LA PROTÉINE CFTR

Sommaire

Dans la suite de mon travail, j'ai étudié l'impact du contexte haplotypique de la délétion p.Phe508del sur la fonction de la protéine CFTR. De nombreuses études se sont déjà focalisées sur l'impact de cette délétion, mais le contexte haplotypique est rarement décrit [Vecchio-Pagán et al., 2016]. Je me suis servi des résultats obtenus à l'aide de GEMPROT sur les données de FrEx, qui ont permis d'identifier deux variants associés en *cis* avec la délétion. Le premier variant, qui transforme une valine en méthionine en position 470 (p.Val470Met), est présent sur tous les haplotypes porteurs de la délétion en position 508. Le second variant, qui transforme une Isoleucine en Thréonine (p.Ile1027Thr), n'est présent que sur quelques haplotypes porteurs de la délétion. J'ai donc cherché à caractériser l'impact de ces variants sur l'expression, la localisation et la fonction du CFTR.

3.1 Matériels et méthodes

3.1.1 Cultures cellulaires et transfection

Les combinaisons de variants sont obtenues par mutagenèse dirigée de l'ADNc du CFTR introduit dans des plasmides qui sont ensuite transfectés de façon transitoire dans des cellules HEK293T. Les HEK293T sont des cellules embryonnaires de reins (Human Embryonic Kidney). Nous avons choisi ce modèle car les HEK293T n'expriment ni l'ARNm ni les protéines CFTR de façon endogène. Il est donc plus aisé de déterminer le phénotype de chaque mutant et combinaison de mutants. C'est aussi un excellent modèle pour étudier la fonction protéique, car ces cellules sont d'origine humaine et expriment facilement les protéines exogènes [Domingue et al., 2014]. Elles permettent donc la maturation de la protéine CFTR. En raison de l'expression de l'antigène T SV40 dans ses cellules, les plasmides transfectés qui portent l'origine de réplication SV40 peuvent se répliquer et conservent de manière transitoire un nombre élevé de copies [DuBridge et al., 1987]. Cela augmente la quantité de protéines recombinantes produites. Les cellules de lignée HEK293T sont cultivées dans un milieu DMEM (pour Dulbecco/Vogt modified Eagle's minimal essential medium) complété avec 10% de sérum de veau fœtal, dans une atmosphère humidifiée sous 5% de CO2, à 37℃.

Le plasmide utilisé pour les transfections est le plasmide pcDNA3.1 conçu pour la surexpression dans les cellules de mammifère. Il contient l'origine de réplication SV40 nécessaire pour augmenter la production de protéine. Il contient également un activateur-accélérateur de cytomégalovirus (CMV) pour l'expression en grande quantité, un grand site de clonage multiple, dans lequel nous avons introduit l'ADNc du CFTR, et le gène de résistance à l'ampicilline pour sélection dans E. coli (Figure 3.1). Nous disposions de deux types d'ADNc du gène *CFTR*, l'un porteur du variant conduisant à la délétion p.Phe508del (508del) et l'autre ne portant pas ce variant (WT). Dans les deux cas, les ADNc comportaient le variant codant pour une méthionine en position 470. Pour obtenir également des versions d'ADNc codant pour une valine en position 470, j'ai réalisé des mutagenèses dirigées pour pouvoir disposer de 4 types de plasmides : 470M-508del, 470V-508del, 470M- WT et 470V-WT. J'ai fait de même avec la position 1027 afin d'obtenir les plasmides : 470M-508del-1027T et 470M-WT-1027T. J'ai ensuite réalisé les transfections transitoires des plasmides dans des cellules HEK293T à l'aide d'un agent de transfection, la LipoD293.

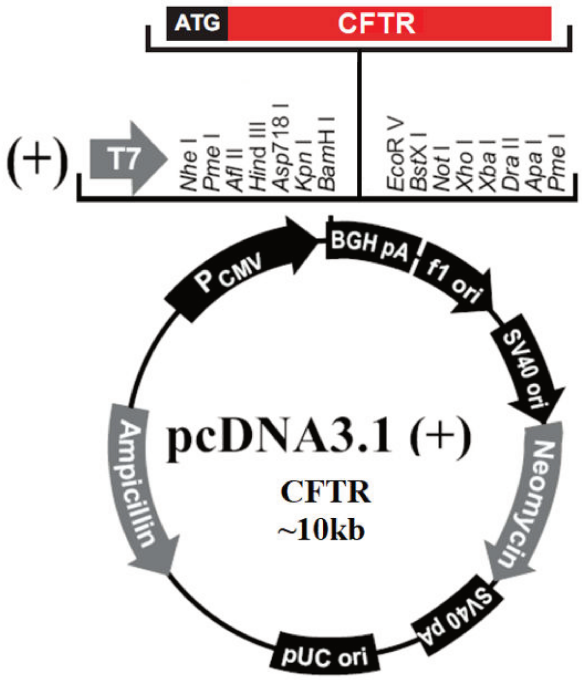


FIGURE 3.1: Représentation du plasmide pcDNA3.1 avec l'insert de l'ADNc du CFTR.

3.1.2 Profil électrophorétique

La technique du western blot va pouvoir nous indiquer, à l'aide de marqueur de taille, la masse moléculaire apparente en kDa de la protéine CFTR présente dans les cellules transfectées. Ainsi comme on l'a vu dans le chapitre sur la biosynthèse de la protéine CFTR, la protéine CFTR possède 3 formes de glycosylation. Une forme non glycosylée de 130 kDa, correspondant à la bande A en western blot, est le polypeptide juste issu de la traduction. Une forme dite immature de 145 kDa, correspondant à la bande B en western blot, est la protéine partiellement glycosylée dans le RE. La dernière forme, Celle mature de 180 kDa, correspondant à la bande C en western blot, est la protéine totalement glycosylée présente à la membrane plasmique. Le profil électrophorétique de la protéine CFTR WT montre la présence d'une bande C et d'une bande B avec une bande C plus intense. Le profil électrophorétique de la protéine CFTR 508del ne laisse apparaitre qu'une bande B représentant la protéine immature.

Après 48h de transfection, les cellules HEK293T sont lysées dans du tampon RIPA (Tris-HCl 25 mM pH 7,5, NaCl 150 mM, Triton X-100 1%, Na-désoxycholate 1%, SDS 0,1%, iodoacétamide 10 mM, PMSF 100 µM) additionné d'inhibiteurs de protéases ajoutés extemporanément à une concentration finale de 40 µl/ ml. Les lysats protéiques sont récupérés et leurs concentrations sont déterminées en utilisant la méthode de Folin-Lowry [Lowry et al., 1951]. 40 µg de protéines sont chargées sur un gel SDS-PAGE à 7,5%. Après migration, les protéines sont transférées sur une membrane de Polyfluorure de vinylidène (PVDF) puis la membrane est saturée dans une solution de

TBST + lait (Tris-Buffered Saline (TBS) 1X + tween20 0,1% + lait 5%). Les membranes sont ensuite incubées avec un anticorps polyclonal anti-CFTR de souris pendant la nuit à 4℃ (M3A7 spécifique au domaine NBD2 C-terminus, 1/500e). Après plusieurs lavages au TBST, les membranes sont incubées avec l'anticorps secondaire couplé à la peroxydase de raifort (HRP : horseradish peroxidase, anti-souris, 1/20000e). Les membranes sont révélées par chimiluminescence (ECL Plus, GE Healthcare). La réaction de HRP avec son substrat donne un dérivé luminescent qui permet de détecter la protéine. Le signal obtenu est proportionnel à la quantité de protéine et permet donc de mesurer la quantité relative de protéine CFTR [Veitch, 2004].

3.1.3 Localisation cellulaire

Dans la technique d'immunofluorescence (IF) indirecte, l'immunomarquage permet de localiser un antigène dans un tissu ou dans une cellule à l'aide d'un anticorps spécifique et d'un anticorps secondaire conjugué à un fluorochrome. La fluorescence est ensuite détectée à l'aide de microscopie à fluorescence. L'immunofluorescence est réalisée après 48h de transfection sur les cellules HEK293T cultivées sur lamelles de verre. Les cellules sont fixées avec du paraformaldéhyde et perméabilisées (Triton 0,1%) pour permettre la pénétration des anticorps. Les cellules sont alors incubées avec l'anticorps primaire anti-CFTR de souris (24-1, spécifique au domaine C-terminus 1/50e) puis incubées avec l'anticorps secondaire anti-souris conjugué à la cyanine 3 (anti-souris/cy3, 1/400e). Les noyaux sont ensuite contre colorés au Hoechst. Un marquage CFTR et un marquage Calnexine (marqueur du RE) est réalisé afin de permettre d'effectuer une comparaison de localisation entre ces deux protéines. Le CFTR 508del est retrouvé dans le RE.

3.1.4 Fonction du canal CFTR

J'ai ensuite étudié la fonction du canal CFTR par deux techniques : la technique FLIPR® et la technique du Patch Clamp. La technique FLIPR® (Fluorescent Imaging Plate Reader) permet de réaliser un criblage rapide de plusieurs cellules et de plusieurs transfections simultanément [Schroeder and Neagle, 1996]. Cette technique est couramment utilisée pour tester les modulateurs de canaux ioniques normaux ou mutants, surexprimés dans les cellules HEK293 [Van Goor et al., 2006, Maitra et al., 2013, Molinski et al., 2015]. Elle est basée sur le suivi de l'activité du canal CFTR en utilisant une molécule fluorescente sensible au potentiel de membrane. Cette technique a été développée spécifiquement pour étudier l'activité des canaux ioniques [Maitra et al., 2013, Ahmadi et al., 2017]. L'activité du canal CFTR est suivie en utilisant une molécule fluorescente ou fluorochrome Blue-Dye (Molecular

Devices). Le fluorochrome est une molécule chimique lipophile qui se répartit dans la cellule à travers la membrane plasmique des cellules vivantes, en fonction du potentiel membranaire. Il est activé lorsque qu'il se lie aux protéines cytosoliques. En absence de stimulation, la membrane de la cellule est à son potentiel de repos et la fluorescence émise, après excitation à 530nm, correspond à la fluorescence basale (RFU : Relative fluorescence units). Au contraire, lorsque la cellule est dépolarisée, une plus grande quantité de fluorochrome y pénètre et se lie aux protéines cytosoliques, ce qui provoque une augmentation de la fluorescence. Ce phénomène est observé lorsque le canal CFTR est activé (Figure 3.2). Lorsque les cellules sont hyperpolarisées, quand on inhibe le CFTR, le fluorochrome sort des cellules ce qui entraîne une diminution de la fluorescence. Les expériences de FLIPR® sont réalisées avec un lecteur de plaques 96 (Varioskan) puis avec une méthode multipoints.

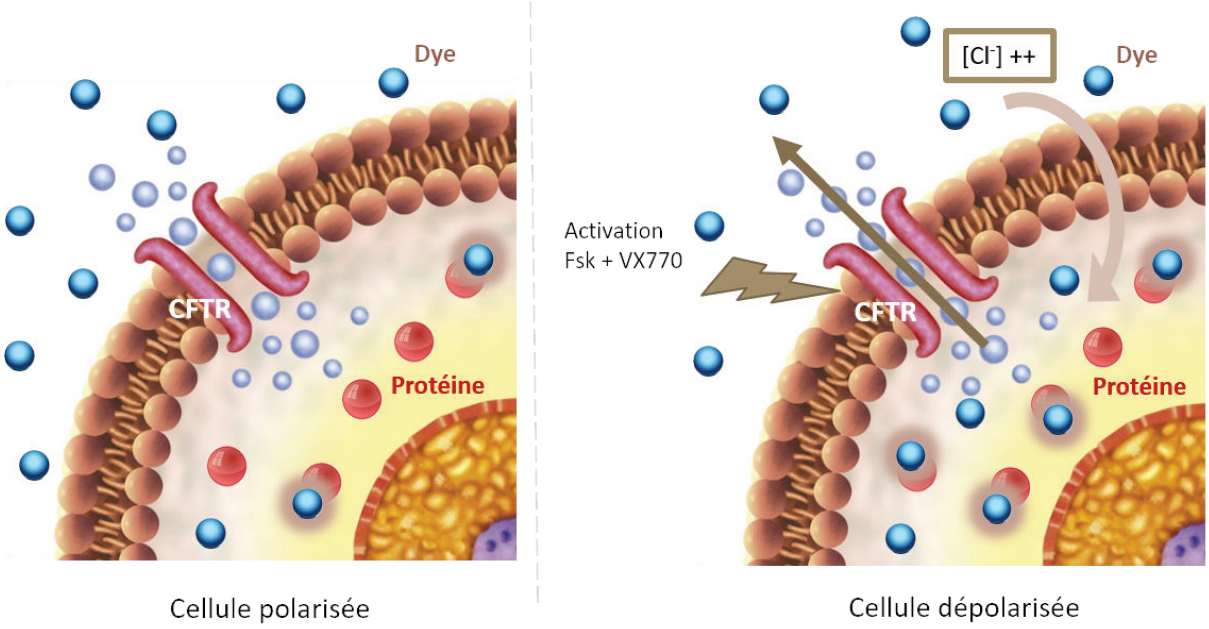


FIGURE 3.2: Technique FLIPR®. A gauche la membrane plasmique est à son potentiel de repos. Le fluorochrome se répartit dans la cellule. La fluorescence, émise lors de sa liaison aux protéines cytosoliques, quand la cellule est polarisée, est la fluorescence basale. A droite la cellule est dépolarisée par l'activation du canal CFTR. Les ions chlorures sortent de la cellule. Une plus grande quantité de fluorochrome pénètre dans la cellule et se lie aux protéines cytosoliques, ce qui provoque une augmentation de la fluorescence.

Après 30h de transfection, on ajoute aux cellules une solution de fluorochrome diluée dans un tampon dépourvu d'ions chlorure et ions sodium (Cl^-/Na^+ free). De cette manière la cellule ne sera pas dépolarisée. En effet comme nous l'avons vu précédemment le canal CFTR permet la sortie des ions chlorures et régule les canaux ENaC responsables du transport d'ions sodium. On obtient ensuite les valeurs de fluorescence pour chacune des transfections. Nous avons mis en place le même protocole que dans l'article de Ahmadi et al. Nous avons donc ajouté au milieu de

culture cellulaire une solution d'activateur du CFTR (forskoline (FSK) et VX-770 1µM), ou d'inhibiteur (CFTRinh172, 10 mM) pour activer ou inhiber le CFTR, respectivement. L'activation du CFTR va entraîner une sortie d'ion chlorure qui va déclencher la dépolarisation de la membrane plasmique. Ainsi, lorsque la protéine CFTR est fonctionnelle, on observera une augmentation de la fluorescence. L'augmentation de la fluorescence reflète le degré de fonction du canal CFTR. Enfin, la même transfection a été réalisée plusieurs fois à chaque expérience et plusieurs mesures ont été prises pour chaque transfection avec une méthode multipoints (Figure 3.3). Ainsi, j'ai plus de valeurs pour augmenter la puissance du test statistique.

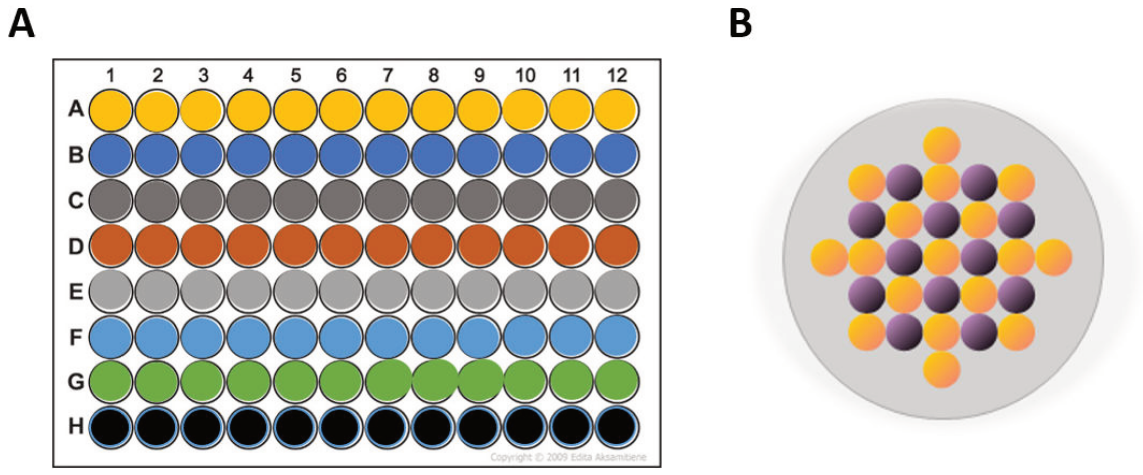


FIGURE 3.3: Schéma de la méthodologie utilisée pour augmenter le nombre de mesure. A.Plan de plaque 96 puits, chaque couleur représente une transfection. B. Méthode multipoint du lecteur de plaque. Chaque point jaune représente une mesure prise par le lecteur.

La technique du patch clamp permet à l'aide d'un appareil d'électrophysiologie automatique (Port-a-Patch, Nanion Technologies GmbH, Allemagne) dans mon cas, d'appliquer des dépolarisations membranaires mais également de contrôler la composition des tampons afin de d'enregistrer le courant ionique passant par le canal CFTR [Farre et al., 2007]. Ces expériences sont effectuées en configuration cellule entière, conformément aux précédents travaux de l'équipe (Figure 3.4) [Benz et al., 2014, Trouvé et al., 2007, Huguet et al., 2016]. Les cellules HEK293T transfectées sont remises en suspension dans une solution extracellulaire contenant 140 mM de NaCl, 2 mM de CaCl2, 1 mM de MgCl2, 10 mM d'HEPES et 5 mM de D-glucose monohydrate (pH 7,4, 298 mOsmol). La solution intracellulaire contient 50 mM de CsCl, 10 mM NaCl, Cs-Fluorure 60 mM, EGTA 20 mM et HEPES / CsOH 10 mM (pH 7,2, 285 mOsmol). Les impulsions de tension-clamp sont générées et les données sont capturées à l'aide du programme Patchmaster (Nanion Technologies). Pour les enregistrements, on applique des dépolarisations de 10mV (100ms) de -80 à + 80mV avec un potentiel

de membrane maintenu à -80mV. Une solution d'inhibiteur de CFTR (CFTRinh172, 10 µM) et une solution d'activateur CFTR (Forskoline, 10 µM et Génistéine, 20 µM), sont ajoutées successivement à la solution extracellulaire pour inhiber ou activer CFTR, afin de s'assurer de la spécificité du signal enregistré. L'intensité du courant ionique, passant par les canaux CFTR de la cellule, est exprimée en ampère (A) et est normalisée par les capacitances. La capacitance membranaire est la capacité électrique associée à la membrane plasmique, exprimée en Farads (F) [Golowasch and Nadim, 2014]. Elle est proportionnelle à la surface de la cellule.

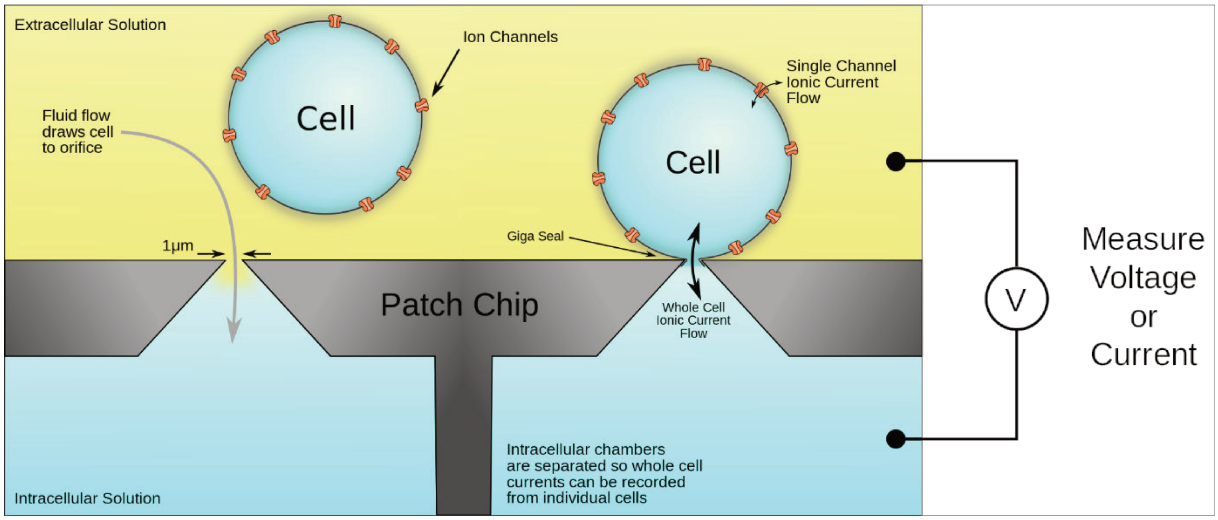


FIGURE 3.4: Schéma de la technique de Patch Clamp montrant une configuration d'enregistrement cellules entières.

3.1.5 Analyses Statistiques

Afin de savoir si la Valine ou la Méthionine ont un impact sur la fonction du CFTR dans un contexte 508del ou WT, j'ai réalisé des tests statistiques sur les résultats des analyses fonctionnelles en utilisant le logiciel R [Ihaka and Gentleman, 1996] Pour les données obtenues par la méthode FLIPR®, j'ai utilisé un modèle mixte afin d'évaluer les différences de fonction du CFTR d'une transfection à l'autre (fonction lmer sous R). Les modèles mixtes permettent d'expliquer une variable dépendante par une ou plusieurs variables explicatives [Oberg and Mahoney, 2007]. On parle de modèles mixtes car ils intègrent à la fois des variables explicatives fixes et aléatoires. Dans notre modèle, le type de transfection qui est la variable explicative qui nous intéresse est introduite avec un effet fixe. En revanche, les analyses étant reproduites plusieurs fois on ne peut pas ignorer le fait que la fluorescence fluctue d'une expérience à l'autre. J'ai donc pris en compte ce possible artefact en introduisant l'effet expérience dans notre modèle sous la forme d'un effet aléatoire. En intégrant ce paramètre, je vais mesurer les différences

entre les transfections tout en tenant compte des différences entre les expériences. La valeur de fluorescence obtenue après activation a été normalisée par la valeur de fluorescence basale.

L'équation du modèle est donc la suivante :

$$\frac{RFUact - RFU_0}{RFU_0} = \alpha + \beta T + \upsilon E + \epsilon$$

- α = Intercept
- β = Effet fixe
- T = Variable lié à la transfection
- υ = Effet aléatoire
- E = Variable lié à l'expérience
- ϵ = Erreur aléatoire

Pour la comparaison des résultats de Patch Clamp, les différences d'intensité de courant entre les transfections ont été évaluées par un test t de Student bidirectionnel non apparié.

Une visualisation graphique des moyennes ± écart type est donnée pour les deux méthodes.

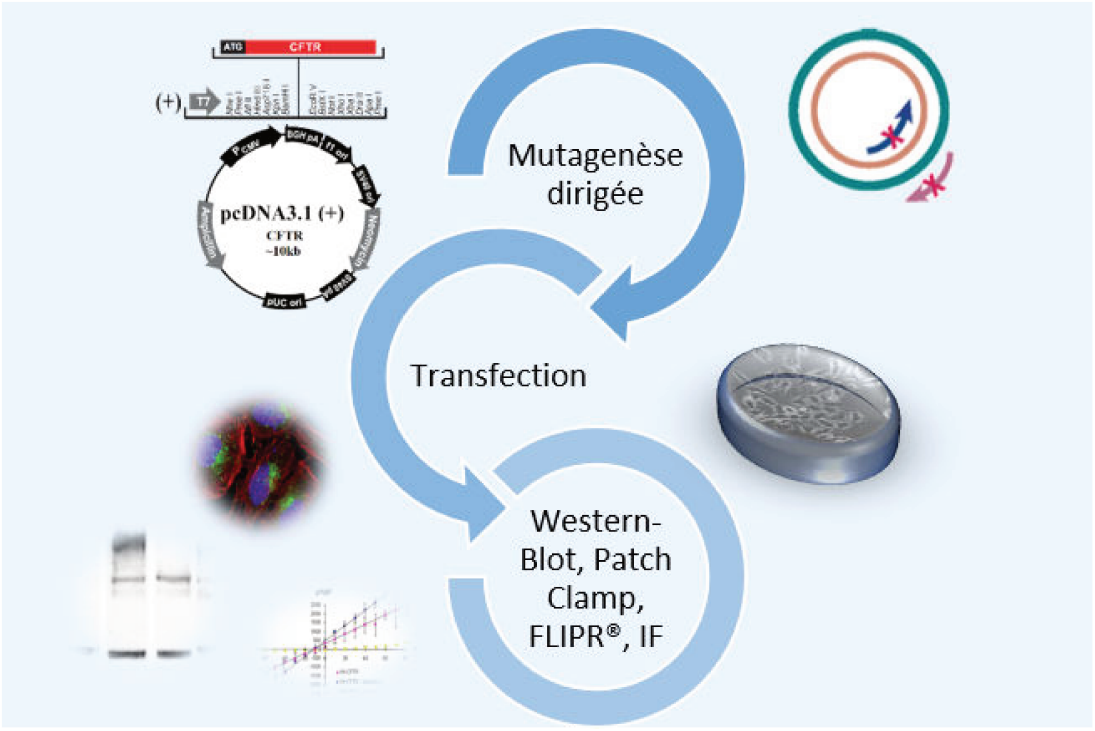


FIGURE 3.5: Protocole expérimental utilisé pour définir les phénotypes des différents variants et combinaison de variants dans les cellules HEK293T transfectées.

3.2 Effet de deux combinaisons de variants sur la protéine CFTR

3.2.1 Effet de la p.Val470Met de la protéine CFTR WT et 508del

3.2.1.1 Profil

J'ai pu mettre en évidence par Western blot qu'il n'y avait pas de différence de profil électrophorétique entre les protéines produites par les plasmides porteurs du variant 470V et 470M. Cela est retrouvé à la fois pour les plasmides porteurs ou non de la délétion p.Phe508del (plamides 508del ou WT) (Figure 3.6). En effet, les protéines CFTR provenant des cellules HEK293T exprimant le CFTR-470M-WT et CFTR-470V-WT ont le même profil avec une bande C plus intense que la bande B. Les protéines CFTR des cellules HEK293T exprimant le CFTR-470M-508del et CFTR-470V-508del ont également le même profil avec une seule bande, la bande B. La présence d'une Valine ou d'une Méthionine à la position 470 de la protéine CFTR ne semble pas affecter sa maturation.

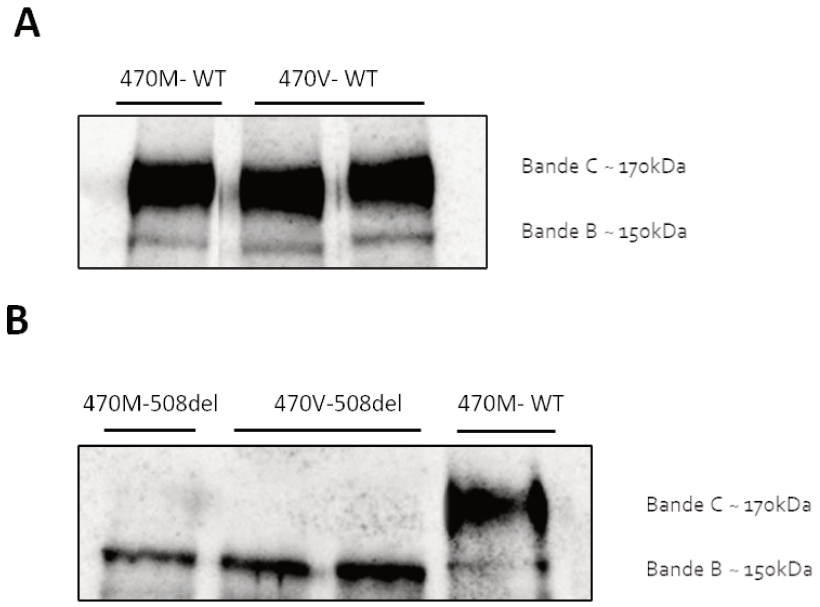


FIGURE 3.6: Analyse par Western blot de l'expression de CFTR dans des cellules HEK293T. A. Cellules HEK293T transfectées exprimant le CFTR-470M-WT ou le CFTR-470V-WT (N=6). B. Cellules HEK293T transfectées exprimant le CFTR-470M-508del ou le CFTR-470V-508del(N=6).

3.2.1.2 Fonction

Pour mesurer la fonction du CFTR, j'ai utilisé dans un premier temps la méthode de FLIPR. Après 30h de transfection, deux lectures de fluorescence sont réalisées : (1) la fluorescence basale, fluorescence de la cellule à son potentiel de repos et (2) la fluorescence après activation. Pour nous assurer que nous mesurions bien la fonction du CFTR, j'ai également mesuré la fluorescence dans des cellules HEK293T non transfectées (NT) et j'ai confirmé l'absence de fonction (rappelons que les cellules HEK293T n'expriment pas le CFTR de manière endogène). J'ai également mesuré la fluorescence des cellules transfectées après inhibition et confirmé que la fonction diminuait bien avec l'ajout d'inhibiteur et cela pour chaque transfection (Figure 3.7).

Après analyse statistique, j'ai pu mettre en évidence qu'il existe une différence significative de fonction du canal CFTR selon que la protéine porte la Valine ou la Méthionine en position 470. Les p-valeurs observées dans le contexte WT et 508del sont respectivement de $3{,}32.10^{-4}$ et de $5{,}57.10^{-9}$. En revanche, l'effet de la valine 470 n'est pas le même selon le variant porté en position 508. Le canal CFTR-WT avec la valine est moins fonctionnel que lorsque la méthionine est présente. Au contraire, c'est la présence de la valine qui confère plus de fonction à la protéine CFTR-508del (rappelons cependant que cette combinaison n'existe pas dans la nature) (Figure 3.7).

Après cette observation j'ai cherché à savoir si l'interaction entre les deux facteurs (WT/508del et M/V) était significative. Une interaction significative décrit une situation dans laquelle deux facteurs n'agissent pas de façon additive sur un troisième. Pour cela j'ai également utilisé un modèle mixte en ajoutant un terme d'interaction. J'ai pu montrer que l'interaction était significative avec une p-valeur de $3{,}63.10^{-7}$ et que la valine 470 avait donc un effet opposé sur la fonction du CFTR-WT et du CFTR-508del (Figure 3.8).

J'ai ensuite cherché à valider nos résultats de FLIPR par la technique la plus utilisée pour connaitre la fonction du CFTR, le Patch Clamp. Les résultats obtenus confirment un effet de l'acide aminé en position 470 sur la fonction du CFTR et que cet effet est différent selon que la phénylalanine en position 508 est présente (WT) ou délétée (508del) (Figure 3.9). Ainsi, la présence d'une Valine en 470 sur un contexte WT rend le canal moins fonctionnel que la Méthionine et, à l'inverse, la Valine 470 semble restaurer un peu de fonction au CFTR-508del. Cette tendance est observable visuellement mais les résultats obtenus ne sont pas statiquement significatifs.

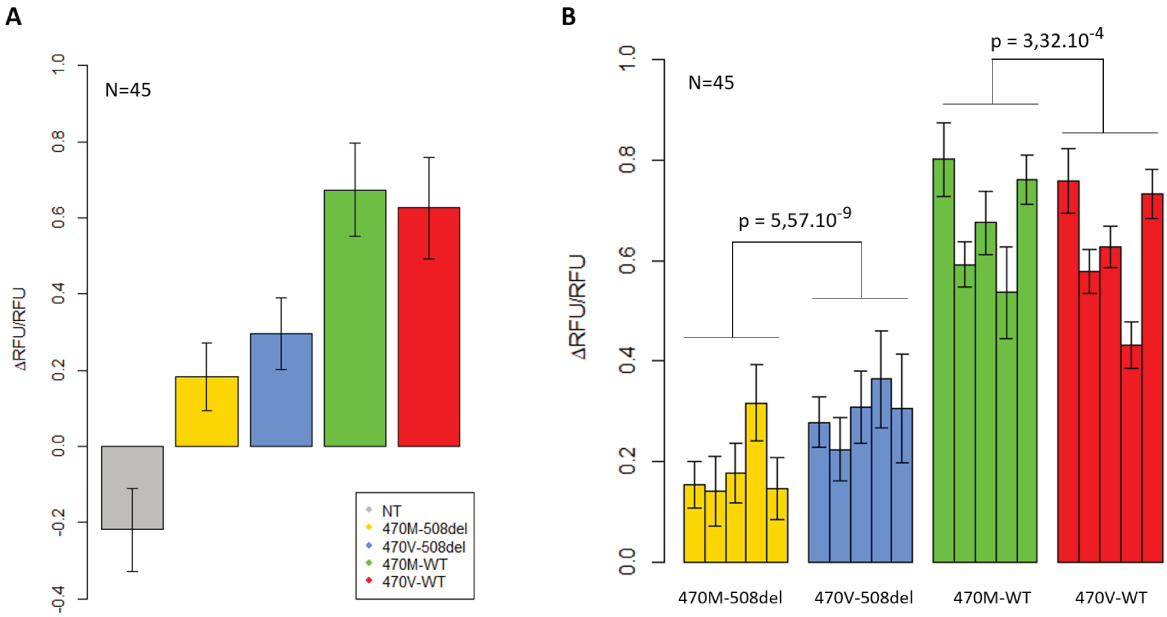


FIGURE 3.7: Représentation graphique des résultats de FLIPR. Différence d'intensité de fluorescence après l'ajout de l'activateur (ΔRFU = RFUactivée - RFUbasale) normalisée par la fluorescence basale (RFU). Le canal CFTR est activé par la FSK et le VX-770 (1µM) dans toutes les transfections. A. Représentation graphique des moyennes de l'ensemble des expériences. B. Représentation graphique des moyennes par expérience (5 expériences)

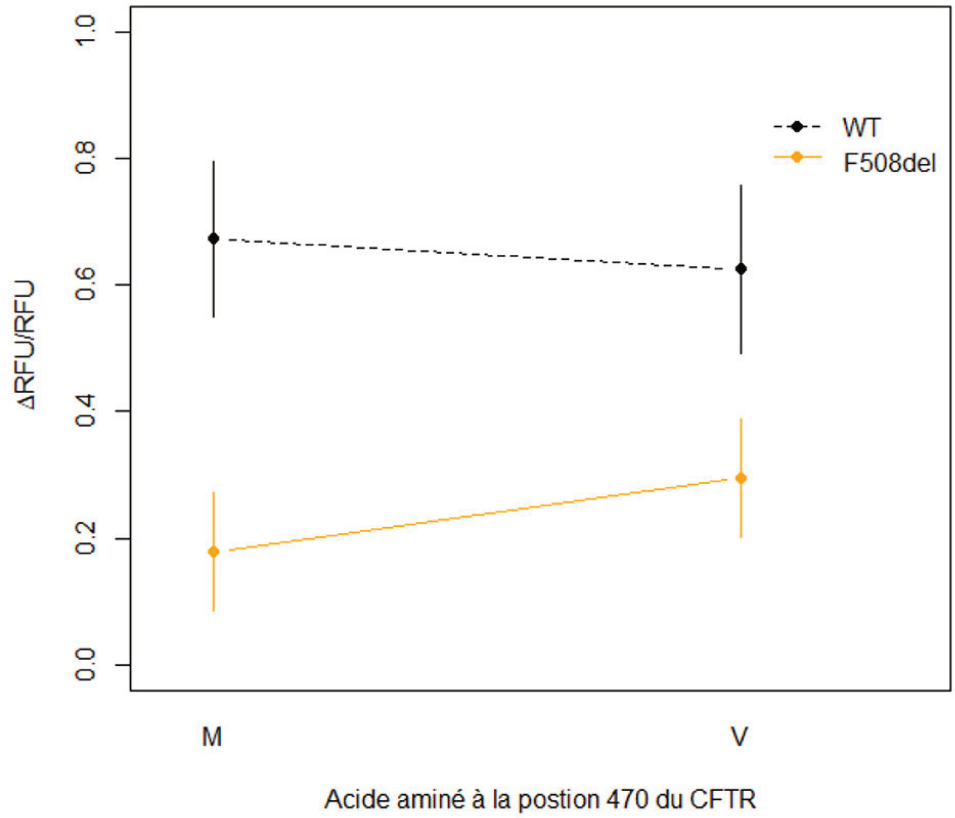


FIGURE 3.8: Graphique d'interaction entre les deux facteurs (WT/508del) et (M/V).

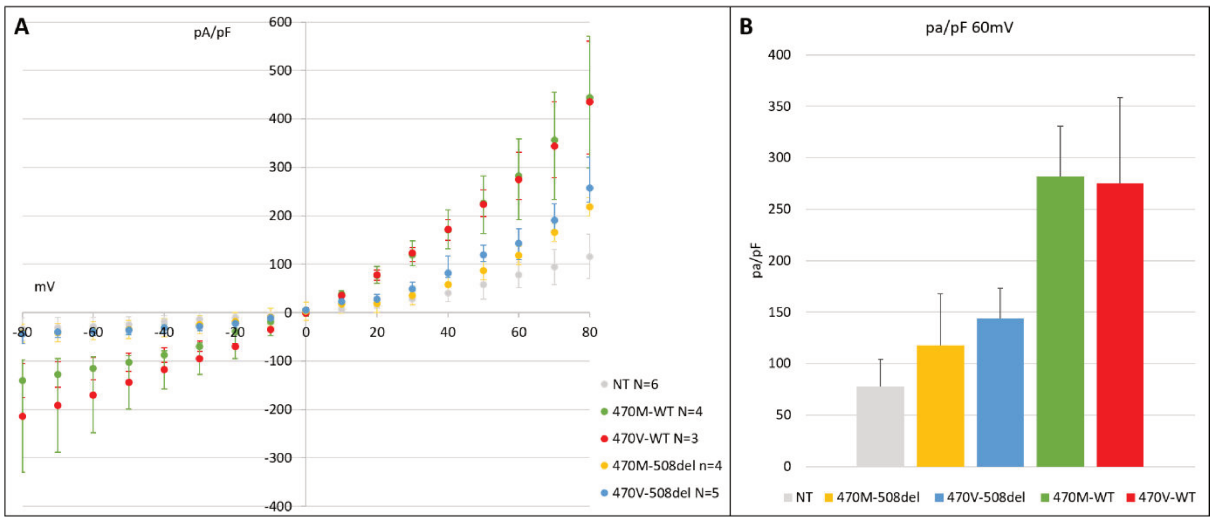


FIGURE 3.9: Résultat du Patch Clamp. L'analyse par Patch-clamp a été effectuée sur des cellules HEK293T NT et dans des cellules HEK293T surexprimant CFTR-470M-WT, CFTR-470V-WT, CFTR-470M-508del ou CFTR-470V-508del. A. Relations courant-tension (A-V) obtenues à partir de courants de cellules entières dans des conditions basales. B. Comparaison des densités de courant à +60mV.

3.2.2 Effet de la p.Ile1027Thr sur la protéine CFTR muté p.Phe508del

3.2.2.1 Profil

J'ai mis en évidence par Western blot qu'il n'y avait pas de différence de profil électrophorétique du CFTR des protéines produites par les cellules WT ou 508del en fonction de la présence ou non du variant p.Ile1027Thr. En effet le profil pour l'allèle complexe CFTR-508del-1027T est semblable à celui de la délétion 508del seule avec une bande B plus intense que pour un CFTR-WT sans bande C (figure 3.10). Il n'y a pas non plus de changement de profil de la protéine muté 1027T dans un contexte WT par rapport à un CFTR avec une Isoleucine à la position 1027. Il semblerait donc qu'une mutation à cette position ne modifie pas l'expression de la protéine.

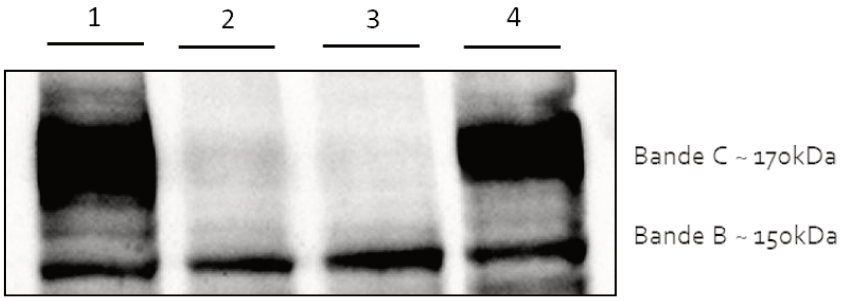


FIGURE 3.10: Analyse par Western blot de l'expression de la protéine CFTR dans des cellules HEK293T. Cellules HEK293T transfectées exprimant les constructions suivantes : piste 1 : CFTR-470M-WT, piste 2 : CFTR-470M-508del, piste 3 : CFTR-470M-508del-1027T, piste 4 : CFTR-470M-WT-1027T (N=4).

3.2.2.2 Localisation

Les expériences d'immunofluorescence, pour les mutants p.Phe508del et p.Phe508del-p.Ile1027Thr, semblent montrer une localisation du CFTR semblable à la localisation de la calnexine, marqueur du RE (Figure 3.11). Un double marquage aurait pu confirmer leur colocalisation dans le RE. La localisation cellulaire de la protéine CFTR n'est donc pas différente selon l'acide aminé en position 1027 dans un contexte 508del : la protéine est retenue dans le RE et très peu de protéine CFTR est exprimé dans la membrane plasmique des cellules.

3.2.2.3 Fonction

Par la technique de FLIPR j'ai pu mettre en évidence une différence significative de fonction du canal CFTR entre les cellules portant ou non la 1027T dans un contexte 508del. La p-valeur observée est de $1{,}19.10^{-7}$. La Thréonine en position 1027 réduirait la fonction du canal par rapport à une Isoleucine (Figure 3.12). Ceci pourrait s'expliquer par la position de l'acide aminé 1027 sur la protéine CFTR. En effet si l'on regarde la structure 3D du CFTR, la position 1027 se situe au niveau du canal sur la 10ième hélice transmembranaire dans le domaine MSD2 (Figure 3.13). La 10ième hélice transmembranaire participe à la formation du canal et est directement liée à la boucle intracellulaire 4 permettant l'interaction entre les domaines MSD2 et NBD1 [Hwang et al., 2018].

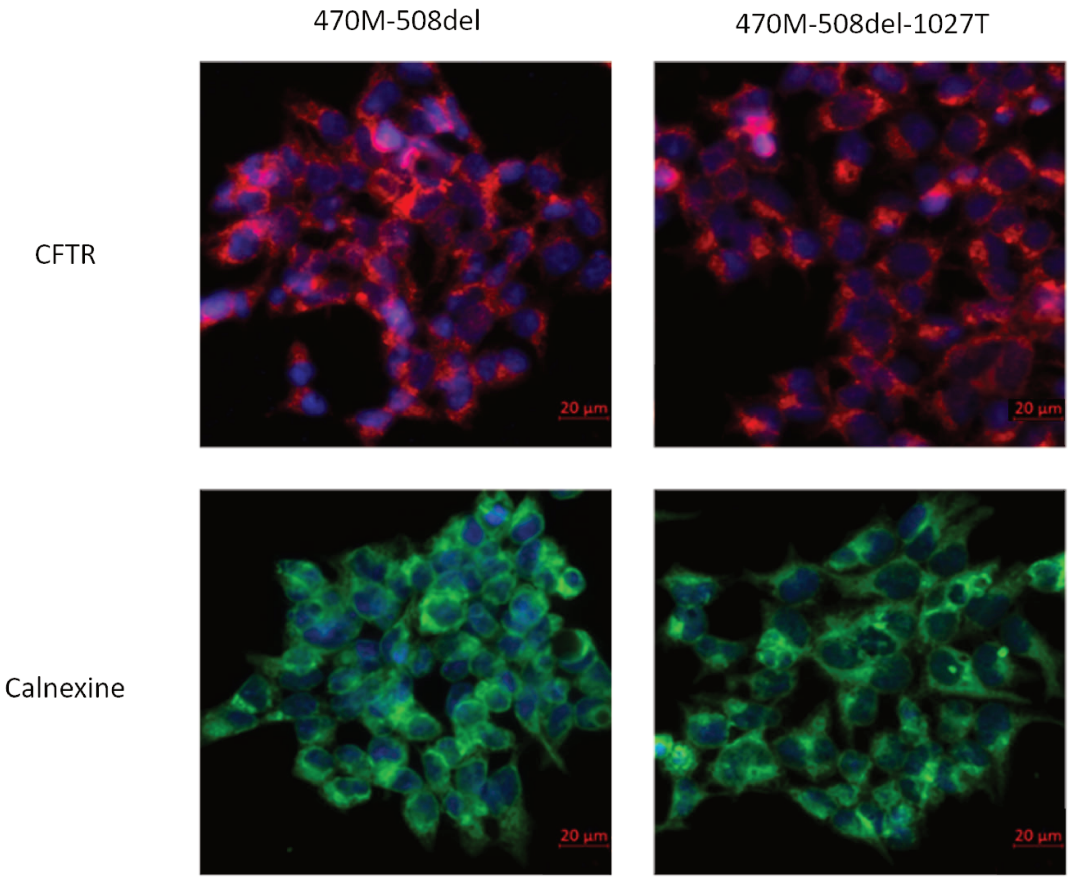


FIGURE 3.11: Analyse par Immunofluorescence de la localisation du CFTR (rouge) et de la calnexine (vert) dans des cellules HEK293T exprimant le CFTR-470M-508del et CFTR-470M-508del-1027T

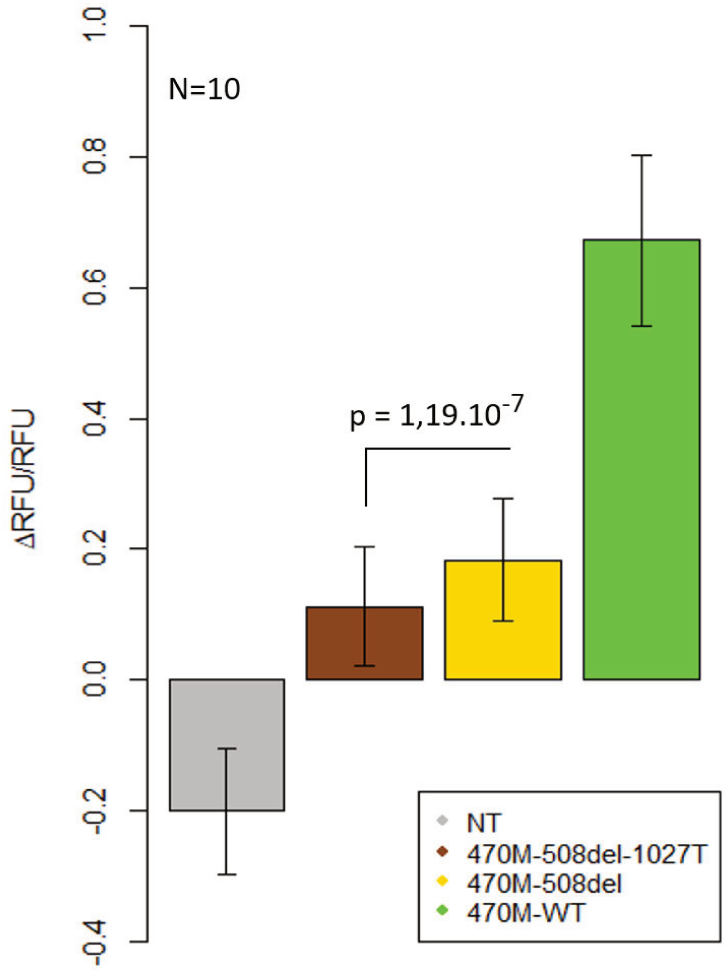


FIGURE 3.12: Représentation graphique des résultats de FLIPR. Différence d'intensité de fluorescence après l'ajout de l'activateur (ΔRFU = RFUactivée - RFUbasale) normalisée par la fluorescence basale (RFU). Le canal CFTR est activé par la FSK et le VX-770 (1µM) dans toutes les transfections.

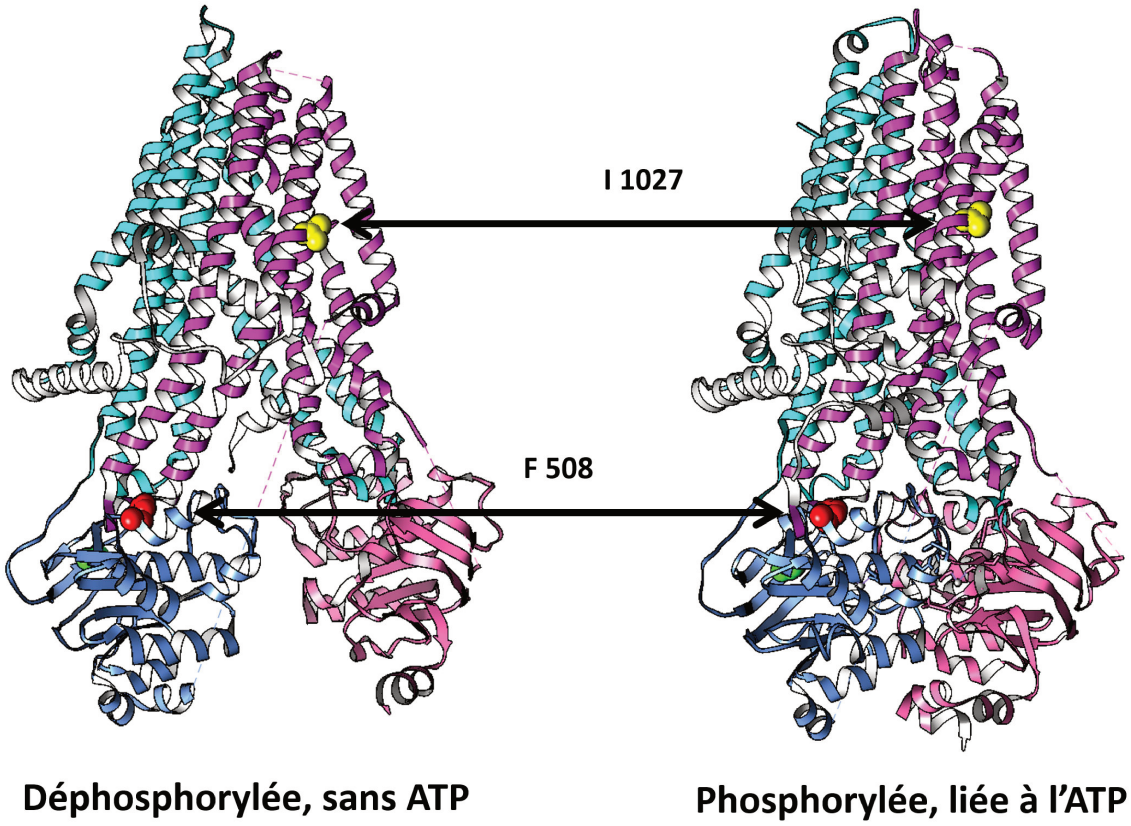


FIGURE 3.13: Représentation 3D de la protéine CFTR. Les positions 508 et 1027 de la protéine sont représentées sous forme de sphères. A gauche représentation 3D de la protéine CFTR déphosphorylée, sans ATP et à droite la protéine CFTR phosphorylée, liée à l'ATP. Représentation obtenue avec l'outil VMD [Humphrey et al., 1996] à partir des structures pdb 5UAK et 6MSM [Liu et al., 2017, Zhang et al., 2018].

CHAPITRE 4

DISCUSSION-PERSPECTIVES

Pour ce projet nous avons combiné des approches *in silico* et *in vitro*. Tout d'abord nous avons mis au point GEMPROT, un outil innovant de visualisation des variations, puis nous avons étudié sur un modèle cellulaire, l'impact sur la fonction de la protéine CFTR de deux variants retrouvés en *cis* avec la délétion de la phénylalanine en position 508.

GEMPROT permet d'obtenir une visualisation globale de la séquence en acides aminés des protéines codées par chaque haplotype d'un individu, en mettant en évidence les domaines fonctionnels de la protéine lorsqu'ils sont connus. Il est ainsi possible de voir directement comment des variants exoniques impactent une séquence protéique et s'ils touchent ses régions fonctionnelles.

La méthode de traduction de l'intégralité de la séquence nucléotidique que j'ai développée dans cet outil permet d'éviter des erreurs d'annotation, notamment pour les variants intervenant sur un même codon que Khan et al. [Khan et al., 2018] désignent par l'acronyme pcSNV pour paired codon SNV. Sur les données du projet FrEx, 114 pcSNV ont ainsi été mis en évidence chez 194 individus [Khan et al., 2018]. En utilisant l'outil de prédiction d'effet SIFT, neuf pcSNVs étaient prédits « damaging » alors que les deux SNV initiaux étaient annotés « tolerated ». Inversement, deux pcSNVs sont annotés « tolerated » ou « neutral », alors qu'ils provenaient de deux SNVs initialement « damaging ». Les pcSNVs sont donc une source de variants potentiellement délétères, qu'il est important de prendre en compte.

GEMPROT permet également la comparaison de distributions d'haplotypes entre différents sous-groupes de populations. Ainsi, il est possible d'identifier des haplotypes différentiellement répartis géographiquement comme nous l'avons vu pour certains haplotypes de la population de l'Asie de l'Est de 1000 Genomes dans le gène *CFTR*. Cela peut également être transposé à deux sous-groupes avec dans le premier groupe des individus malades (cas) et, dans le second groupe, des témoins (personnes n'ayant pas développé cette maladie). En comparant les distributions entre les cas et les témoins, il est possible, d'identifier des combinaisons de variants pouvant expliquer la maladie qui auraient pu être éliminées lors des premières étapes de filtre. En

effet, les variants fréquents en population sont généralement retirés de l'analyse, voire même bien souvent du fichier VCF. Ces variants, plus fréquents en population que ceux généralement retenus dans l'étude des maladies, peuvent cependant, lorsqu'ils sont retrouvés en combinaison avec un autre variant, devenir des bons candidats pour expliquer la maladie.

Différentes extensions de GEMPROT peuvent être envisagées. Une première extension serait la prise en compte du phénomène d'échappement au NMD. Ce processus de défense des cellules fait que des transcrits, trop courts du fait de l'existence d'un codon stop prématuré, sont détruits et ne produisent pas de protéine. Des études ont montré que la position d'apparition du codon stop prématuré, au sein de l'ARNm, détermine s'il est pris en charge ou non par le NMD [Nagy and Maquat, 1998]. Ainsi, lorsque le codon stop est situé à moins de 50 paires de bases de la dernière jonction exon-exon, l'ARNm peut échapper au NMD et donner une protéine tronquée. Lorsque c'est une mutation ponctuelle qui engendre un codon stop, il est facile d'en déduire sa position et de dire si oui ou non l'ARNm est pris en charge par le NMD. En revanche, si c'est une insertion/délétion située en amont dans le gène qui entraîne un codon stop, il est beaucoup plus difficile de déterminer où apparaitra le prochain codon stop.

Coban-Akdemir et al. ont développé un outil qui permet, en donnant la position du « frameshift », de déterminer si oui ou non il peut engendrer une protéine tronquée [Coban-Akdemir et al., 2018]. Cet outil en ligne nécessite cependant de travailler variant par variant. De manière très intéressante, GEMPROT permet également d'obtenir cette information qui pourrait facilement être ajoutée dans le compte rendu répertoriant les variants présents chez un individu. Cette information est importante car l'impact pour la cellule peut être très différent selon que l'ARNm est pris en charge ou non par le NMD. Si l'ARNm est bien pris en charge par le NMD, il en résulte une réduction de la quantité de protéine. En revanche s'il n'est pas pris en charge par le système de défense, une protéine tronquée et potentiellement mutée est traduite qui peut conduire à des maladies de gain de fonction comme décrit par exemple dans le syndrome de Robinow [White et al., 2015, White et al., 2016, Poli et al., 2018].

Un des aspects qui a été fréquemment soulevé durant mes travaux de thèse est la structure 3D de la protéine qui est un élément très important, surtout lorsque l'on traite des combinaisons de variants. La structure que prend la protéine lorsque qu'elle est repliée peut en effet réunir deux variants qui sont éloignés sur la séquence protéique. Il existe plusieurs outils de visualisation de structure 3D [Glusman et al., 2017]. Nous n'avons, pour le moment, essayé que quelques outils de visualisation 3D dont l'outil MuPIT (pour Mutation Position Imaging Toolbox). Le pipeline de MuPIT aligne les

séquences de protéines des structures PDB sur les séquences de protéines humaines disponibles dans UniProtKB [UniProt Consortium, 2019]. Les séquences de protéines UniProtKB sont alignées sur les séquences de transcription de l'ARNm RefSeq et les séquences de transcription de l'ARNm sont alignées sur l'ADN génomique humain [Niknafs et al., 2013]. Cet outil MuPIT est proposé sous forme d'un navigateur web qui localise ainsi automatiquement les SNV sur les coordonnées des structures protéiques tridimensionnelles disponibles. Il permet aussi de localiser les éléments de fonctions et les molécules qui entrent en interactions avec la protéine. Par exemple, pour la protéine CFTR, il est possible de localiser où se fixent les deux ATP permettant l'ouverture du canal (Figure 4.1). Dans son état actuel, cet outil ne nous permet pas de fournir nos propres fichiers de structure (fichier pdb). Cet outil MuPIT devrait pouvoir assez facilement être raccordé à GEMPROT moyennant quelques développements informatiques supplémentaires que nous n'avons cependant pas eu le temps de réaliser au cours de cette thèse.

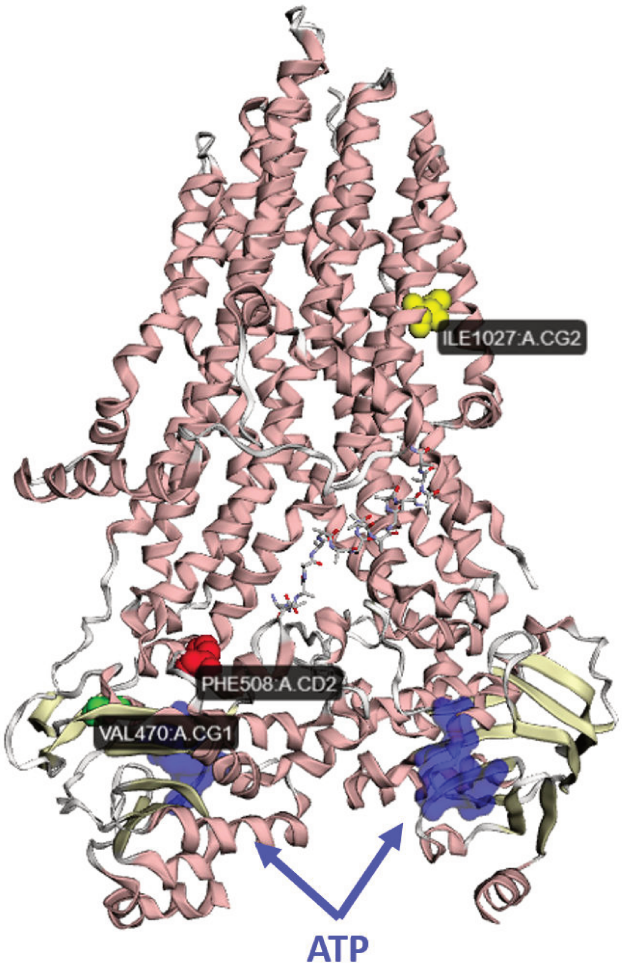


FIGURE 4.1: Représentation de la structure 3D du canal CFTR fermé, issus du serveur web MuPIT (pdb :5UAK [Liu et al., 2017]). Les trois acides aminés touchés par les variants étudiés 470V (vert), 508F (rouge) et 1027I (jaune) sont représentés sous forme de sphère. Les ATP sont localisés au niveau de leurs sites de liaisons aux deux domaines NBD.

Pour la visualisation de la protéine, nous avons fait le choix dans GEMPROT de nous affranchir des variants présents dans les zones non-codantes du génome c'est-à-dire dans les introns ou en dehors des gènes. Cependant, la plupart des variants trouvés lors du séquençage d'un génome se situent dans ces zones non codantes. Cela peut s'expliquer en grande partie par le fait que les régions codantes ne représentent qu'environ 1% du génome [Xuan et al., 2013]. Les régions non codantes sont à la fois plus nombreuses mais également moins soumises à la pression de sélection par l'effet moins délétère de leurs variants sur la fonction des gènes. Très longtemps délaissées au profit des exons, ces régions non codantes font aujourd'hui l'objet de nombreuses études car elles contiennent des séquences régulatrices de l'expression des gènes qui peuvent se révéler indispensables au bon fonctionnement des cellules. La prise en compte de l'effet des variants présents dans les régions régulatrices sur la production et la fonction des protéines serait très intéressante à étudier dans ce contexte mais il s'agit de problèmes encore assez mal compris que nous n'avons donc pas pu aborder dans notre travail.

De même, on sait que certains variants situés dans les introns et les exons peuvent modifier l'épissage et donc changer la séquence protéique [Dujardin et al., 2011]. Un exemple très connu dans le gène *CFTR* est la présence d'une séquence polypyrimidique (TGmTn) dans l'intron 9 au niveau du site d'épissage [Chu et al., 1993, Cuppens et al., 1998, Claustres, 2005]. En fonction du nombre de répétition de Thymines (Tn), dans cette séquence intronique, un épissage alternatif du gène *CFTR* se produit, entraînant un saut de l'exon 10. Il en résulte un pourcentage d'ARNm du CFTR délétés de l'exon 10 inversement corrélé au nombre de Thymines : moins il y a de Thymines, plus il y a de CFTR sans exon 10. L'excision de l'exon 10 entraîne la production d'une protéine CFTR non fonctionnelle. Ces différences de séquence dans l'intron 9 seraient associées à la survenue de CBAVD [Claustres, 2005]. Nous avons d'ailleurs été confrontés à ce phénomène d'épissage alternatif lors des expériences de séquençage des lignées cellulaires présentes au laboratoire (Annexe IV).

Nos connaissances dans le domaine de l'épissage alternatif et de ses mécanismes de contrôle sont encore très incomplètes et il est difficile de prédire l'épissage. Nous avons donc fait le choix, dans notre travail, de partir d'un transcrit déjà épissé et de laisser à l'utilisateur le choix de ce transcrit pour nous affranchir du problème de l'épissage. Il pourrait cependant être intéressant de tenir compte dans GEMPROT des variants impliqués dans l'épissage. Pour cela, il serait nécessaire d'utiliser des outils de prédiction de sites modifiant l'épissage mais également, après avoir détecté les sites, de pouvoir déterminer le transcrit qui en résulte.

Lors de mon travail expérimental, j'ai caractérisé l'impact de deux combinaisons de variants sur l'expression et la fonction du canal CFTR. Le variant 470V ne modifie

pas l'expression du CFTR. En effet, lorsqu'on analyse les résultats de Western Blot, le profil électrophorétique des protéines WT et 508del ne change pas entre une construction portant la 470V et une construction portant la 470M. En revanche, les résultats de fonction par la technique FLIPR permettent d'identifier une différence entre les deux constructions. Nous n'avons cependant pas pu confirmer ce résultat par la technique du patch-clamp qui n'a pas montré de différences significatives. Ces résultats peuvent peut-être s'expliquer par le fait que nous travaillons sur des transfections transitoires. En effet, lorsqu'on utilise des transfections transitoires, un mélange de cellules transfectées et non transfectées est présent dans la culture cellulaire. Pour déterminer la fonction, surtout en patch clamp, nous ne sommes pas à l'abri de mesurer le courant d'une cellule non transfectée. C'est pour cette raison que nous avons commencé nos analyses par la technique de FLIPR. Nous avons obtenu plusieurs mesures par culture cellulaire permettant de faire une moyenne de fluorescence par culture de cellules transfectées. Ainsi, par cette méthode de criblage il a été possible de s'affranchir des cellules non transfectées. La protéine WT portant la Valine semble moins fonctionnelle que la protéine portant la Méthionine. Cela conforte les résultats obtenus par Cuppens et al. [Cuppens et al., 1998] qui montre une différence de fonction dans les cellules COS transfectées avec le CFTR portant la 470V ou la 470M, en faveur de la Méthionine (Figure 4.2).

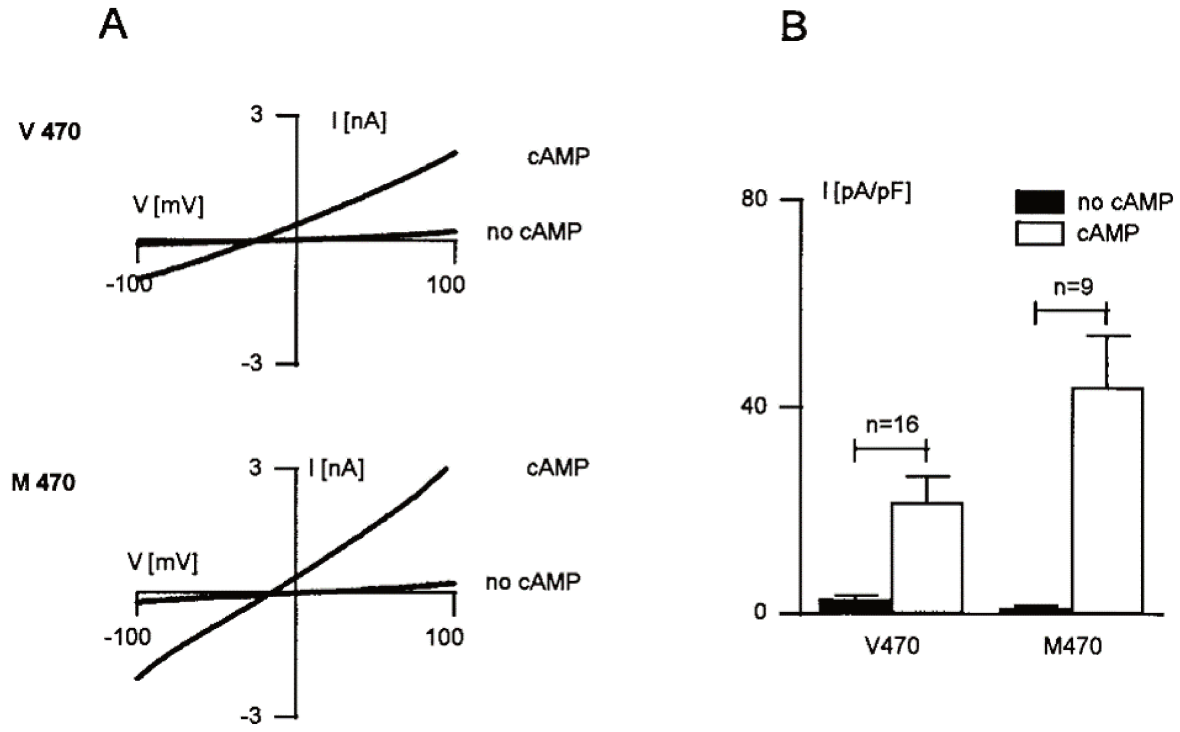


FIGURE 4.2: Courants de cellules entières exprimant de manière transitoire le CFTR portant la 470M ou 470V. (A) Relations courant-tension (I-V) obtenues à partir de courants de cellules entières COS dans des conditions basales et après activation par l'AMPc. (B) Comparaison des densités de courant (avec ou sans activation par l'AMPc) à +100 mV (d'après Cuppens et al 1998).

Par nos analyses FLIPR, nous avons pu montrer que pour un haplotype portant la 508del, la différence de fonction entre la 470M et la 470V est également significative mais de manière opposée. Les résultats de fluorescence montrent que la Valine confère plus de fonction que la Méthionine. Les positions 470 et 508 sont toutes les deux situées dans le domaine NBD1. Comme nous l'avons vu précédemment, la mutation 508del déstabilise l'ensemble de la conformation du CFTR [Lewis et al., 2005, Du et al., 2005, Serohijos et al., 2008, Mornon et al., 2008, Mendoza et al., 2012]. Nous pouvons donc émettre l'hypothèse que la 470V confère une conformation plus stable à la protéine 508del, que la 470M, permettant une amélioration de la fonction. A l'inverse, la 470V peut légèrement modifier la conformation de la protéine WT réduisant sa fonction. En parallèle nous avons également regardé l'effet de La Valine ou de la Méthionine sur un haplotype portant la p.Gly551Asp (551D) sur la fonction du CFTR. Les résultats de Western Blot et de FLIPR n'ont pas montré de différence de profil électrophorétique et de fonction de la protéine CFTR (Annexe V).

La méthode classique pour mesurer l'expression fonctionnelle de la protéine CFTR est la méthode de Patch Clamp. Cette méthode est précise pour mesurer la fonction du CFTR de différents mutants. Les mesures prises par cette méthode se font sur cellule unique. Il est donc nécessaire, dans notre cas, de la reproduire un nombre de fois assez conséquent afin d'avoir une puissance statistique suffisante pour pouvoir faire une estimation de la fonction du mutant. Nous n'avons pas pu réaliser suffisamment de réplicats lors de notre travail de thèse et il faudrait réaliser de nouvelles mesures par la méthode de Patch Clamp pour s'assurer de cette différence de fonction d'une protéine CFTR mutée F508del selon l'acide aminé présent en position 470.

D'un point de vue clinique, ces résultats ne présentent pas grand intérêt pour les patients porteurs de la délétion 508 car aucun d'eux ne porte la 470V. Cependant lorsqu'il s'agit de l'analyse de fonction, par exemple dans le but de déterminer *in vitro* la réponse à un traitement, il est important de réaliser les analyses sur le bon fond génétique. La question se pose dans l'exemple qui nous intéresse car la Valine 470 correspond à l'allèle de référence dans le génome humain et lors de la conception de plasmide, c'est la séquence de référence qui est la plus utilisée. À la suite de l'observation de ce déséquilibre de liaison complet entre la 508del et la 470M dans les populations FrEx et 1000G, j'ai séquencé l'ADNc CFTR exprimé dans les différents types cellulaires utilisés dans mon laboratoire, dans le cadre de l'étude de la mucoviscidose (Annexe IV). Tous les plasmides portant la délétion F508del étaient bien porteurs du variant conduisant à une Méthionine en position 470. Ces plasmides sont utilisés depuis de très nombreuses années au laboratoire et ont été construits à partir de l'ADNc issu de cellules de patient 508del immortalisées. Il semble que

maintenant la position 470 soit prise en compte de manière plus systématique dans les tests fonctionnels. Par exemple dans l'étude de Han et al. les plasmides ont été modifiés pour inclure le 470M pour les variants connus pour le porter en *cis* [Han et al., 2018].

La position variante 1027 ne modifie pas le profil électrophorétique de la protéine CFTR que ce soit sur un fond WT ou 508del. En effet, les résultats montrent que la protéine CFTR-470M-508del-1027T a le même profil électrophorétique que la protéine CFTR-470M-508del. Nous avons ainsi en Western Blot, dans les deux cas, une bande B et une absence de bande C. En immunofluorescence, il semble également que la localisation des protéines, portant ou non la 1027T, soit la même. La protéine est localisée dans le RE. Cela confirme les résultats obtenus par Baatallah et al. qui ont évalué l'impact de différentes combinaisons de variants portant la 508del sur l'efficacité du traitement combinant le correcteur VX-809 et le potentialisateur VX-770 (Orkambi) [Baatallah et al., 2018]. Ils n'ont pas trouvé de différence de profil électrophorétique, entre la 508del seule et en présence de la 1027T, en Western Blot, avec le ratio des bandes C/(B+C) ($\sim$20%) (Figure 4.3.A).

Ils n'ont pas trouvé de différence significative de la fonction du canal CFTR entre les deux constructions (Figure 4.3.C). Nos résultats de FLIPR semblent cependant indiquer que la thréonine en position 1027 réduit la fonction du canal CFTR lorsqu'elle est associée à la 508del. Ces différences peuvent peut-être s'expliquer par l'allèle présent en position 470. En effet, Baatallah et al. ont utilisé l'ADNc du CFTR présent dans une construction procaryote fournit par Transgene (pTG5960) qui semble présenter le polymorphisme 470V [Vankeerberghen et al., 1998]. S'ils n'ont pas pris en compte que la 508del est toujours associée à la 470M alors les résultats de cette étude peuvent ne pas être le reflet de l'effet de la mutation sur la protéine dans son contexte physiologique. Il serait intéressant de vérifier quel est l'acide aminé à la position 470. Il serait également intéressant que nous puissions comparer l'effet de la mutation I1027T sur une haplotype 470V-508del ce que nous n'avons pas encore fait.

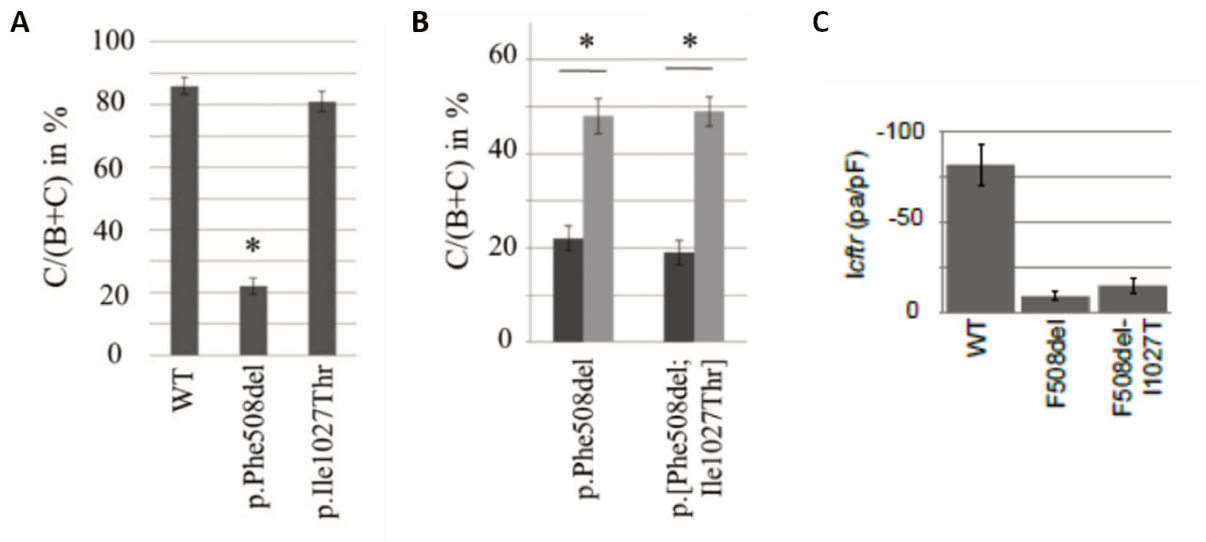


FIGURE 4.3: Quantification de l'efficacité de maturation du CFTR et mesure de la fonction de la protéine CFTR obtenues par Baatallah et al. A. Quantification de l'efficacité de maturation du CFTR avec le ratio C/(B+C) B. Quantification de l'efficacité de maturation du CFTR avant (noir) et après traitement correcteur (gris). C. Mesures de Patch exprimées en courants de chlorure CFTR en pA/pF mesurés à -60 mV à partir de cellules HEK293 transfectées de manière transitoire avec les différents plasmides. Les courants ont été mesurés à -82 ± 23 pA/pF (n = 6) pour CFTR-WT, -9 ± 5 pA/pF (n = 4) pour CFTR-Phe508del et -15 ± 8 pA/pF (n = 3) pour CFTR-p.[Phe508del ; Ile1027Th]. (Adapté de Baatallah et al)

Des expériences de FLIPR avec l'ajout de correcteur VX-809 ont été entreprises. L'ajout de correcteur lors de la culture cellulaire améliore la maturation et la fonction du canal CFTR-508del. Nous pourrons ainsi déterminer s'il existe des différences de réponse au VX-809 en fonction des combinaisons de variants étudiées. Cela va permettre d'appuyer nos propos quant à l'importance de prendre en compte le fond génétique lors de la conception de médicaments dirigés vers une protéine. Des études sur l'action de certains médicaments ont déjà montré l'importance du contexte haplotypique. Un variant présent dans une protéine cible peut entraîner des différences de réponses au traitement. C'est par exemple le cas pour la molécule MEDI-2843 qui module la réponse immunitaire en ciblant le récepteur TLR4. La séquence de référence de la protéine TLR4 a été choisie pour la conception de la molécule. Cependant, la réponse au traitement n'a été que partielle chez certains individus qui présentent tous un variant 299D>G non présent sur la protéine de référence. Ce variant est responsable de la perte d'activité de la molécule. Dans le panel de 1000 Genomes, 12,1% des individus ont au moins une copie de ce variant 299D>G (Figure 4.4). La prise en compte de ces informations en amont aurait pu permettre une meilleure conception des traitements. Par ailleurs, des études récentes ont montré que pour un gène sur sept, l'haplotype le plus commun n'est pas celui qui est présent dans la séquence de référence humaine [Spooner et al., 2018].

Le manque de connaissance de l'association de variants rares peut aussi entraîner

des erreurs de génotypage [Claustres et al., 2004]. Plusieurs erreurs ont d'ailleurs été mises en évidence dans le registre clinique de la mucoviscidose [Claustres et al., 2017]. Ce résultat est en partie dû à la façon dont est réalisée la recherche des variants lors du diagnostic en se focalisant sur seulement une trentaine de variants (les plus courants) et en ne séquençant donc pas le gène dès lors que deux mutations « causales » sont trouvées. Il est fort probable que ce manque de prise en compte des combinaisons de variants en *cis* soit généralisable à de nombreuses maladies génétiques. En effet une étude a montré qu'environ 60% des gènes mutés dans les autosomes ont leurs variants présents en *cis*, et environ 40% en *trans* [Hoehe et al., 2019].

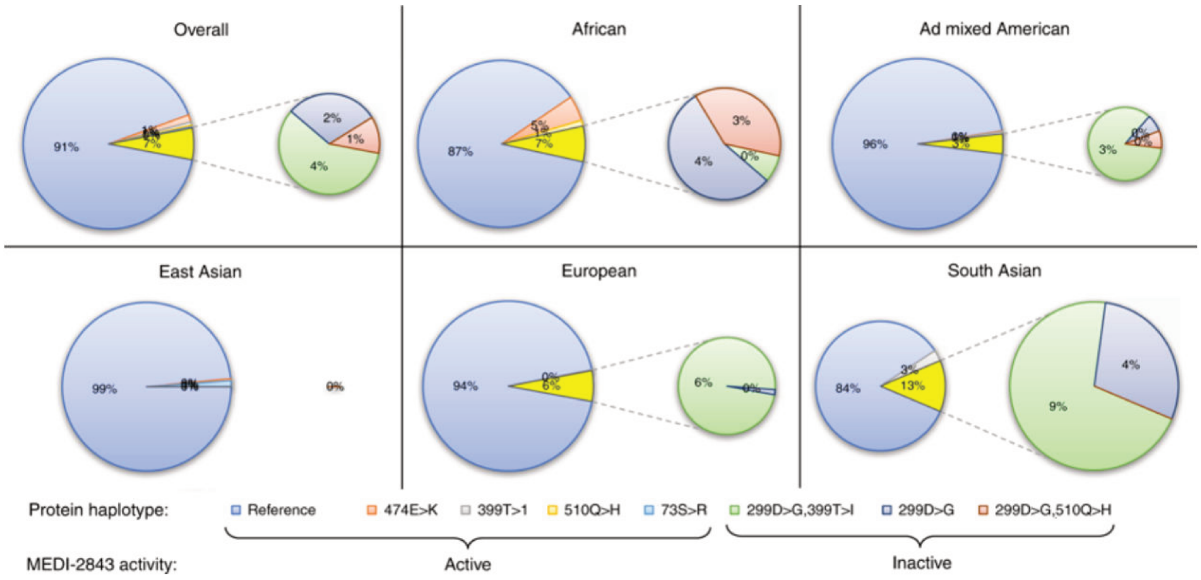


FIGURE 4.4: Fréquence des haplotypes de la protéine TLR4 des différentes sous-populations de 1000 Genomes. Les haplotypes [299D>G,399T>I], 299D>G, 299D>G, 510Q>H qui portent 299D>G et entraînent la perte d'activité du MEDI-2843 sont présents à droite.

Comme nous l'avons vu précédemment, des variants en *cis*, en dehors des régions exoniques peuvent avoir des conséquences sur la fonction du CFTR et moduler le phénotype et/ou la gravité clinique [Chu et al., 1993, Cuppens et al., 1998, Claustres, 2005]. Ces éléments, appelés cis-régulateurs, peuvent être localisés également dans des régions éloignées du CFTR [Yang et al., 2016, Moisan et al., 2016, Vecchio-Pagán et al., 2016]. Le phénotype, des patients atteints de la mucoviscidose ayant un même génotype, peut également être influencé par des gènes modificateurs [Zielenski et al., 1999, Drumm, 2001, Castaldo et al., 2001, Salvatore et al., 2002, Mekus et al., 2003] Des protéines ont été étudiées et reconnues comme ayant un effet modulateur sur le phénotype. On peut citer parmi ces protéines : des protéines associées à la défense antimicrobienne (MBL, MBL2) [Garred et al., 1999, Davies et al., 2000, Gabolde et al., 2001]. , des protéines associées à la réponse inflammatoire (TGF-β, ACE, IL-10)

[Arkwright et al., 2000, Arkwright et al., 2003]. Une interaction directe ou indirecte avec la protéine CFTR a pu être mise en évidence, suggérant un possible rôle modulateur de certaines protéines sur le phénotype. La protéine Annexine A5 (ANXA5), par exemple, a une interaction directe avec le domaine NBD1 du CFTR. Sa surexpression dans les cellules permet une augmentation de la présence à la membrane et de la fonction du canal CFTR [Trouvé et al., 2007, Le Drévo et al., 2008]. GEMPROT pourrait être un outil utile pour identifier des variants dans les gènes codant pour ces protéines. Il se pourrait que certains de ces variants expliquent les différences qui existent entre deux patients ayant un même génotype.

CHAPITRE 5

CONCLUSION

Au cours de ce travail, j'ai réalisé un programme qui est prêt à l'utilisation, j'ai écrit une documentation complète de ce programme et une note le décrivant publiée dans la revue Bioinformatics. J'ai par ailleurs présenté mes travaux sous la forme de posters à l'occasion des 9èmes Assises de Génétique Humaine et Médicale et au 19e colloque français des jeunes chercheurs « Vaincre la mucoviscidose » et en présentation orale aux Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM). J'ai également pu commencer à mettre en évidence l'intérêt d'un tel outil dans le cadre des maladies génétiques en réalisant un travail expérimental sur la protéine CFTR impliquée dans la mucoviscidose.

En mettant en évidence une différence de fonction de la protéine CFTR-508Del selon les variants portés en *cis*, nous avons montré l'intérêt de prendre en compte l'ensemble des mutations sur le gène. Des combinaisons de variants pourraient également expliquer une partie de la variabilité de réponses des patients aux traitements. Il est donc important, lors de la conception de molécules thérapeutiques, de connaître tous les disfonctionnements résultants de variations génétiques pour mieux adapter leurs actions.

Notre outil GEMPROT est facile d'utilisation et peut être appliqué sur tous les gènes. Lorsqu'aucun variant dans le (ou les) gène(s) connu(s) dans une pathologie n'a pu être mis en évidence, il pourrait être utile au diagnostic pour explorer des hypothèses plus complexes comme l'impact de combinaison de variants situés en *cis* ou de variants gain de fonction dans ces gènes qui pourraient conduire à une protéine tronquée. Ces hypothèses sont difficiles à explorer en pratique faute d'outils adaptés et nous pensons que GEMPROT peut apporter une première solution pour une prise en compte plus systématique de ces situations complexes.

BIBLIOGRAPHIE

[1000 Genomes Project Consortium et al., 2010] 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319) :1061–1073.

[1000 Genomes Project Consortium et al., 2015] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571) :68–74.

[Adzhubei et al., 2013] Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, 0 7 :Unit7.20.

[Ahmadi et al., 2017] Ahmadi, S., Bozoky, Z., Di Paola, M., Xia, S., Li, C., Wong, A. P., Wellhauser, L., Molinski, S. V., Ip, W., Ouyang, H., Avolio, J., Forman-Kay, J. D., Ratjen, F., Hirota, J. A., Rommens, J., Rossant, J., Gonska, T., Moraes, T. J., and Bear, C. E. (2017). Phenotypic profiling of CFTR modulators in patient-derived respiratory epithelia. *NPJ Genomic Medicine*, 2.

[Altmann et al., 2012] Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E. B., and Müller-Myhsok, B. (2012). A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Human Genetics*, 131(10) :1541–1554.

[Andersen, 1938] Andersen, D. H. (1938). Cystic fibrosis of the pancreas and its relation to celiac disease. A clinical and pathologic study. *American Journal of Diseases of Children*, 56 :344–396.

[Anderson et al., 1991] Anderson, M. P., Gregory, R. J., Thompson, S., Souza, D. W., Paul, S., Mulligan, R. C., Smith, A. E., and Welsh, M. J. (1991). Demonstration that CFTR is a chloride channel by alteration of its anion selectivity. *Science*, 253(5016) :202–205.

[Anna and Monika, 2018] Anna, A. and Monika, G. (2018). Splicing mutations in human genetic disorders : examples, detection, and confirmation. *Journal of Applied Genetics*, 59(3) :253–268.

[Arkwright et al., 2000] Arkwright, P., Laurie, S., Super, M., Pravica, V., Schwarz, M., Webb, A., and Hutchinson, I. (2000). TGF-1 genotype and accelerated decline in lung function of patients with cystic fibrosis. *Thorax*, 55(6) :459–462.

[Arkwright et al., 2003] Arkwright, P. D., Pravica, V., Geraghty, P. J., Super, M., Webb, A. K., Schwarz, M., and Hutchinson, I. V. (2003). End-Organ Dysfunction in Cystic Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 167(3) :384–389.

[Baatallah et al., 2018] Baatallah, N., Bitam, S., Martin, N., Servel, N., Costes, B., Mekki, C., Chevalier, B., Pranke, I., Simonin, J., Girodon, E., Hoffmann, B., Mornon, J.-P., Callebaut, I., Sermet-Gaudelus, I., Fanen, P., Edelman, A., and Hinzpeter, A. (2018). Cis variants identified in F508del complex alleles modulate CFTR channel rescue by small molecules. *Human Mutation*, 39(4) :506–514.

[Bamshad et al., 2011] Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews. Genetics*, 12(11) :745–755.

[Basso et al., 2003] Basso, C., Vergani, P., Nairn, A. C., and Gadsby, D. C. (2003). Prolonged Nonhydrolytic Interaction of Nucleotide with CFTR's NH2-terminal Nucleotide Binding Domain and its Role in Channel Gating. *The Journal of General Physiology*, 122(3) :333–348.

[Bellenguez et al., 2017] Bellenguez, C., Charbonnier, C., Grenier-Boley, B., Quenez, O., Le Guennec, K., Nicolas, G., Chauhan, G., Wallon, D., Rousseau, S., Richard, A. C., Boland, A., Bourque, G., Munter, H. M., Olaso, R., Meyer, V., Rollin-Sillaire, A., Pasquier, F., Letenneur, L., Redon, R., Dartigues, J.-F., Tzourio, C., Frebourg, T., Lathrop, M., Deleuze, J.-F., Hannequin, D., Genin, E., Amouyel, P., Debette, S., Lambert, J.-C., Campion, D., and CNR MAJ collaborators (2017). Contribution to Alzheimer's disease risk of rare variants in TREM2, SORL1, and ABCA7 in 1779 cases and 1273 controls. *Neurobiology of Aging*, 59 :220.e1–220.e9.

[Benz et al., 2014] Benz, N., Le Hir, S., Norez, C., Kerbiriou, M., Calvez, M.-L., Becq, F., Trouvé, P., and Férec, C. (2014). Improvement of chloride transport defect by gonadotropin-releasing hormone (GnRH) in cystic fibrosis epithelial cells. *PloS One*, 9(2) :e88964.

[Black, 2003] Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry*, 72 :291–336.

[Bompadre et al., 2008] Bompadre, S. G., Li, M., and Hwang, T.-C. (2008). Mechanism of G551d-CFTR (cystic fibrosis transmembrane conductance regulator) potentiation by a high affinity ATP analog. *The Journal of Biological Chemistry*, 283(9) :5364–5369.

[Bompadre et al., 2007] Bompadre, S. G., Sohma, Y., Li, M., and Hwang, T.-C. (2007). G551d and G1349d, two CF-associated mutations in the signature sequences of CFTR, exhibit distinct gating defects. *The Journal of General Physiology*, 129(4) :285–298.

[Boyle et al., 2014] Boyle, M. P., Bell, S. C., Konstan, M. W., McColley, S. A., Rowe, S. M., Rietschel, E., Huang, X., Waltz, D., Patel, N. R., Rodman, D., and VX09-809-102 study group (2014). A CFTR corrector (lumacaftor) and a CFTR potentiator (ivacaftor) for treatment of patients with cystic fibrosis who have a phe508del CFTR mutation : a phase 2 randomised controlled trial. *The Lancet. Respiratory Medicine*, 2(7) :527–538.

[Browning and Browning, 2007] Browning, S. and Browning, B. (2007). Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *American Journal of Human Genetics*, 81(5) :1084–1097.

[Brugnon et al., 2008] Brugnon, F., Bilan, F., Heraud, M.-C., Grizard, G., Janny, L., and Creveaux, I. (2008). Outcome of intracytoplasmic sperm injection for a couple in which the man is carrier of CFTR p.[R74w ;V201m ;D1270n] and p.P841r mutations and his spouse a heterozygous carrier of p.F508del mutation of the cystic fibrosis transmembrane conductance regulator gene. *Fertility and Sterility*, 90(5) :2004.e23–26.

[Cant et al., 2014] Cant, N., Pollock, N., and Ford, R. C. (2014). CFTR structure and cystic fibrosis. *The International Journal of Biochemistry & Cell Biology*, 52 :15–25.

[Castaldo et al., 2001] Castaldo, G., Fuccio, A., Salvatore, D., Raia, V., Santostasi, T., Leonardi, S., Lizzi, N., La Rosa, M., Rigillo, N., and Salvatore, F. (2001). Liver expression in cystic fibrosis could be modulated by genetic factors different from the cystic fibrosis transmembrane regulator genotype. *American Journal of Medical Genetics*, 98(4) :294–297.

[Castellani and Assael, 2017] Castellani, C. and Assael, B. M. (2017). Cystic fibrosis : a clinical view. *Cellular and molecular life sciences : CMLS*, 74(1) :129–140.

[Castellani and CFTR2 team, 2013] Castellani, C. and CFTR2 team (2013). CFTR2 : How will it help care ? *Paediatric Respiratory Reviews*, 14 Suppl 1 :2–5.

[Centre de Référence Mucoviscidose de Lyon, 2017] Centre de Référence Mucoviscidose de Lyon, . (2017). Protocole National de Diagnostic et de Soins (PNDS) Mucoviscidose.

[Chamary et al., 2006] Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence : non-neutral evolution at synonymous sites in mammals. *Nature Reviews. Genetics*, 7(2) :98–108.

[Chang et al., 2007] Chang, Y.-F., Imam, J. S., and Wilkinson, M. F. (2007). The nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry*, 76 :51–74.

[Chanoux and Rubenstein, 2012] Chanoux, R. A. and Rubenstein, R. C. (2012). Molecular Chaperones as Targets to Circumvent the CFTR Defect in Cystic Fibrosis. *Frontiers in Pharmacology*, 3.

[Chen and Schrijver, 2011] Chen, N. and Schrijver, I. (2011). Allelic discrimination of cis-trans relationships by digital polymerase chain reaction : GJB2 (p.V27i/p.E114g) and CFTR (p.R117h/5t). *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, 13(12) :1025–1031.

[Choi et al., 2018] Choi, Y., Chan, A. P., Kirkness, E., Telenti, A., and Schork, N. J. (2018). Comparison of phasing strategies for whole human genomes. *PLoS genetics*, 14(4) :e1007308.

[Chu et al., 1993] Chu, C. S., Trapnell, B. C., Curristin, S., Cutting, G. R., and Crystal, R. G. (1993). Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA. *Nature Genetics*, 3(2) :151–156.

[Cingolani et al., 2012] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff : SNPs in the genome of Drosophila melanogaster strain w1118 ; iso-2 ; iso-3. *Fly*, 6(2) :80–92.

[Clain et al., 2001] Clain, J., Fritsch, J., Lehmann-Che, J., Bali, M., Arous, N., Goossens, M., Edelman, A., and Fanen, P. (2001). Two mild cystic fibrosis-associated mutations result in severe cystic fibrosis when combined in cis and reveal a residue important for cystic fibrosis transmembrane conductance regulator processing and function. *The Journal of Biological Chemistry*, 276(12) :9045–9049.

[Claustres, 2005] Claustres, M. (2005). Molecular pathology of the CFTR locus in male infertility. *Reproductive Biomedicine Online*, 10(1) :14–41.

[Claustres et al., 2004] Claustres, M., Altiéri, J.-P., Guittard, C., Templin, C., Chevalier-Porst, F., and Des Georges, M. (2004). Are p.I148t, p.R74w and p.D1270n cystic fibrosis causing mutations ? *BMC medical genetics*, 5 :19.

[Claustres et al., 2017] Claustres, M., Thèze, C., des Georges, M., Baux, D., Girodon, E., Bienvenu, T., Audrezet, M.-P., Dugueperoux, I., Férec, C., Lalau, G., Pagin, A., Kitzis, A., Thoreau, V., Gaston, V., Bieth, E., Malinge, M.-C., Reboul, M.-P., Fergelot, P., Lemonnier, L., Mekki, C., Fanen, P., Bergougnoux, A., Sasorith, S., Raynal, C., and Bareil, C. (2017). CFTR-France, a national relational patient database for sharing genetic and phenotypic data associated with rare CFTR variants. *Human Mutation*, 38(10) :1297–1315.

[Coban-Akdemir et al., 2018] Coban-Akdemir, Z., White, J. J., Song, X., Jhangiani, S. N., Fatih, J. M., Gambin, T., Bayram, Y., Chinn, I. K., Karaca, E., Punetha, J., Poli, C., Baylor-Hopkins Center for Mendelian Genomics, Boerwinkle, E., Shaw, C. A., Orange, J. S., Gibbs, R. A., Lappalainen, T., Lupski, J. R., and Carvalho, C. M. B. (2018). Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *American Journal of Human Genetics*, 103(2) :171–187.

[Collaco et al., 2016] Collaco, J. M., Blackman, S. M., Raraigh, K. S., Corvol, H., Rommens, J. M., Pace, R. G., Boelle, P.-Y., McGready, J., Sosnay, P. R., Strug, L. J., Knowles, M. R., and Cutting, G. R.

(2016). Sources of Variation in Sweat Chloride Measurements in Cystic Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 194(11) :1375–1382.

[Cuppens et al., 1998] Cuppens, H., Lin, W., Jaspers, M., Costes, B., Teng, H., Vankeerberghen, A., Jorissen, M., Droogmans, G., Reynaert, I., Goossens, M., Nilius, B., and Cassiman, J. J. (1998). Polyvariant mutant cystic fibrosis transmembrane conductance regulator genes. The polymorphic (Tg)m locus explains the partial penetrance of the T5 polymorphism as a disease mutation. *The Journal of Clinical Investigation*, 101(2) :487–496.

[Cystic Fibrosis Foundation, 2017] Cystic Fibrosis Foundation, . (2017). 2017 Patient Registry Annual Data Report. page 96.

[Dalemans et al., 1991] Dalemans, W., Barbry, P., Champigny, G., Jallat, S., Dott, K., Dreyer, D., Crystal, R. G., Pavirani, A., Lecocq, J. P., and Lazdunski, M. (1991). Altered chloride ion channel kinetics associated with the delta F508 cystic fibrosis mutation. *Nature*, 354(6354) :526–528.

[Danecek et al., 2011] Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15) :2156–2158.

[Davies et al., 2000] Davies, J., Neth, O., Alton, E., Klein, N., and Turner, M. (2000). Differential binding of mannose-binding lectin to respiratory pathogens in cystic fibrosis. *Lancet (London, England)*, 355(9218) :1885–1886.

[Davies, 2002] Davies, J. C. (2002). Pseudomonas aeruginosa in cystic fibrosis : pathogenesis and persistence. *Paediatric Respiratory Reviews*, 3(2) :128–134.

[Davis, 2006] Davis, P. B. (2006). Cystic fibrosis since 1938. *American Journal of Respiratory and Critical Care Medicine*, 173(5) :475–482.

[de Prada Merino et al., 2010] de Prada Merino, A., Bütschi, F. N., Bouchardy, I., Beckmann, J. S., Morris, M. A., Hafen, G. M., and Fellmann, F. (2010). [R74w ;R1070w ;D1270n] : a new complex allele responsible for cystic fibrosis. *Journal of Cystic Fibrosis : Official Journal of the European Cystic Fibrosis Society*, 9(6) :447–449.

[Delaneau et al., 2014] Delaneau, O., Marchini, J., and The 1000 Genomes Project Consortium (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, 5(1) :3934.

[Delaneau et al., 2012] Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2) :179–181.

[Delaneau et al., 2013] Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10(1) :5–6.

[den Dunnen et al., 2016] den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., Roux, A.-F., Smith, T., Antonarakis, S. E., and Taschner, P. E. M. (2016). HGVS Recommendations for the Description of Sequence Variants : 2016 Update. *Human Mutation*, 37(6) :564–569.

[Diana et al., 2016] Diana, A., Polizzi, A. M., Santostasi, T., Ratclif, L., Pantaleo, M. G., Leonetti, G., Iusco, D. R., Gallo, C., Conese, M., and Manca, A. (2016). The novel complex allele [A238v ;F508del] of the CFTR gene : clinical phenotype and possible implications for cystic fibrosis etiological therapies. *Journal of Human Genetics*, 61(6) :473–481.

[Domingue et al., 2014] Domingue, J. C., Ao, M., Sarathy, J., George, A., Alrefai, W. A., Nelson, D. J., and Rao, M. C. (2014). HEK-293 cells expressing the cystic fibrosis transmembrane conductance regulator (CFTR) : a model for studying regulation of Cl- transport. *Physiological Reports*, 2(9).

[Drumm, 2001] Drumm, M. L. (2001). Modifier genes and variation in cystic fibrosis. *Respiratory Research*, 2(3) :125–128.

[Du et al., 2005] Du, K., Sharma, M., and Lukacs, G. L. (2005). The DeltaF508 cystic fibrosis mutation impairs domain-domain interactions and arrests post-translational folding of CFTR. *Nature Structural & Molecular Biology*, 12(1) :17–25.

[DuBridge et al., 1987] DuBridge, R. B., Tang, P., Hsia, H. C., Leong, P. M., Miller, J. H., and Calos, M. P. (1987). Analysis of mutation in human cells by using an Epstein-Barr virus shuttle system. *Molecular and Cellular Biology*, 7(1) :379–387.

[Dujardin et al., 2011] Dujardin, G., Commandeur, D., Le Jossic-Corcos, C., Ferec, C., and Corcos, L. (2011). Splicing defects in the CFTR gene : minigene analysis of two mutations, 1811+1g>C and 1898+3a>G. *Journal of Cystic Fibrosis : Official Journal of the European Cystic Fibrosis Society*, 10(3) :212–216.

[El-Gebali et al., 2019] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1) :D427–D432.

[El-Seedy et al., 2012] El-Seedy, A., Girodon, E., Norez, C., Pajaud, J., Pasquet, M.-C., de Becdelièvre, A., Bienvenu, T., des Georges, M., Cabet, F., Lalau, G., Bieth, E., Blayau, M., Becq, F., Kitzis, A., Fanen, P., and Ladeveze, V. (2012). CFTR mutation combinations producing frequent complex alleles with different clinical and functional outcomes. *Human Mutation*, 33(11) :1557–1565.

[Ewing et al., 1998] Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8(3) :175–185.

[Fanen et al., 1999] Fanen, P., Clain, J., Labarthe, R., Hulin, P., Girodon, E., Pagesy, P., Goossens, M., and Edelman, A. (1999). Structure-function analysis of a double-mutant cystic fibrosis transmembrane conductance regulator protein occurring in disorders related to cystic fibrosis. *FEBS letters*, 452(3) :371–374.

[Fanen et al., 1992] Fanen, P., Ghanem, N., Vidaud, M., Besmond, C., Martin, J., Costes, B., Plassa, F., and Goossens, M. (1992). Molecular characterization of cystic fibrosis : 16 novel mutations identified by analysis of the whole cystic fibrosis conductance transmembrane regulator (CFTR) coding regions and splice site junctions. *Genomics*, 13(3) :770–776.

[Farber et al., 1943] Farber, S., Shwachman, H., and Maddock, C. L. (1943). PANCREATIC FUNCTION AND DISEASE IN EARLY LIFE. I. PANCREATIC ENZYME ACTIVITY AND THE CELIAC SYNDROME 1. *Journal of Clinical Investigation*, 22(6) :827–838.

[Farre et al., 2007] Farre, C., Stoelzle, S., Haarmann, C., George, M., Brüggemann, A., and Fertig, N. (2007). Automated ion channel screening : patch clamping made easy. *Expert Opinion on Therapeutic Targets*, 11(4) :557–565.

[Farrell et al., 2017] Farrell, P. M., White, T. B., Ren, C. L., Hempstead, S. E., Accurso, F., Derichs, N., Howenstine, M., McColley, S. A., Rock, M., Rosenfeld, M., Sermet-Gaudelus, I., Southern, K. W., Marshall, B. C., and Sosnay, P. R. (2017). Diagnosis of Cystic Fibrosis : Consensus Guidelines from the Cystic Fibrosis Foundation. *The Journal of Pediatrics*, 181S :S4–S15.e1.

[Fay et al., 2018] Fay, J. F., Aleksandrov, L. A., Jensen, T. J., Cui, L. L., Kousouros, J. N., He, L., Aleksandrov, A. A., Gingerich, D. S., Riordan, J., and Chen, J. Z. (2018). Cryo-EM visualization of an active high open probability CFTR ion channel. preprint, Biophysics.

[Fichou et al., 2008] Fichou, Y., Génin, E., Le Maréchal, C., Audrézet, M.-P., Scotet, V., and Férec, C. (2008). Estimating the age of CFTR mutations predominantly found in Brittany (Western France). *Journal of Cystic Fibrosis : Official Journal of the European Cystic Fibrosis Society*, 7(2) :168–173.

[Finn et al., 2016] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database : towards a more sustainable future. *Nucleic Acids Research*, 44(D1) :D279–285.

[Fraser-Pitt and O'Neil, 2015] Fraser-Pitt, D. and O'Neil, D. (2015). Cystic fibrosis - a multiorgan protein misfolding disease. *Future science OA*, 1(2) :FSO57.

[Fu et al., 2013] Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Rieder, M. J., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., NHLBI Exome Sequencing Project, and Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431) :216–220.

[Gabolde et al., 2001] Gabolde, M., Hubert, D., Guilloud-Bataille, M., Lenaerts, C., Feingold, J., and Besmond, C. (2001). The mannose binding lectin gene influences the severity of chronic liver disease in cystic fibrosis. *Journal of Medical Genetics*, 38(5) :310–311.

[Garred et al., 1999] Garred, P., Pressler, T., Madsen, H. O., Frederiksen, B., Svejgaard, A., Høiby, N., Schwartz, M., and Koch, C. (1999). Association of mannose-binding lectin gene heterogeneity with severity of lung disease and survival in cystic fibrosis. *The Journal of Clinical Investigation*, 104(4) :431–437.

[Genin et al., 2008] Genin, E., Feingold, J., and Clerget-Darpoux, F. (2008). Identifying modifier genes of monogenic disease : strategies and difficulties. *Human Genetics*, 124(4) :357–368.

[Genin et al., 2017] Genin, E., Redon, R., Deleuze, J.-F., Campion, D., Lambert, J.-C., Dartigues, J.-F., and Consortium, F. (2017). The French Exome (FREX) Project : a population-based panel of exomes to help filter out common local variants.

[Gibson and Cooke, 1959] Gibson, L. E. and Cooke, R. E. (1959). A test for concentration of electrolytes in sweat in cystic fibrosis of the pancreas utilizing pilocarpine by iontophoresis. *Pediatrics*, 23(3) :545–549.

[Gibson-Corley et al., 2016] Gibson-Corley, K. N., Meyerholz, D. K., and Engelhardt, J. F. (2016). Pancreatic pathophysiology in cystic fibrosis. *The Journal of Pathology*, 238(2) :311–320.

[Girodon-Boulandet and Costa, 2005] Girodon-Boulandet, E. and Costa, C. (2005). Génétique de la mucoviscidose. *Médecine thérapeutique / Pédiatrie*, 8(3) :126–134.

[Glusman et al., 2017] Glusman, G., Rose, P. W., Prlić, A., Dougherty, J., Duarte, J. M., Hoffman, A. S., Barton, G. J., Bendixen, E., Bergquist, T., Bock, C., Brunk, E., Buljan, M., Burley, S. K., Cai, B., Carter, H., Gao, J., Godzik, A., Heuer, M., Hicks, M., Hrabe, T., Karchin, R., Leman, J. K., Lane, L., Masica, D. L., Mooney, S. D., Moult, J., Omenn, G. S., Pearl, F., Pejaver, V., Reynolds, S. M., Rokem, A., Schwede, T., Song, S., Tilgner, H., Valasatava, Y., Zhang, Y., and Deutsch, E. W. (2017). Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation : a proposed framework. *Genome Medicine*, 9(1) :113.

[Golowasch and Nadim, 2014] Golowasch, J. and Nadim, F. (2014). Capacitance, Membrane. In Jaeger, D. and Jung, R., editors, *Encyclopedia of Computational Neuroscience*, pages 1–5. Springer New York, New York, NY.

[Gregory et al., 1990] Gregory, R. J., Cheng, S. H., Rich, D. P., Marshall, J., Paul, S., Hehir, K., Ostedgaard, L., Klinger, K. W., Welsh, M. J., and Smith, A. E. (1990). Expression and characterization of the cystic fibrosis transmembrane conductance regulator. *Nature*, 347(6291) :382–386.

[Han et al., 2018] Han, S. T., Rab, A., Pellicore, M. J., Davis, E. F., McCague, A. F., Evans, T. A., Joynt, A. T., Lu, Z., Cai, Z., Raraigh, K. S., Hong, J. S., Sheppard, D. N., Sorscher, E. J., and Cutting, G. R. (2018). Residual function of cystic fibrosis mutants predicts response to small molecule CFTR modulators. *JCI insight*, 3(14).

[Hanzén et al., 2016] Hanzén, S., Vielfort, K., Yang, J., Roger, F., Andersson, V., Zamarbide-Forés, S., Andersson, R., Malm, L., Palais, G., Biteau, B., Liu, B., Toledano, M. B., Molin, M., and Nyström, T. (2016). Lifespan Control by Redox-Dependent Recruitment of Chaperones to Misfolded Proteins. *Cell*, 166(1) :140–151.

[Hattori, 2005] Hattori, M. (2005). [Finishing the euchromatic sequence of the human genome]. *Tanpakushitsu Kakusan Koso. Protein, Nucleic Acid, Enzyme*, 50(2) :162–168.

[Haute Autorité de Santé, 2012] Haute Autorité de Santé, . (2012). KALYDECO (ivacaftor), potentiateur de la protéine CFTR.

[Haute Autorité de Santé, 2019] Haute Autorité de Santé, . (2019). ORKAMBI (lumacaftor / ivacaftor), correcteur et potentialisateur du gène CFTR.

[Haws et al., 1996] Haws, C. M., Nepomuceno, I. B., Krouse, M. E., Wakelee, H., Law, T., Xia, Y., Nguyen, H., and Wine, J. J. (1996). Delta F508-CFTR channels : kinetics, activation by forskolin, and potentiation by xanthines. *The American Journal of Physiology*, 270(5 Pt 1) :C1544–1555.

[Hayden, 2014] Hayden, E. C. (2014). Is the $1,000 genome for real ? *Nature News*.

[He et al., 2015] He, L., Aleksandrov, A. A., An, J., Cui, L., Yang, Z., Brouillette, C. G., and Riordan, J. R. (2015). Restoration of NBD1 thermal stability is necessary and sufficient to correct F508 CFTR folding and assembly. *Journal of molecular biology*, 427(1) :106–120.

[Heath et al., 2008] Heath, S. C., Gut, I. G., Brennan, P., McKay, J. D., Bencko, V., Fabianova, E., Foretova, L., Georges, M., Janout, V., Kabesch, M., Krokan, H. E., Elvestad, M. B., Lissowska, J., Mates, D., Rudnai, P., Skorpen, F., Schreiber, S., Soria, J. M., Syvänen, A.-C., Meneton, P., Herçberg, S., Galan, P., Szeszenia-Dabrowska, N., Zaridze, D., Génin, E., Cardon, L. R., and Lathrop, M. (2008). Investigation of the fine structure of European populations with applications to disease association studies. *European Journal of Human Genetics*, 16(12) :1413–1429.

[Helenius and Aebi, 2001] Helenius, A. and Aebi, M. (2001). Intracellular functions of N-linked glycans. *Science (New York, N.Y.)*, 291(5512) :2364–2369.

[Hoehe et al., 2019] Hoehe, M. R., Herwig, R., Mao, Q., Peters, B. A., Drmanac, R., Church, G. M., and Huebsch, T. (2019). Significant abundance of cis configurations of coding variants in diploid human genomes. *Nucleic Acids Research*, 47(6) :2981–2995.

[Hubert et al., 1996] Hubert, D., Bienvenu, T., Desmazes-Dufeu, N., Fajac, I., Lacronique, J., Matran, R., Kaplan, J. C., and Dusser, D. J. (1996). Genotype-phenotype relationships in a cohort of adult cystic fibrosis patients. *The European Respiratory Journal*, 9(11) :2207–2214.

[Huguet et al., 2016] Huguet, F., Calvez, M. L., Benz, N., Le Hir, S., Mignen, O., Buscaglia, P., Horgen, F. D., Férec, C., Kerbiriou, M., and Trouvé, P. (2016). Function and regulation of TRPM7, as well as intracellular magnesium content, are altered in cells expressing F508-CFTR and G551d-CFTR. *Cellular and molecular life sciences : CMLS*, 73(17) :3351–3373.

[Humphrey et al., 1996] Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD : visual molecular dynamics. *Journal of Molecular Graphics*, 14(1) :33–38, 27–28.

[Hwang et al., 2018] Hwang, T.-C., Yeh, J.-T., Zhang, J., Yu, Y.-C., Yeh, H.-I., and Destefano, S. (2018). Structural mechanisms of CFTR function and dysfunction. *The Journal of General Physiology*, 150(4) :539–570.

[Ihaka and Gentleman, 1996] Ihaka, R. and Gentleman, R. (1996). R : A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3) :299–314.

[Inoue et al., 2004] Inoue, K., Khajavi, M., Ohyama, T., Hirabayashi, S.-i., Wilson, J., Reggin, J. D., Mancias, P., Butler, I. J., Wilkinson, M. F., Wegner, M., and Lupski, J. R. (2004). Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nature Genetics*, 36(4) :361–369.

[International Human Genome Sequencing Consortium, 2001] International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921.

[Joshi and Fass, 2011] Joshi, N. and Fass, J. (2011). Sickle : A sliding-window, adaptive, quality-based trimming tool for FastQ files [Software].

[Karakachoff et al., 2015] Karakachoff, M., Duforet-Frebourg, N., Simonet, F., Le Scouarnec, S., Pellen, N., Lecointe, S., Charpentier, E., Gros, F., Cauchi, S., Froguel, P., Copin, N., D.E.S.I.R. Study Group, Le Tourneau, T., Probst, V., Le Marec, H., Molinaro, S., Balkau, B., Redon, R., Schott, J.-J., Blum, M. G., Dina, C., and D E S I R Study Group (2015). Fine-scale human genetic structure in Western France. *European journal of human genetics : EJHG*, 23(6) :831–836.

[Karki et al., 2015] Karki, R., Pandya, D., Elston, R. C., and Ferlini, C. (2015). Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC Medical Genomics*, 8.

[Kerem et al., 1989] Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., and Tsui, L. C. (1989). Identification of the cystic fibrosis gene : genetic analysis. *Science (New York, N.Y.)*, 245(4922) :1073–1080.

[Kerem et al., 1990] Kerem, E., Corey, M., Kerem, B. S., Rommens, J., Markiewicz, D., Levison, H., Tsui, L. C., and Durie, P. (1990). The relation between genotype and phenotype in cystic fibrosis–analysis of the most common mutation (delta F508). *The New England Journal of Medicine*, 323(22) :1517–1522.

[Kerem et al., 2014] Kerem, E., Konstan, M. W., De Boeck, K., Accurso, F. J., Sermet-Gaudelus, I., Wilschanski, M., Elborn, J. S., Melotti, P., Bronsveld, I., Fajac, I., Malfroot, A., Rosenbluth, D. B., Walker, P. A., McColley, S. A., Knoop, C., Quattrucci, S., Rietschel, E., Zeitlin, P. L., Barth, J., Elfring, G. L., Welch, E. M., Branstrom, A., Spiegel, R. J., Peltz, S. W., Ajayi, T., Rowe, S. M., and Cystic Fibrosis Ataluren Study Group (2014). Ataluren for the treatment of nonsense-mutation cystic fibrosis : a randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet. Respiratory Medicine*, 2(7) :539–547.

[Kessler and Andersen, 1951] Kessler, W. R. and Andersen, D. H. (1951). Heat prostration in fibrocystic disease of the pancreas and other conditions. *Pediatrics*, 8(5) :648–656.

[Khan et al., 2018] Khan, W., Varma Saripella, G., Ludwig, T., Cuppens, T., Thibord, F., Génin, E., Deleuze, J.-F., and Trégouët, D.-A. (2018). MACARON : a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data. *Bioinformatics (Oxford, England)*, 34(19) :3396–3398.

[Kitts and Sherry, 2002] Kitts, A. and Sherry, S. (2002). The Single Nucleotide Polymorphism Database (dbSNP) of Nucleotide Sequence Variation. *NCBI Handbook*, page 30.

[Kosova et al., 2010] Kosova, G., Pickrell, J. K., Kelley, J. L., McArdle, P. F., Shuldiner, A. R., Abney, M., and Ober, C. (2010). The CFTR Met 470 allele is associated with lower birth rates in fertile men from a population isolate. *PLoS genetics*, 6(6) :e1000974.

[Kriegenburg et al., 2012] Kriegenburg, F., Ellgaard, L., and Hartmann-Petersen, R. (2012). Molecular chaperones in targeting misfolded proteins for ubiquitin-dependent degradation. *The FEBS journal*, 279(4) :532–542.

[Kumar et al., 2009] Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7) :1073–1081.

[Landrum et al., 2016] Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W., and Maglott, D. R. (2016). ClinVar : public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1) :D862–868.

[Landrum et al., 2018] Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K., Katz, K., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G., Zhou, G., Holmes, J. B., Kattman, B. L., and Maglott, D. R. (2018). ClinVar : improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1) :D1062–D1067.

[Le Drévo et al., 2008] Le Drévo, M.-A., Benz, N., Kerbiriou, M., Giroux-Metges, M.-A., Pennec, J.-P., Trouvé, P., and Claude, F. (2008). Annexin A5 increases the cell surface expression and the chloride channel function of the F508-cystic fibrosis transmembrane regulator. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1782(10) :605.

[Lee et al., 2012] Lee, C. M., Flynn, R., Hollywood, J. A., Scallan, M. F., and Harrison, P. T. (2012). Correction of the F508 Mutation in the Cystic Fibrosis Transmembrane Conductance Regulator Gene by Zinc-Finger Nuclease Homology-Directed Repair. *BioResearch Open Access*, 1(3) :99–108.

[Lee and Rio, 2015] Lee, Y. and Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, 84 :291–323.

[Lek et al., 2016] Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H.-H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., MacArthur, D. G., and Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616) :285–291.

[Leslie et al., 2015] Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., Wellcome Trust Case Control Consortium 2, International Multiple Sclerosis Genetics Consortium, Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., and Bodmer, W. (2015). The fine-scale genetic structure of the British population. *Nature*, 519(7543) :309–314.

[Lewis et al., 2005] Lewis, H. A., Zhao, X., Wang, C., Sauder, J. M., Rooney, I., Noland, B. W., Lorimer, D., Kearins, M. C., Conners, K., Condon, B., Maloney, P. C., Guggino, W. B., Hunt, J. F., and Emtage, S. (2005). Impact of the deltaF508 mutation in first nucleotide-binding domain of human cystic fibrosis

transmembrane conductance regulator on domain folding and structure. *The Journal of Biological Chemistry*, 280(2) :1346–1353.

[Li, 2013] Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv :1303.3997 [q-bio]*. arXiv : 1303.3997.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14) :1754–1760.

[Liu et al., 2017] Liu, F., Zhang, Z., Csanády, L., Gadsby, D. C., and Chen, J. (2017). Molecular Structure of the Human CFTR Ion Channel. *Cell*, 169(1) :85–95.e8.

[Loh et al., 2016] Loh, P.-R., Palamara, P. F., and Price, A. L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, 48(7) :811–816.

[Lowry et al., 1951] Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951). Protein measurement with the Folin phenol reagent. *The Journal of Biological Chemistry*, 193(1) :265–275.

[Lucarelli et al., 2010] Lucarelli, M., Narzi, L., Pierandrei, S., Bruno, S. M., Stamato, A., d'Avanzo, M., Strom, R., and Quattrucci, S. (2010). A new complex allele of the CFTR gene partially explains the variable phenotype of the L997f mutation. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, 12(9) :548–555.

[Lukacs and Verkman, 2012] Lukacs, G. L. and Verkman, A. S. (2012). CFTR : folding, misfolding and correcting the F508 conformational defect. *Trends in Molecular Medicine*, 18(2) :81–91.

[Lázaro et al., 1999] Lázaro, C., de Cid, R., Sunyer, J., Soriano, J., Giménez, J., Alvarez, M., Casals, T., Antó, J. M., and Estivill, X. (1999). Missense mutations in the cystic fibrosis gene in adult patients with asthma. *Human Mutation*, 14(6) :510–519.

[MacArthur et al., 2014] MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., Adams, D. R., Altman, R. B., Antonarakis, S. E., Ashley, E. A., Barrett, J. C., Biesecker, L. G., Conrad, D. F., Cooper, G. M., Cox, N. J., Daly, M. J., Gerstein, M. B., Goldstein, D. B., Hirschhorn, J. N., Leal, S. M., Pennacchio, L. A., Stamatoyannopoulos, J. A., Sunyaev, S. R., Valle, D., Voight, B. F., Winckler, W., and Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508(7497) :469–476.

[Maitra et al., 2013] Maitra, R., Sivashanmugam, P., and Warner, K. (2013). A rapid membrane potential assay to monitor CFTR function and inhibition. *Journal of Biomolecular Screening*, 18(9) :1132–1137.

[Mardis, 2008] Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG*, 24(3) :133–141.

[Martin et al., 2016] Martin, C., Hamard, C., Kanaan, R., Boussaud, V., Grenet, D., Abély, M., Hubert, D., Munck, A., Lemonnier, L., and Burgel, P.-R. (2016). Causes of death in French cystic fibrosis patients : The need for improvement in transplantation referral strategies ! *Journal of Cystic Fibrosis : Official Journal of the European Cystic Fibrosis Society*, 15(2) :204–212.

[Matthews et al., 1964] Matthews, L. W., Doershuk, C. F., Wise, M., Eddy, G., Nudelman, H., and Spector, S. (1964). A THERAPEUTIC REGIMEN FOR PATIENTS WITH CYSTIC FIBROSIS. *The Journal of Pediatrics*, 65 :558–575.

[Maxam and Gilbert, 1977] Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2) :560–564.

[Mayer and Bukau, 2005] Mayer, M. P. and Bukau, B. (2005). Hsp70 chaperones : cellular functions and molecular mechanism. *Cellular and molecular life sciences : CMLS*, 62(6) :670–684.

[McDonald et al., 2017] McDonald, C. M., Campbell, C., Torricelli, R. E., Finkel, R. S., Flanigan, K. M., Goemans, N., Heydemann, P., Kaminska, A., Kirschner, J., Muntoni, F., Osorio, A. N., Schara, U., Sejersen, T., Shieh, P. B., Sweeney, H. L., Topaloglu, H., Tulinius, M., Vilchez, J. J., Voit, T., Wong, B., Elfring, G., Kroger, H., Luo, X., McIntosh, J., Ong, T., Riebling, P., Souza, M., Spiegel, R. J., Peltz, S. W., Mercuri, E., Clinical Evaluator Training Group, and ACT DMD Study Group (2017). Ataluren in patients with nonsense mutation Duchenne muscular dystrophy (ACT DMD) : a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet (London, England)*, 390(10101) :1489–1498.

[McEvoy et al., 2006] McEvoy, B., Beleza, S., and Shriver, M. D. (2006). The genetic architecture of normal variation in human pigmentation : an evolutionary perspective and model. *Human Molecular Genetics*, 15 Spec No 2 :R176–181.

[McLaren et al., 2016] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1) :122.

[Mekus et al., 2003] Mekus, F., Laabs, U., Veeze, H., and Tümmler, B. (2003). Genes in the vicinity of CFTR modulate the cystic fibrosis phenotype in highly concordant or discordant F508del homozygous sib pairs. *Human Genetics*, 112(1) :1–11.

[Menardi et al., 2002] Menardi, G., Perotti, L., Prucca, M., Martini, S., Prandi, G., and Peano, G. (2002). A very rare association of three mutations of the HFE gene for hemochromatosis. *Genetic Testing*, 6(4) :331–334.

[Mendoza et al., 2012] Mendoza, J. L., Schmidt, A., Li, Q., Caspa, E., Barrett, T., Bridges, R. J., Feranchak, A. P., Brautigam, C. A., and Thomas, P. J. (2012). Requirements for Efficient Correction of F508 CFTR Revealed by Analyses of Evolved Sequences. *Cell*, 148(1-2) :164–174.

[Moisan et al., 2016] Moisan, S., Berlivet, S., Ka, C., Gac, G. L., Dostie, J., and Férec, C. (2016). Analysis of long-range interactions in primary human cells identifies cooperative CFTR regulatory elements. *Nucleic Acids Research*, 44(6) :2564–2576.

[Molinski et al., 2015] Molinski, S. V., Ahmadi, S., Hung, M., and Bear, C. E. (2015). Facilitating Structure-Function Studies of CFTR Modulator Sites with Efficiencies in Mutagenesis and Functional Screening. *Journal of Biomolecular Screening*, 20(10) :1204–1217.

[Mornon et al., 2015] Mornon, J.-P., Hoffmann, B., Jonic, S., Lehn, P., and Callebaut, I. (2015). Full-open and closed CFTR channels, with lateral tunnels from the cytoplasm and an alternative position of the F508 region, as revealed by molecular dynamics. *Cellular and molecular life sciences : CMLS*, 72(7) :1377–1403.

[Mornon et al., 2008] Mornon, J.-P., Lehn, P., and Callebaut, I. (2008). Atomic model of human cystic fibrosis transmembrane conductance regulator : membrane-spanning domains and coupling interfaces. *Cellular and molecular life sciences : CMLS*, 65(16) :2594–2612.

[Mornon et al., 2009] Mornon, J.-P., Lehn, P., and Callebaut, I. (2009). Molecular models of the open and closed states of the whole human CFTR protein. *Cellular and molecular life sciences : CMLS*, 66(21) :3469–3486.

[Myers and Miller, 1988] Myers, E. W. and Miller, W. (1988). Optimal alignments in linear space. *Computer applications in the biosciences : CABIOS*, 4(1) :11–17.

[Nagy and Maquat, 1998] Nagy, E. and Maquat, L. E. (1998). A rule for termination-codon position within intron-containing genes : when nonsense affects RNA abundance. *Trends in Biochemical Sciences*, 23(6) :198–199.

[Ng et al., 2010] Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics*, 42(1) :30–35.

[Niknafs et al., 2013] Niknafs, N., Kim, D., Kim, R., Diekhans, M., Ryan, M., Stenson, P. D., Cooper, D. N., and Karchin, R. (2013). MuPIT interactive : webserver for mapping variant positions to annotated, interactive 3d structures. *Human Genetics*, 132(11) :1235–1243.

[Norman, 1964] Norman, A. P. (1964). ANTIBIOTICS IN CYSTIC FIBROSIS. *Postgraduate Medical Journal*, 40 :SUPPL :131–132.

[Novembre et al., 2008] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M., and Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, 456(7218) :98–101.

[Nyrén et al., 1993] Nyrén, P., Pettersson, B., and Uhlén, M. (1993). Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Analytical Biochemistry*, 208(1) :171–175.

[Oberg and Mahoney, 2007] Oberg, A. L. and Mahoney, D. W. (2007). Linear mixed effects models. *Methods in Molecular Biology (Clifton, N.J.)*, 404 :213–234.

[O'Connell et al., 2014] O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., Navarro, P., Zagury, J.-F., Wilson, J. F., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M. S., and Marchini, J. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics*, 10(4) :e1004234.

[Okiyoneda et al., 2010] Okiyoneda, T., Barrière, H., Bagdány, M., Rabeh, W. M., Du, K., Höhfeld, J., Young, J. C., and Lukacs, G. L. (2010). Peripheral protein quality control removes unfolded CFTR from the plasma membrane. *Science (New York, N.Y.)*, 329(5993) :805–810.

[Okiyoneda et al., 2013] Okiyoneda, T., Veit, G., Dekkers, J. F., Bagdany, M., Soya, N., Xu, H., Roldan, A., Verkman, A. S., Kurth, M., Simon, A., Hegedus, T., Beekman, J. M., and Lukacs, G. L. (2013). Mechanism-based corrector combination restores F508-CFTR folding and function. *Nature chemical biology*, 9(7).

[Ong et al., 1993] Ong, T., Marshall, S. G., Karczeski, B. A., Sternen, D. L., Cheng, E., and Cutting, G. R. (1993). Cystic Fibrosis and Congenital Absence of the Vas Deferens. In Adam, M. P., Ardinger, H. H., Pagon, R. A., Wallace, S. E., Bean, L. J., Stephens, K., and Amemiya, A., editors, *GeneReviews®*. University of Washington, Seattle, Seattle (WA).

[Ooi and Durie, 2012] Ooi, C. Y. and Durie, P. R. (2012). Cystic fibrosis transmembrane conductance regulator (CFTR) gene mutations in pancreatitis. *Journal of Cystic Fibrosis : Official Journal of the European Cystic Fibrosis Society*, 11(5) :355–362.

[Ostedgaard et al., 2000] Ostedgaard, L. S., Baldursson, O., Vermeer, D. W., Welsh, M. J., and Robertson, A. D. (2000). A functional R domain from cystic fibrosis transmembrane conductance regulator is predominantly unstructured in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10) :5657–5662.

[Palacios, 2013] Palacios, I. M. (2013). Nonsense-mediated mRNA decay : from mechanistic insights to impacts on human health. *Briefings in Functional Genomics*, 12(1) :25–36.

[Patrick et al., 2011] Patrick, A. E., Karamyshev, A. L., Millen, L., and Thomas, P. J. (2011). Alteration of CFTR transmembrane span integration by disease-causing mutations. *Molecular Biology of the Cell*, 22(23) :4461–4471.

[Paul and Apgar, 2005] Paul, P. and Apgar, J. (2005). Single-molecule dilution and multiple displacement amplification for molecular haplotyping. *BioTechniques*, 38(4) :553–554, 556, 558–559.

[Penmatsa et al., 2009] Penmatsa, H., Frederick, C. A., Nekkalapu, S., Conoley, V. G., Zhang, W., Li, C., Kappes, J., Stokes, D. C., and Naren, A. P. (2009). Clinical and Molecular Characterization of S1118f-CFTR. *Pediatric pulmonology*, 44(10) :1003.

[Poli et al., 2018] Poli, M. C., Ebstein, F., Nicholas, S. K., de Guzman, M. M., Forbes, L. R., Chinn, I. K., Mace, E. M., Vogel, T. P., Carisey, A. F., Benavides, F., Coban-Akdemir, Z. H., Gibbs, R. A., Jhangiani, S. N., Muzny, D. M., Carvalho, C. M. B., Schady, D. A., Jain, M., Rosenfeld, J. A., Emrick, L., Lewis, R. A., Lee, B., Undiagnosed Diseases Network members, Zieba, B. A., Küry, S., Krüger, E., Lupski, J. R., Bostwick, B. L., and Orange, J. S. (2018). Heterozygous Truncating Variants in POMP Escape Nonsense-Mediated Decay and Cause a Unique Immune Dysregulatory Syndrome. *American Journal of Human Genetics*, 102(6) :1126–1142.

[Pruitt et al., 2009] Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., Hart, E., Suner, M.-M., Landrum, M. J., Aken, B., Ayling, S., Baertsch, R., Fernandez-Banet, J., Cherry, J. L., Curwen, V., Dicuccio, M., Kellis, M., Lee, J., Lin, M. F., Schuster, M., Shkeda, A., Amid, C., Brown, G., Dukhanina, O., Frankish, A., Hart, J., Maidak, B. L., Mudge, J., Murphy, M. R., Murphy, T., Rajan, J., Rajput, B., Riddick, L. D., Snow, C., Steward, C., Webb, D., Weber, J. A., Wilming, L., Wu, W., Birney, E., Haussler, D., Hubbard, T., Ostell, J., Durbin, R., and Lipman, D. (2009). The consensus coding sequence (CCDS) project : Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, 19(7) :1316–1323.

[Rabbani et al., 2014] Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics*, 59(1) :5–15.

[Radpour et al., 2008] Radpour, R., Gourabi, H., Dizaj, A. V., Holzgreve, W., and Zhong, X. Y. (2008). Genetic investigations of CFTR mutations in congenital absence of vas deferens, uterus, and vagina as a cause of infertility. *Journal of Andrology*, 29(5) :506–513.

[Ramsey et al., 2011] Ramsey, B. W., Davies, J., McElvaney, N. G., Tullis, E., Bell, S. C., Dřevínek, P., Griese, M., McKone, E. F., Wainwright, C. E., Konstan, M. W., Moss, R., Ratjen, F., Sermet-Gaudelus, I., Rowe, S. M., Dong, Q., Rodriguez, S., Yen, K., Ordoñez, C., Elborn, J. S., and VX08-770-102 Study Group (2011). A CFTR potentiator in patients with cystic fibrosis and the G551d mutation. *The New England Journal of Medicine*, 365(18) :1663–1672.

[Raraigh et al., 2018] Raraigh, K. S., Han, S. T., Davis, E., Evans, T. A., Pellicore, M. J., McCague, A. F., Joynt, A. T., Lu, Z., Atalar, M., Sharma, N., Sheridan, M. B., Sosnay, P. R., and Cutting, G. R. (2018). Functional Assays Are Essential for Interpretation of Missense Variants Associated with Variable Expressivity. *American Journal of Human Genetics*, 102(6) :1062–1077.

[Regan et al., 2015] Regan, J. F., Kamitaki, N., Legler, T., Cooper, S., Klitgord, N., Karlin-Neumann, G., Wong, C., Hodges, S., Koehler, R., Tzonev, S., and McCarroll, S. A. (2015). A Rapid Molecular Approach for Chromosomal Phasing. *PLOS ONE*, 10(3) :e0118270.

[Rhoads and Au, 2015] Rhoads, A. and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, 13(5) :278–289.

[Rich et al., 1990] Rich, D. P., Anderson, M. P., Gregory, R. J., Cheng, S. H., Paul, S., Jefferson, D. M., McCann, J. D., Klinger, K. W., Smith, A. E., and Welsh, M. J. (1990). Expression of cystic fibrosis transmembrane conductance regulator corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. *Nature*, 347(6291) :358–363.

[Richards et al., 2015] Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., and ACMG Laboratory Quality

Assurance Committee (2015). Standards and guidelines for the interpretation of sequence variants : a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, 17(5) :405–424.

[Riordan, 2005] Riordan, J. R. (2005). Assembly of functional CFTR chloride channels. *Annual Review of Physiology*, 67 :701–718.

[Riordan et al., 1989] Riordan, J. R., Rommens, J. M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., and Chou, J. L. (1989). Identification of the cystic fibrosis gene : cloning and characterization of complementary DNA. *Science (New York, N.Y.)*, 245(4922) :1066–1073.

[Rogan et al., 2011] Rogan, M. P., Stoltz, D. A., and Hornick, D. B. (2011). Cystic fibrosis transmembrane conductance regulator intracellular processing, trafficking, and opportunities for mutation-specific treatment. *Chest*, 139(6) :1480–1490.

[Rohlfs et al., 2002] Rohlfs, E. M., Zhou, Z., Sugarman, E. A., Heim, R. A., Pace, R. G., Knowles, M. R., Silverman, L. M., and Allitto, B. A. (2002). The I148t CFTR allele occurs on multiple haplotypes : a complex allele is associated with cystic fibrosis. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, 4(5) :319–323.

[Rommens et al., 1989] Rommens, J. M., Iannuzzi, M. C., Kerem, B., Drumm, M. L., Melmer, G., Dean, M., Rozmahel, R., Cole, J. L., Kennedy, D., and Hidaka, N. (1989). Identification of the cystic fibrosis gene : chromosome walking and jumping. *Science (New York, N.Y.)*, 245(4922) :1059–1065.

[Saint Pierre and Génin, 2014] Saint Pierre, A. and Génin, E. (2014). How important are rare variants in common disease ? *Briefings in Functional Genomics*, 13(5) :353–361.

[Salvatore et al., 2002] Salvatore, F., Scudiero, O., and Castaldo, G. (2002). Genotype-phenotype correlation in cystic fibrosis : the role of modifier genes. *American Journal of Medical Genetics*, 111(1) :88–95.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12) :5463–5467.

[Santamaría et al., 2013] Santamaría, I., Menéndez, S. T., and Balbín, M. (2013). EGFR L858r mutation may go undetected because of P848l in cis mutation. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 31(26) :e420–421.

[Schroeder and Neagle, 1996] Schroeder, K. S. and Neagle, B. D. (1996). FLIPR : A New Instrument for Accurate, High Throughput Optical Screening. *Journal of Biomolecular Screening*, 1(2) :75–80.

[Serohijos et al., 2008] Serohijos, A. W. R., Hegedus, T., Aleksandrov, A. A., He, L., Cui, L., Dokholyan, N. V., and Riordan, J. R. (2008). Phenylalanine-508 mediates a cytoplasmic-membrane domain contact in the CFTR 3d structure crucial to assembly and channel function. *Proceedings of the National Academy of Sciences of the United States of America*, 105(9) :3256–3261.

[Shabalina et al., 2013] Shabalina, S. A., Spiridonov, N. A., and Kashina, A. (2013). Sounds of silence : synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Research*, 41(4) :2073–2094.

[Shwachman, 1960] Shwachman, H. (1960). Therapy of cystic fibrosis of the pancreas. *Pediatrics*, 25 :155–163.

[Spooner et al., 2018] Spooner, W., McLaren, W., Slidel, T., Finch, D. K., Butler, R., Campbell, J., Eghobamien, L., Rider, D., Kiefer, C. M., Robinson, M. J., Hardman, C., Cunningham, F., Vaughan,

T., Flicek, P., and Huntington, C. C. (2018). Haplosaurus computes protein haplotypes for use in precision drug design. *Nature Communications*, 9(1) :4128.

[Stajich et al., 2002] Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit : Perl modules for the life sciences. *Genome Research*, 12(10) :1611–1618.

[Stankovic et al., 2008] Stankovic, M., Nikolic, A., Divac, A., Tomovic, A., Petrovic-Stanojevic, N., Andjelic, M., Dopudja-Pantic, V., Surlan, M., Vujicic, I., Ponomarev, D., Mitic-Milikic, M., Kusic, J., and Radojkovic, D. (2008). The CFTR M470v gene variant as a potential modifier of COPD severity : study of Serbian population. *Genetic Testing*, 12(3) :357–362.

[Suarez et al., 2012] Suarez, C. J., Boyle, T., Chiang, T., and Schrijver, I. (2012). Actions and consequences : characterization of a deletion in the CFTR gene that encompasses a splice site. *Grand Rounds*, 12(1) :36–39.

[The International HapMap Consortium, 2005] The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063) :1299–1320.

[Timmreck et al., 2003] Timmreck, L. S., Gray, M. R., Handelin, B., Allito, B., Rohlfs, E., Davis, A. J., Gidwani, G., and Reindollar, R. H. (2003). Analysis of cystic fibrosis transmembrane conductance regulator gene mutations in patients with congenital absence of the uterus and vagina. *American Journal of Medical Genetics. Part A*, 120A(1) :72–76.

[Trouvé et al., 2007] Trouvé, P., Le Drévo, M.-A., Kerbiriou, M., Friocourt, G., Fichou, Y., Gillet, D., and Férec, C. (2007). Annexin V is directly involved in cystic fibrosis transmembrane conductance regulator's chloride channel function. *Biochimica Et Biophysica Acta*, 1772(10) :1121–1133.

[Uguen et al., 2017] Uguen, K., Scotet, V., Ka, C., Gourlaouen, I., L'hostis, C., Merour, M.-C., Cuppens, T., Ferec, C., and Le Gac, G. (2017). Diagnostic value of targeted next-generation sequencing in suspected hemochromatosis patients with a single copy of the HFE p.Cys282tyr causative allele. *American Journal of Hematology*, 92(12) :E664–E666.

[UniProt Consortium, 2019] UniProt Consortium (2019). UniProt : a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1) :D506–D515.

[Van Goor et al., 2009] Van Goor, F., Hadida, S., Grootenhuis, P. D. J., Burton, B., Cao, D., Neuberger, T., Turnbull, A., Singh, A., Joubran, J., Hazlewood, A., Zhou, J., McCartney, J., Arumugam, V., Decker, C., Yang, J., Young, C., Olson, E. R., Wine, J. J., Frizzell, R. A., Ashlock, M., and Negulescu, P. (2009). Rescue of CF airway epithelial cell function in vitro by a CFTR potentiator, VX-770. *Proceedings of the National Academy of Sciences of the United States of America*, 106(44) :18825–18830.

[Van Goor et al., 2006] Van Goor, F., Straley, K. S., Cao, D., González, J., Hadida, S., Hazlewood, A., Joubran, J., Knapp, T., Makings, L. R., Miller, M., Neuberger, T., Olson, E., Panchenko, V., Rader, J., Singh, A., Stack, J. H., Tung, R., Grootenhuis, P. D. J., and Negulescu, P. (2006). Rescue of DeltaF508-CFTR trafficking and gating in human cystic fibrosis airway primary cultures by small molecules. *American Journal of Physiology. Lung Cellular and Molecular Physiology*, 290(6) :L1117–1130.

[Vankeerberghen et al., 1998] Vankeerberghen, A., Wei, L., Jaspers, M., Cassiman, J. J., Nilius, B., and Cuppens, H. (1998). Characterization of 19 disease-associated missense mutations in the regulatory domain of the cystic fibrosis transmembrane conductance regulator. *Human Molecular Genetics*, 7(11) :1761–1769.

[Vecchio-Pagán et al., 2016] Vecchio-Pagán, B., Blackman, S. M., Lee, M., Atalar, M., Pellicore, M. J., Pace, R. G., Franca, A. L., Raraigh, K. S., Sharma, N., Knowles, M. R., and Cutting, G. R. (2016). Deep resequencing of CFTR in 762 F508del homozygotes reveals clusters of non-coding variants associated with cystic fibrosis disease traits. *Human Genome Variation*, 3 :16038.

[Veitch, 2004] Veitch, N. C. (2004). Horseradish peroxidase : a modern view of a classic enzyme. *Phytochemistry*, 65(3) :249–259.

[Walsh et al., 2006] Walsh, A., Dixon, J. L., Ramm, G. A., Hewett, D. G., Lincoln, D. J., Anderson, G. J., Subramaniam, V. N., Dodemaide, J., Cavanaugh, J. A., Bassett, M. L., and Powell, L. W. (2006). The clinical relevance of compound heterozygosity for the C282y and H63d substitutions in hemochromatosis. *Clinical Gastroenterology and Hepatology : The Official Clinical Practice Journal of the American Gastroenterological Association*, 4(11) :1403–1410.

[Wang et al., 2010] Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16) :e164.

[Welsh and Smith, 1993] Welsh, M. J. and Smith, A. E. (1993). Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. *Cell*, 73(7) :1251–1254.

[White et al., 2015] White, J., Mazzeu, J. F., Hoischen, A., Jhangiani, S. N., Gambin, T., Alcino, M. C., Penney, S., Saraiva, J. M., Hove, H., Skovby, F., Kayserili, H., Estrella, E., Vulto-van Silfhout, A. T., Steehouwer, M., Muzny, D. M., Sutton, V. R., Gibbs, R. A., Baylor-Hopkins Center for Mendelian Genomics, Lupski, J. R., Brunner, H. G., van Bon, B. W. M., and Carvalho, C. M. B. (2015). DVL1 frameshift mutations clustering in the penultimate exon cause autosomal-dominant Robinow syndrome. *American Journal of Human Genetics*, 96(4) :612–622.

[White et al., 2016] White, J. J., Mazzeu, J. F., Hoischen, A., Bayram, Y., Withers, M., Gezdirici, A., Kimonis, V., Steehouwer, M., Jhangiani, S. N., Muzny, D. M., Gibbs, R. A., Baylor-Hopkins Center for Mendelian Genomics, van Bon, B. W. M., Sutton, V. R., Lupski, J. R., Brunner, H. G., and Carvalho, C. M. B. (2016). DVL3 Alleles Resulting in a -1 Frameshift of the Last Exon Mediate Autosomal-Dominant Robinow Syndrome. *American Journal of Human Genetics*, 98(3) :553–561.

[Williams et al., 2012] Williams, A. L., Patterson, N., Glessner, J., Hakonarson, H., and Reich, D. (2012). Phasing of many thousands of genotyped samples. *American Journal of Human Genetics*, 91(2) :238–251.

[Xuan et al., 2013] Xuan, J., Yu, Y., Qing, T., Guo, L., and Shi, L. (2013). Next-generation sequencing in the clinic : promises and challenges. *Cancer Letters*, 340(2) :284–295.

[Yang et al., 2016] Yang, R., Kerschner, J. L., Gosalia, N., Neems, D., Gorsic, L. K., Safi, A., Crawford, G. E., Kosak, S. T., Leir, S.-H., and Harris, A. (2016). Differential contribution of cis-regulatory elements to higher order chromatin structure and expression of the CFTR locus. *Nucleic Acids Research*, 44(7) :3082–3094.

[Yu and Sharma, 2019] Yu, E. and Sharma, S. (2019). Cystic Fibrosis. In *StatPearls*. StatPearls Publishing, Treasure Island (FL).

[Yuan et al., 2019] Yuan, P., Liang, Z. K., Liang, H., Zheng, L. Y., Li, D., Li, J., Zhang, J., Tian, J., Lai, L. H., Zhang, K., He, Z. Y., Zhang, Q. X., and Wang, W. J. (2019). Expanding the phenotypic and genetic spectrum of Chinese patients with congenital absence of vas deferens bearing CFTR and ADGRG2 alleles. *Andrology*, 7(3) :329–340.

[Zhang and Chen, 2016] Zhang, Z. and Chen, J. (2016). Atomic Structure of the Cystic Fibrosis Transmembrane Conductance Regulator. *Cell*, 167(6) :1586–1597.e9.

[Zhang et al., 2017] Zhang, Z., Liu, F., and Chen, J. (2017). Conformational Changes of CFTR upon Phosphorylation and ATP Binding. *Cell*, 170(3) :483–491.e8.

[Zhang et al., 2018] Zhang, Z., Liu, F., and Chen, J. (2018). Molecular structure of the ATP-bound, phosphorylated human CFTR. *Proceedings of the National Academy of Sciences*, 115(50) :12757–12762.

[Ziedalski et al., 2006] Ziedalski, T. M., Kao, P. N., Henig, N. R., Jacobs, S. S., and Ruoss, S. J. (2006). Prospective analysis of cystic fibrosis transmembrane regulator mutations in adults with bronchiectasis or pulmonary nontuberculous mycobacterial infection. *Chest*, 130(4) :995–1002.

[Zielenski et al., 1999] Zielenski, J., Corey, M., Rozmahel, R., Markiewicz, D., Aznarez, I., Casals, T., Larriba, S., Mercier, B., Cutting, G. R., Krebsova, A., Macek, M., Langfelder-Schwind, E., Marshall, B. C., DeCelie-Germana, J., Claustres, M., Palacio, A., Bal, J., Nowakowska, A., Ferec, C., Estivill, X., Durie, P., and Tsui, L. C. (1999). Detection of a cystic fibrosis modifier locus for meconium ileus on human chromosome 19q13. *Nature Genetics*, 22(2) :128–129.

[Zielenski and Tsui, 1995] Zielenski, J. and Tsui, L. C. (1995). Cystic fibrosis : genotypic and phenotypic variations. *Annual Review of Genetics*, 29 :777–807.

ANNEXES

Annexe I : Diagnostic value of targeted next-generation sequencing in suspected hemochromatosis patients with a single copy of the HFE p.Cys282Tyr causative allele, Uguen et al., 2017, publié dans l'American Journal of Hematology

Hematology, Department of Cellular Biotechnologies and Hematology, Policlinico Umberto 1, Sapienza University, Rome, Italy

Correspondence
Matteo Molica, MD, Hematology, Department of Cellular Biotechnologies and Hematology, Policlinico Umberto 1, Sapienza University, Via Benevento 6, 00161 Rome, Italy.
Email: molica@bce.uniroma1.it

REFERENCES

[1] Pfirrmann M, Baccarani M, Saussele S, et al. Prognosis of long-term survival considering disease-specific death in patients with chronic myeloid leukemia. *Leukemia*. 2016;30:48–56.

[2] Castagnetti F, Gugliotta G, Breccia M, et al. The Eutos long-term survival score is predictive for response and outcome in CML patients treated frontline with Nilotinib-based regimens. *Abstract*. 2016;S434:EHA.

[3] Deininger M, O'brien SG, Guilhot F, et al. International randomized study of interferon vs STI571 (IRIS) 8-year follow up: sustained survival and low risk for progression or events in patients with newly diagnosed chronic myeloid leukemia in chronic phase (CML-CP) treated with imatinib. *Blood*. 2009;114:462.

[4] Etienne G, Dulucq S, Nicolini FE, et al. Achieving deeper molecular response is associated with a better clinical outcome in chronic myeloid leukemia patients on imatinib front-line therapy. *Haematologica*. 2014; 99:458–464.

[5] Sokal JE, Cox EB, Baccarani M, et al. Prognostic discrimination in "good-risk" chronic granulocytic leukemia. *Blood*. 1984;63:789–799.

[6] Hasford J, Pfirrmann M, Hehlmann R, et al. A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. *J Natl Cancer Inst*. 1998;90:850–858.

Received: 15 September 2017 | Accepted: 19 September 2017

DOI 10.1002/ajh.24912

Diagnostic value of targeted next-generation sequencing in suspected hemochromatosis patients with a single copy of the *HFE* p.Cys282Tyr causative allele

To the Editor:

The best-known and most prevalent form of hemochromatosis is an adult-onset autosomal recessive condition usually associated with the *HFE* p.[Cys282Tyr];[Cys282Tyr] genotype. In Caucasians, this genotype is carried by approximately 1 person in 200–300. A second *HFE* variant, p.His63Asp, has proved to be enriched on the non-p.Cys282Tyr alleles of hemochromatosis patients, and it has been argued that the p.[Cys282Tyr];[His63Asp], but not the p.[His63Asp]+[His63Asp], compound heterozygous genotype increases the risk of disease.[1]

Apart from the widespread p.His63Asp allele, rare *HFE* mutations can be found *in trans* with the causative p.Cys282Tyr substitution. Complex alleles associating the p.His63Asp variant and a second amino-acid alteration have also been observed.[2] Sequencing the entire *HFE* coding region in suspected hemochromatosis patients where the standard *HFE* genetic test reveals a heterozygote status for p.Cys282Tyr is, however, rarely performed.

We used targeted next-generation sequencing to investigate all hemochromatosis genes (*HFE*, *HFE2*, *HAMP*, *TFR2*, and *SLC40A1*) and *BMP6*, which has recently been associated with mild to moderate late-onset iron overload,[3] in 42 patients with unexplained hyperferritinemia (men $\geq$300 μg/L and women $\geq$200 μg/L). All had serum transferrin saturation of at least 60% (men) or at least 50% (women), consistent with the natural history of hemochromatosis. The patient group was composed of 7 p.Cys282Tyr heterozygotes and 35 p.Cys282Tyr/p.His63Asp compound heterozygotes.

The splice site c.76 + 2T>C and nonsense p.Trp155* *HFE* mutations were separately detected in two females firstly identified as simple p.Cys282Tyr heterozygotes. They both had a severe phenotype, manifested by very high levels of serum ferritin and amounts of iron removed by phlebotomy to restore normal iron indices greater than 3 g in the absence of comorbid factors (Table 1). The c.76 + 2T>C splicing mutation is present in the GnomAD database but at a very low frequency (8.19×10^{-6}) as it was only seen in two individuals (one Latino and one European). Here, we demonstrate its functional impact on *HFE* pre-mRNA splicing: it results in intron retention and may give rise to a truncated and inactive protein (Supporting Information Figure 1). To the best of our knowledge, the p.Trp155* nonsense mutation has never previously been reported. It was not detected in 1,460 control chromosomes from the western part of France and, thus, it is also expected to be very rare or private.

We identified a rare *HFE* missense mutation, which changes glutamic acid at position 168 to glutamine (p.Glu168Gln; p.E168Q), in a 51-year-old man with 61% transferrin saturation and 487 μg/L serum ferritin at diagnosis (Table 1). On clinical interview, the patient declared moderate but daily alcohol consumption (approximately 16 U per week). Cloning genomic *HFE* sequences, from exon 2 to exon 3 and from exon 3 to exon 4, revealed that the p.Glu168Gln substitution was *in cis* with the p.His63Asp allele (Supporting Information Figure 2). This observation is consistent with the study of an Italian pedigree, where the p.Glu168Gln and p.His63Asp variants were inherited together (*in cis*) and independently of the p.Cys282Tyr-bearing chromosome.[4] By sequencing the chromosomes of 730 healthy French adults, we found the frequency of the complex allele to be 0.00274.

There is at present moderate evidence to support a causal role for the *HFE* p.Glu168Gln missense mutation. This consists in the c.502G>C transversion reported by the GnomAD Consortium at very low frequency in non-Finnish Europeans (<0.001; rs146519482). Apart from the Italian pedigree, it has been reported only twice in the literature,[2] and genotype-phenotype correlation is impossible. *In silico* evaluation does not clarify the contribution of this rare genetic variant to disease. Indeed, substitution of glutamic acid by glutamine is a biochemically moderate change with a Grantham distance of 29 (0–215).

TABLE 1 Clinical and biological data of the three patients with a rare *HFE* variation *in trans* with the predominant p.Cys282Tyr missense mutation

Genotype								
HFE	BMP6	Gender	Age at diagnosis (years)	Transferrin saturation (%)	Serum ferritin microg/L	AIR (g)	Alcohol glasses/week	BMI kg/m^2
p.[Cys282Tyr];[Trp155*]	p.[=];[=]	Female	67	78	1538	4.4	0	26.8
p.[Cys282Tyr];[p.Leu25fs]	p.[=];[=]	Female	55	75	919	3.2	0	25.2
p.[Cys282Tyr];[His63Asp;Glu168Gln]	p.[=];[=]	Male	51	61	487	1.4	16	25.2
p.[Cys282Tyr];[His63Asp]	p.[Leu96Pro];[=]	Female	43	69	402	1.1	0	20.7

This change at position 168 of the HFE protein is predicted to be neutral by Align GVGD (score: C0; GV: 54.02-GD: 0.00), Mutation Taster (polymorphism; *P* value: 0.707) and SIFT (tolerated; score: 0.06), whereas it is predicted to be probably damaging by PolyPhen-2 (score: 0.980; HumVar model). More importantly, however, is perhaps the finding that glutamic acid 168 is part of a region that may contribute considerably to the interaction between HFE and transferrin receptor 1 (TfR1).[5]

We have previously hypothesized that the p.Leu96Pro missense mutation is enriched in French patients with hyperferritinemia and no mutations in genes commonly associated with hemochromatosis.[6] This mutation has been proven to decrease BMP6 secretion in opossum kidney (OK) cells.[3] We report here, for the first time, the case of a 43-year-old woman bearing both the p.[Cys282Tyr];[His63Asp] compound heterozygous *HFE* genotype and the p.[Leu96Pro];[=] heterozygous *BMP6* genotype (Table 1). The patient presented with a moderate degree of iron accumulation in the body as revealed by slightly elevated transferrin saturation (69%) and serum ferritin (402 μg/L) levels and also evaluated from the amount of iron removed by phlebotomy (1.1 g). No familial investigations were carried out to identify symptomatic or asymptomatic relatives with identical genotypes. Intrafamilial phenotypic heterogeneity has been observed in pedigrees with *BMP6* propeptide missense mutations.[3] On the other hand, subjects with the *HFE* p.[Cys282Tyr];[His63Asp] genotype and no comorbid factors dot not usually develop iron overload.[1] Both HFE and BMP6 modulate hepcidin synthesis in response to body iron status[7]. An interesting question is whether mild cases of iron overload can involve the two disease loci.

To conclude, our study highlights the benefit of investigating the whole *HFE* coding sequence in p.Cys282Tyr heterozygotes after exclusion of the most common causes of acquired hyperferritinemia. We report an unusual case of moderate iron overload associated with a complex *HFE* allele (p.[His63Asp;Glu168Gln]). This further indicates that screening p.His63Asp substitution in p.[Cys282Tyr];[His63Asp] heterozygotes with a well-defined iron overload phenotype has the potential to reveal rare *HFE* variants. If correctly interpreted, these uncommon variants may reduce diagnostic uncertainty and improve genetic counseling. Last but not least, we revealed disease-associated alleles in two different genes that interact within the same pathway. Further observations will help to determine whether mutations in *BMP6* may partially explain the large phenotypic heterogeneity observed in *HFE* p.[Cys282Tyr];[His63Asp] individuals and confer higher risk of iron overload.

ACKNOWLEDGMENTS

The authors sincerely thank the physicians from the French Blood Agency who agreed to participate to the EMSAI study. This work has been funded by the French Blood Agency (grants APR-2010 and APR-2013).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

G.LG. designed the study. K.U., C.K., and I.G. performed molecular analyses. G.LG., V.S., C.H., T.C., MC.M., and C.F. analyzed the data. G.LG. wrote the paper.

ORCID

Gerald Le Gac http://orcid.org/0000-0003-3236-7280

Kevin Uguen[1], Virginie Scotet[2,3], Chandran Ka[2,4,5], Isabelle Gourlaouen[2,3,5], Carine L'hostis[2,6], Marie-Christine Merour[3], Tania Cuppens[1,2], Claude Ferec[1,2,3,5*], Gerald Le Gac[1,2,3,4,5]

[1]*Université Bretagne Loire, Université de Bretagne Occidentale, IBSAM, Brest, France*

[2]*Inserm U1078, Brest, France*

[3]*Etablissement Français du Sang, Brest, France*

[4]*Laboratory of Excellence GR-Ex, Paris, France*

[5]*Laboratoire de Genetique Moleculaire et Histocompatibilité, CHRU de Brest, Hopital Morvan, Brest, France*

[6]*Association Gaetan Saleun, Brest, France*

Correspondence

G. Le Gac, Centre Hospitalier Regional Universitaire de Brest, Hopital Morvan, Bat 5bis, Laboratoire de Genetique Moleculaire et d'Histocompatibilité, 2 avenue Foch, 29200 Brest, France.

Email: gerald.legac@univ-brest.fr

Funding information

Grant sponsor: French Blood Agency; Grant sponsor(s): APR-2010 and APR-2013.

Abbreviation

AIR, amount of iron removed by phlebotomy

*On Behalf of the FREX Consortium (Genin E, Campion D, Dartigues JF, Deleuze JF, Lambert JC, Redon R, Ludwig T, Grenier-Boley B, Letort S, Lindenbaum P, Meyer V, Quenez O, Dina C, Bellenguez C, -Le Clezio CC, Giemza J, Chatel S, Ferec C, Le Marec H, Letenneur L, Nicolas G, Rouault K, Bacq D, Boland A, Lechner D)

REFERENCES

[1] European Association for the Study of the Liver. EASL clinical practice guidelines for HFE hemochromatosis. *J Hepatol*. 2010;53(1):3–22. https://doi.org/10.1016/j.jhep.2010.03.001.

[2] Barton JC, Edwards CQ, Acton RT. HFE gene: structure, function, mutations, and associated iron abnormalities. *Gene*. 2015;574(2):179–192. https://doi.org/10.1016/j.gene.2015.10.009.

[3] Daher R, Kannengiesser C, Houamel D, et al. Heterozygous mutations in BMP6 propeptide lead to inappropriate hepcidin synthesis and moderate iron overload in humans. *Gastroenterology*. 2016;150(3):672–683. https://doi.org/10.1053/j.gastro.2015.10.049.

[4] Menardi G, Perotti L, Prucca M, Martini S, Prandi G, Peano G. A very rare association of three mutations of the HFE gene for hemochromatosis. *Genet Test*. 2002;6(4):331–334. https://doi.org/10.1089/10906570260471895.

[5] Bennett MJ, Lebrón JA, Bjorkman PJ. Crystal structure of the hereditary haemochromatosis protein HFE complexed with transferrin receptor. *Nature*. 2000;403(6765):46–53. https://doi.org/10.1038/47417.

[6] Le Gac G, Gourlaouen I, Ka C, Férec C. The p.Leu96Pro missense mutation in the BMP6 gene is repeatedly associated with hyperferritinemia in patients of French origin. *Gastroenterology*. 2016;151(4):769–770. https://doi.org/10.1053/j.gastro.2016.03.054.

[7] Rishi G, Wallace DF, Subramaniam VN. Hepcidin: regulation of the master iron regulator. *Biosci Rep*. 2015;35(3):e00192. https://doi.org/10.1042/BSR20150014.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

Received: 29 June 2017 | Revised: 21 September 2017 | Accepted: 24 September 2017

DOI 10.1002/ajh.24918

Noninvasive monitoring of liver fibrosis in sickle cell disease: Longitudinal observation of a cohort of adult patients

To the Editor:

Acute vaso-occlusive events (VOCs) and liver iron overload (LIO), or viral hepatitis due to multiple blood transfusions, are the main factors in Sickle Cell Disease (SCD) associated with hepatic damage resulting in fibrosis/cirrhosis. There is a high frequency of cirrhosis (around 11%-14%) as a result of these complications based on autopsy findings at the time of death in patients with SCD.[1] Acute VOCs with hepatic damage can result with catastrophic consequences, including acute hepatic failure, and they may contribute to early mortality.[2,3] Unfortunately, due to the paucity of sizeable or controlled studies on sickle hepatopathy, as well as incomplete information regarding additional causes of liver disease, it is very difficult to evaluate the contribution of single factors in liver damage.

Hepatic function in 37 adult patients with SCD (68% HbSβ°, mean age 40 ± 14 years, 46% male, 13% non-Caucasian) who attended our center was retrospectively evaluated using a combination of biochemical markers (every six months), vibration-controlled transient elastography (VCTE; every two years), liver imaging (ultrasound, Magnetic Resonance Imaging [MRI]), and noninvasive iron measurement (MRI-T2*). Data were collected from 01 January 2002 to 31 December 2016. The chart review was approved by the local Ethics Committee. All participants gave

TABLE 1 Characteristics of 37 patients with sickle cell disease

Characteristics	Mean ± SD	Range	Normal values
Age (years)	40 ± 14	17–73	-
BMI (kg/m2)	23 ± 4	14.7–30.5	18.5–24.9
Hb (g/dL)	10.2 ± 1.2	8.4–13.4	14–17.5
AST (IU/L)	40 ± 18	14.2–106.2	4–37
ALT (IU/L)	32 ± 19	12–91	4–41
GGT (IU/L)	39 ± 32	10.5–163	8–61
Total bilirubin (mg/dL)	2.3 ± 1.3	0.63-6.2	0.1-1.0
Conjugated bilirubin (mg/dL)	0.7 ± 0.4	0.2-2.1	0.1-0.3
Ferritin (ng/mL)	1146 ± 1707	17–8094	30–400
ALP (IU/L)	95 ± 43	37–220	40–130
Albumin (g/L)	4.1 ± 0.4	3.4-4.8	3.5-5.2
Prothrombin time % (sec)	83 ± 10	54–100	70–120
LDH (IU/L)	784 ± 574	319–3565	135–250
Reticulocytes (%)	8.2 ± 5.5	1.8–23.3	0.5-2.5
HbF (%)	8 ± 8	0.3–29.7	<1.0
HbS (%)	61 ± 11	35–81	-
Liver stiffness (kPa)—total	9 ± 6	3.8–32.8	<8.7
Liver stiffness (kPa)—HbSS	8 ± 5	3.8–24.0	<8.7
Liver stiffness (kPa)—HbSβ°	9 ± 7	4.0–32.8	<8.7
MRI-T2* liver (msec)	10 ± 6	1.6–23.4	>6.3

Abbreviations: ALP, alkaline phosphatase; ALT, alanine transaminase; AST, aspartate transaminase; BMI, body mass index; GGT, γ-glutamyl transferase; Hb, hemoglobin; HbF, fetal hemoglobin; HbS, sickle hemoglobin; kPa, kiloPascal; LDH, lactate dehydrogenase; MRI, Magnetic Resonance Imaging; SD, standard deviation..

Annexe II : MACARON : a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data, Khan et al., 2018, publié dans Bioinformatics

Bioinformatics, 34(19), 2018, 3396–3398
doi: 10.1093/bioinformatics/bty382
Advance Access Publication Date: 3 May 2018
Applications Note

Sequence analysis

MACARON: a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data

Waqasuddin Khan[1,2], Ganapathi Varma Saripella[1,2], Thomas Ludwig[3], Tania Cuppens[3], Florian Thibord[1,2], FREX Consortium, Emmanuelle Génin[3], Jean-Francois Deleuze[4] and David-Alexandre Trégouët[1,2,*] on behalf of the GENMED Consortium

[1]Sorbonne Universités, UPMC Université Paris 06, INSERM UMR_S 1166, F-75013 Paris, France, [2]ICAN Institute for Cardiometabolism and Nutrition, F-75013 Paris, France, [3]INSERM U1078, Génétique, Génomique Fonctionnelle et Biotechnologies, Université de Bretagne Occidentale, CHU Brest, F-29238 Brest, France and [4]Centre National de Recherche en Génomique Humaine (CNRGH), Direction de la Recherche Fondamentale, CEA, Institut de Biologie François Jacob, F-91000 Evry, France

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received and revised on April 16, 2018; editorial decision on May 1, 2018; accepted on May 2, 2018

Abstract

Summary: Predicted deleteriousness of coding variants is a frequently used criterion to filter out variants detected in next-generation sequencing projects and to select candidates impacting on the risk of human diseases. Most available dedicated tools implement a base-to-base annotation approach that could be biased in presence of several variants in the same genetic codon. We here proposed the MACARON program that, from a standard VCF file, identifies, re-annotates and predicts the amino acid change resulting from multiple single nucleotide variants (SNVs) within the same genetic codon. Applied to the whole exome dataset of 573 individuals, MACARON identifies 114 situations where multiple SNVs within a genetic codon induce an amino acid change that is different from those predicted by standard single SNV annotation tool. Such events are not uncommon and deserve to be studied in sequencing projects with inconclusive findings.

Availability and implementation: MACARON is written in python with codes available on the GENMED website (www.genmed.fr).

Contact: david-alexandre.tregouet@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Variant annotation is a crucial step in whole genome/exome sequencing analyses aimed at identifying putative causal variants, especially in a clinical context (Ding *et al.*, 2014). For example, for a rare inherited disease, one often starts to filter out detected variants according to the anticipated mode of inheritance, the type of variations (e.g. synonymous, non-synonymous, stop gain/loss, splice, etc.), allele frequencies and their predicted deleteriousness. There is a plethora of annotation tools (Cingolani *et al.*, 2012; McLaren *et al.*, 2016; Yang and Wang, 2015) but most of them implement a base-to-base approach to annotate single-nucleotide variants (SNVs). However, the presence of several SNVs at the same locus, in particular within the same genetic codon, may bias annotations. For example, two synonymous SNVs in the same codon can generate a non-synonymous variation that would be missed by standard annotation tools. To our knowledge, there is only one program, MAC

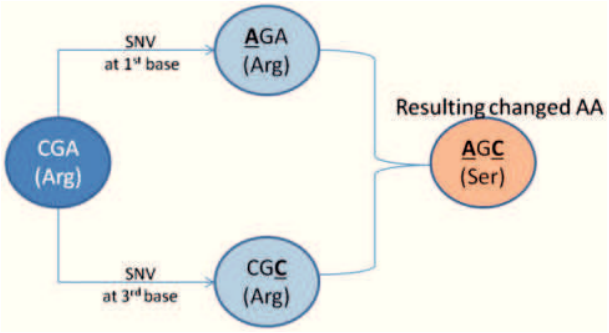


Fig. 1. Illustration of the impact of the presence of two single nucleotide variations within the same genetic codon on the resulting amino acid change

(Wei *et al.*, 2015), that accommodates multiple SNVs simultaneously. However, it is restricted to adjacent SNVs and cannot then properly address the situation when two SNVs affect the first and the third base of a genetic codon. In addition, it does not use the information on genetic code triplet structure. As a consequence, it considers the same way two SNVs affecting the adjacent bases of a genetic codon, and two SNVs affecting the last base of a codon and the first base of the next codon. To fill these gaps, we propose a simple python-based algorithm, MACARON (for **M**ulti-b**A**se **C**odon-**A**ssociated variant **R**e-annotati**ON**) to identify and to more accurately annotate multiple SNVs occurring within the same genetic codon (Fig. 1). We illustrate MACARON's relevance by an application to whole exome sequencing data of 573 subjects.

2 Implementation and application

2.1 Workflow

The overall algorithmic steps of MACARON are given below and illustrated as Supplementary Figure S1. The algorithm of MACARON is written in python language and can run on any LINUX/UNIX-like environment. Two pre-installed software, GATK (McKenna *et al.*, 2010) and SnpEff (Cingolani *et al.*, 2012) should be available for a complete run of MACARON. Briefly, MACARON starts with a VCF file as an input with no restriction on file format specifications. After identifying a list of candidate SNVs that occur within the same genetic codon along with their corrected amino acid changes, a second step consists in reading through the original BAM files to extract reads information and to confirm the presence of multiple SNVs on the same reads.

First, starting with a VCF file, MACARON utilizes GATK's VariationFiltration walker (Van der Auwera *et al.*, 2013) with parameters of –clusterSize 2 and –clusterWindowSize 3 followed by the SelectVariants tool to identify adjacent SNVs and SNVs that are 2 bps apart. Then, coding SNVs are selected based on the SnpEff functional annotation classes: SILENT, MISSENSE and NONSENSE (temp_file1). At the third step, SNVs that cluster within the same genetic codon are kept and new amino acid (AA) changes are written in temp_file2 and temp_file3. Next, clustered SNVs whose resulting AA changes are different from the original ones are stored in temp_file4. In case of a multi-sample VCF file, a scan is then performed on temp_file4 to identify clustered SNVs that are present in at least one individual. Results are stored in a final output text file containing all those SNVs identified within the same genetic codon and for which the allelic status is heterozygous or homozygous compared to the reference. At the final step, in order to confirm that identified clustered SNVs are harbored on the same reads, we used an in-house BASH-shell script (available with MACARON code) to read through the original BAM files that have been used for VCF file generation and to report the number of reads that harbor all variant alleles at the identified clustered SNVs. This script needs a subset of BAM files covering 50 bps over each clustered SNVs.

2.2 Results

MACARON was applied to the whole exome sequencing data of 573 healthy individuals as part of the FREX initiative in which 625 984 exonic SNVs were identified (Genin *et al.*, 2017). MACARON identified 114 multi-base affected codons in 194 participants. All identified affected codons were impacted by two SNVs (these were referred to as paired codon SNVs, pcSNVs) and no codon was identified that was simultaneously affected at all its 3 bases. From the identified pcSNVs, 83 were affecting codon positions 1 and 2, 23 codons were affected at positions 2 and 3 and the remaining 8 were affected at positions 1 and 3. Detailed distribution of the identified pcSNVs according to different criteria including allele frequencies, amino acid changes and predicted deleteriousness is given in Supplementary Table S1. Several observations could be made. For example, of these pcSNVs, 30 involved two rare [i.e. never reported or reported with minor allele frequency <0.01 in the gnomeAD database (Lek *et al.*, 2016)] SNVs, 15 involved one rare and one common SNV and 69 based on two common SNVs. These types of pcSNVs were referred to as 'double-rare', 'single-rare' and 'double-common' pcSNVs, respectively. The number of private (i.e. present in only one individual) pcSNVs were 16 (53%), 11 (~73%) and 3 (~4%) ~ among 'double-rare', 'single-rare' and 'double-common' pcSNVs, respectively. No pcSNV was generated from two synonymous SNVs but 26 were defined from one synonymous and one non-synonymous SNV. For 114 pcSNVs, the resulting amino acid change was different from the two original SNVs. Using the popular functional effect prediction tool SIFT (Ng and Henikoff, 2003), we observed that nine pcSNVs were predicted to be 'damaging' while the two original SNVs were predicted to be 'tolerated'. Conversely, two pcSNVs were predicted to be 'tolerated' or 'neutral' while the two original SNVs were predicted to be 'damaging'. For this application, MACARON took ~1 h on an Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60 GHz processor ×32 cores machine equipped with 64 GB of RAM on UBUNTU 16.04 LTS operating system to screen, re-annotate pcSNVs and validate them from BAM files.

3 Conclusion

MACARON is a new annotation tool for characterizing multiple SNVs within a same codon detected in WGS/WES studies. Its application to real data suggests that the frequency of pcSNVs is underappreciated and that inaccurate annotation of such genetic variations could contribute to explain inconclusive findings in DNA sequencing analyses.

Acknowledgements

Members of the GENMED and FREX consortia are listed in supplements.

Funding

This work was supported by the GENMED Laboratory of Excellence on Medical Genomics [ANR-10-LABX-0013 to WK, GV-S, FT] and the France Genomique National Infrastructure [ANR- 10-INBS-0009 to FREX consortium].

Conflict of Interest: none declared.

References

Cingolani,P. *et al*. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.

Ding,L. *et al*. (2014) Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet*., **15**, 556–570.

Genin,E. *et al*. (2017) The French Exome (FREX) Project: a population-based panel of exomes to help filter out common local variants. *Genet. Epidemiol*., **41**, 691–691.

Lek,M. *et al*. (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.

McKenna,A. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.

McLaren,W. *et al*. (2016) The ensembl variant effect predictor. *Genome Biol*., **17**, 122.

Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*., **31**, 3812–3814.

Van der Auwera,G.A. *et al*. (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinf*., **43**, 11.10. 1–11.10.33.

Wei,L. *et al*. (2015) MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics*, **16**, 569.

Yang,H. and Wang,K. (2015) Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc*., **10**, 1556–1566.

Annexe III : Conséquences des variations par Ensembl

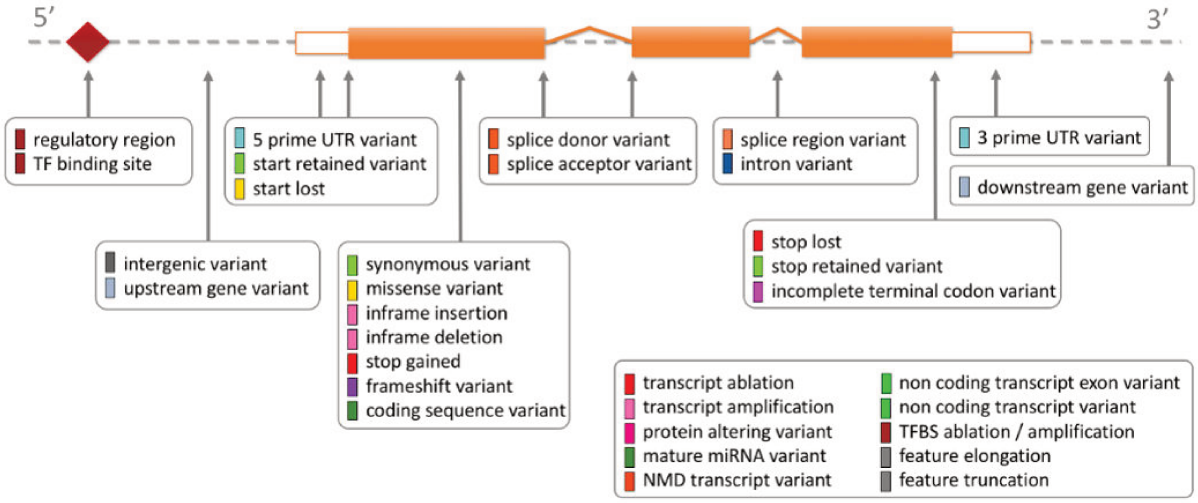


FIGURE 5.1: Ensemble des termes définissant la conséquence d'un variant utilisé par Ensembl.

Annexe IV : Séquençage CFTR des différentes lignées

J'ai vérifié par séquençage Sanger la séquence CFTR des lignées cellulaires pertinentes dans le cadre de l'étude de la mucoviscidose et qui sont utilisées au laboratoire. J'ai fait de même pour les différents plasmides portant l'ADNc du CFTR. J'ai regardé en particulier les positions qui nous ont intéressées dans le cadre de cette thèse à savoir les positions 470, 508 et 1027 (Figure 5.2).

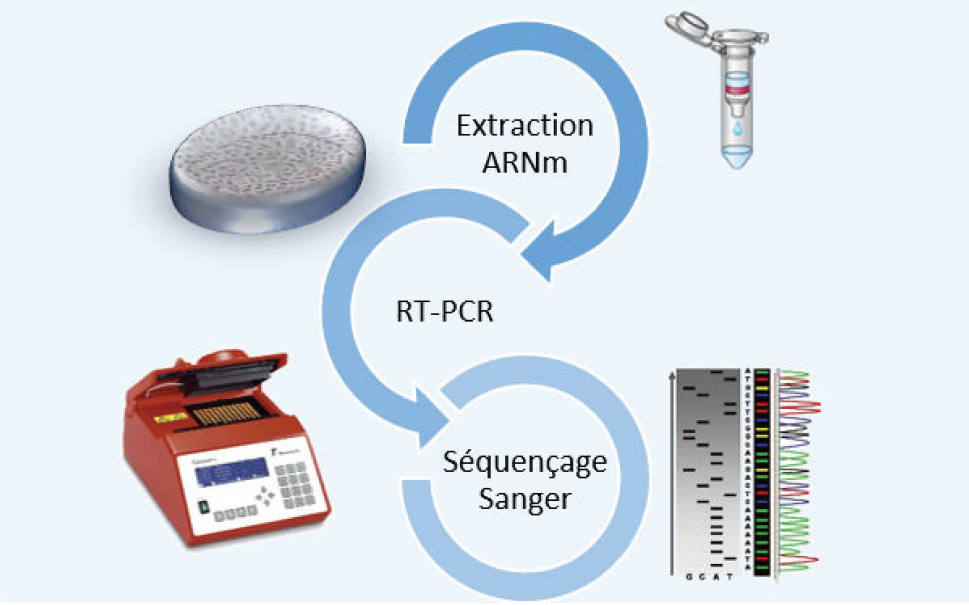


FIGURE 5.2: Protocole expérimentale utilisé pour séquencer l'ADNc du CFTR de différentes lignées.

J'ai pour cela mis en culture les lignées 16HBE, CFBE41o$^-$, CFBE41o$^-$/WT, CFBE41o$^-$/F508del. Ces lignées sont cultivées dans un milieu EMEM (Eagle's minimal essential medium) complété avec 10% de sérum veau fœtal et 1% de L-glutamine. La lignée 16HBE est issue de cellules de bronche d'un individu sain immortalisé. La lignée CFBE41o$^-$ est une lignée de bronche de patient homozygote p.F508del immortalisée. Afin d'étudier les différences entre des cellules WT et des cellules 508del, les cellules natives CFBE41o$^-$ ont été transfectées stablement par un virus SV40 avec un CFTR homozygote WT (CFBE41o$^-$/WT) ou un CFTR homozygote 508del (CFBE41o$^-$/F508del).

J'ai extrait l'ARN des différentes lignées avec le Kit NucleoSpin RNA Plus (Macherey Nagel), puis réalisé une reverse transcription des ARN messagers extraits par la SuperScript™ II RT. Une PCR d'amplification avec les amorces centrées sur la zone du CFTR portant la position 470 (c.1408G>A) et 508 (c.1521_1523delCTT) est réalisé et vérifiée par migration sur gel agarose 1% TBE 1x BET 5%. D'autres amorces ont été utilisées pour la position 1027. Un nettoyage enzymatique des produits de PCR permettant l'élimination des amorces et des dNTP non incorporés restants en fin de PCR est effectué par ExoSAP-IT. Avant d'être purifiés sur colonne autoseq et

séquencés par séquençage Sanger, les produits de PCR sont amplifiés à l'aide du kit BigDye™ Terminator Cycle Sequencing (ThermoFisher). Le séquençage a révélé que toutes les lignées portent une Isoleucine en position 1027. Les lignées CFBE41o$^-$ ont la 470M et la lignée 16HBE a la 470V (Table 5.1).

16HBE	470V	508F	1027I
CFBE41o$^-$	470M	508del	1027I
CFBE41o$^-$/ WT	470M	508F	1027I
CFBE41o$^-$/ 508del	470M	508del	1027I

TABLE 5.1: Résultat du séquençage des différentes lignées aux positions 470,508 et 1027

Lors de la vérification de l'amplification par migration sur gel agarose, un profil différent pour l'ADNc de la lignée 16HBE a été mis en évidence. Le plus petit produit de PCR observé pourrait représenter un produit d'épissage déjà décrit, résultant du saut d'exon 10 (Figure 5.3) [Chu et al., 1993, Claustres, 2005, Lee et al., 2012]. Ce phénomène serait dû à des différences du nombre de répétition de TG et de T (TGmTn) au niveau du site accepteur d'épissage de l'intron 9. Cette séquence est constituée de 5, 7 ou 9 Thymidines (5T, 7T ou 9T). Ainsi, deux types d'ARNm de CFTR, avec ou sans exon 10. La proportion de transcrits délétés est inversement corrélée avec le nombre de Tn. La protéine produite par le transcrit excisé de l'exon 10 n'est pas fonctionnelle.

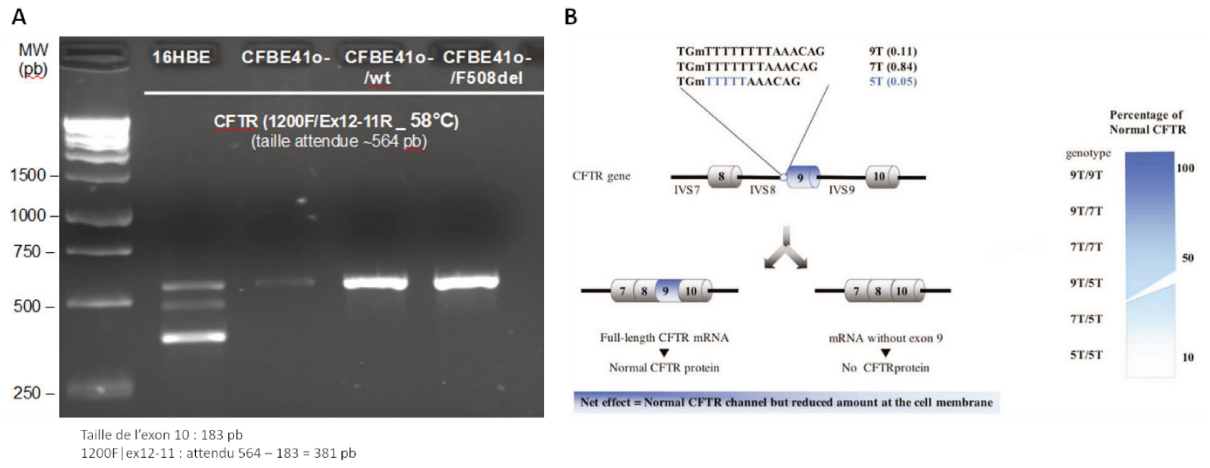


FIGURE 5.3: Représentation de l'excision de l'exon 10 (exon 9 dans l'ancienne annotation). A. Migration sur gel d'agarose des produits de PCR des quatre lignées cellulaires. Le produit attendu est de 564pb. B. Dans la population générale, trois allèles se trouvent au locus Tn, 5T (5% des allèles), 7T (84%) et 9T (11%). Les transcrits porteurs de cinq Thymidines (5T) à ce locus présentent des taux élevés de saut d'exon 10, tandis que ceux qui contiennent sept ou neuf thymidines (7T et 9T respectivement) présentent des taux plus faibles (d'après Claustres, 2005).

Annexe V : Effet de la p.Val470Met sur la protéine CFTR muté p.Gly551Asp

Le variant génétique qui transforme la Glycine en Acide Aspartique en position 551 (551D) est présent sur plus de 4% des allèles des patients atteint de mucoviscidose. Il est associé en cis avec le variant qui donne une Méthionine en position 470, suggérant que, tout comme la délétion p.Phe508Del, ce variant est apparu sur un chromosome portant 470M. En effet, le déséquilibre de liaison est assez fort dans cette région et les individus porteurs de 551D dans le panel de référence français (projet FrEx) ont une Méthionine en position 470. J'ai donc réalisé les mêmes expériences qu'avec le 508del, pour tester si la Valine ou la Méthionine modifiait l'expression et la fonction du canal CFTR. La mutation 551D est une mutation de classe III. Elle est présente à la membrane cellulaire mais n'est pas fonctionnel. Son profil électrophorétique est le même que la protéine CFTR-WT. Le potentiateur VX-770 permet d'activer le canal CFTR-551D augmentant ainsi sa fonction [Van Goor et al., 2009, Ramsey et al., 2011]. Les résultats ont montré qu'il n'y avait ni de différence de profil électrophorétique ni de différence de fonction entre la protéine CFTR-470V-551D et la protéine CFTR-470M-551D (Figure 5.4).

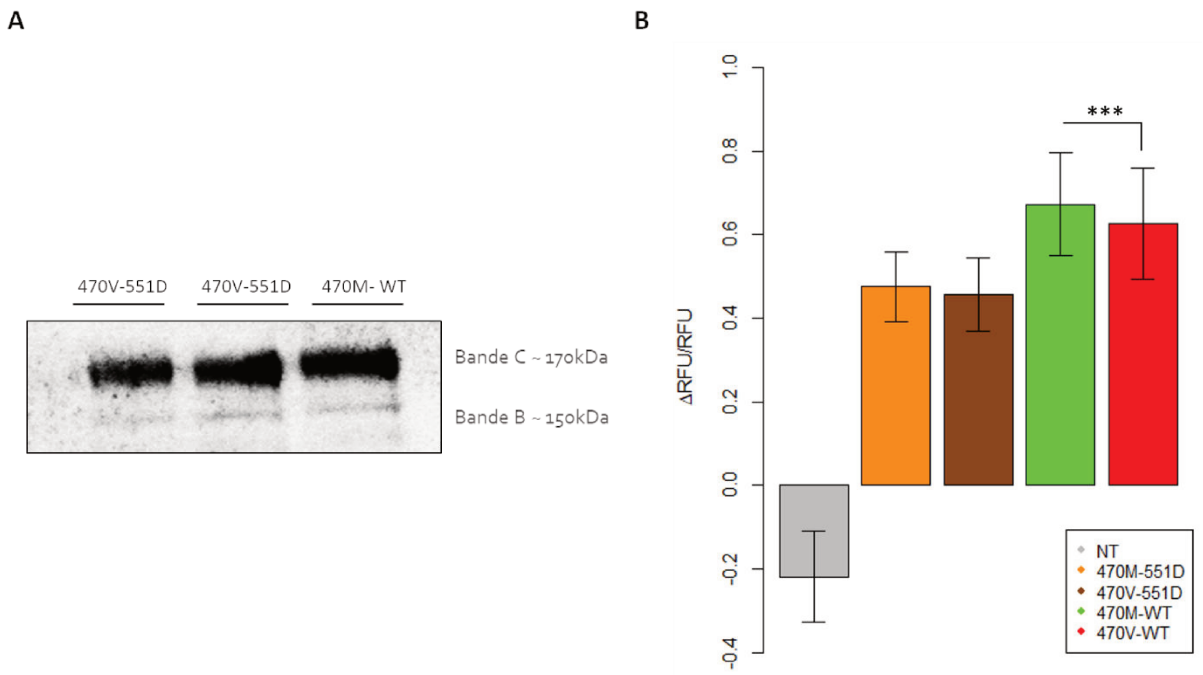


FIGURE 5.4: Analyse par Western blot de l'expression (A) et par FLIPR de la fonction (B) de la protéine CFTR dans des cellules HEK293T . A. Cellules HEK293T transfectées exprimant le CFTR-470M-551D ou le CFTR-470V-551D (N=4) B. Différence d'intensité de fluorescence après l'ajout de l'activateur (ΔRFU = RFUactivée - RFUbasale) normalisée par la fluorescence basale (RFU). Le canal CFTR est activé par la FSK et le VX-770 (1µM) (N=42). Les astérisques indiquent une significativité statistique en utilisant un modèle mixte (* p <0,05, ** p <0,01, *** p <0,001).

Titre : Conséquences du contexte haplotypique sur la fonction des protéines : application à la mucoviscidose

Mots clés : Outil bioinformatique, Contexte haplotypique, Visualisation, Mucoviscidose, Fonction protéique

Résumé : Notre génome contient des centaines de milliers de variants génétiques, qui pour la plupart, n'ont aucun impact sur notre santé. Après séquençage, il faut les filtrer pour ne conserver que ceux qui sont potentiellement impliquées dans une maladie. On utilise des annotateurs qui prédisent l'impact des variants. Ces prédictions sont faites sans tenir compte des variants en cis dans le même gène. Pourtant, des variants neutres peuvent, lorsqu'ils sont réunis chez un individu, devenir délétères. J'ai donc développé l'outil bioinformatique GEMPROT qui permet de visualiser l'effet des variants génétiques sur la séquence protéique et de mettre en évidence les combinaisons de variants touchant un même domaine fonctionnel.

J'ai ensuite étudié l'impact de deux variants associés à la p.Phe508del (508del) sur la protéine CFTR. Le variant p.Val470M est présent sur tous les haplotypes portant la délétion mais pas sur la séquence de référence, qui est généralement utilisée pour la construction de plasmides. Nous avons montré des différences de fonction de la protéine CFTR selon l'acide aminé en position 470. La fonction est augmentée avec une Valine et il convient donc de s'assurer, lors de la construction de plasmides, que le contexte haplotypique des variants étudiés est bien respecté. Le variant p.Ile1027Thr conduit à une dégradation de la fonction de la protéine 508del. Ce variant n'est présent que sur une partie des haplotypes 508del et pourrait donc avoir un effet modificateur de l'expression de la délétion. En conclusion, nous montrons l'importance de la prise en compte des contextes haplotypiques dans l'étude des maladies et proposons un outil bioinformatique pour le faire.

Title : Consequences of the haplotype context on protein function: application to cystic fibrosis

Keywords : Bioinformatic tool, Haplotype context, Visualization, Cystic Fibrosis, Protein function

Abstract : We all carry hundreds of thousands genetic variations in our genome that, for the most of them, have no impact on our health. After sequencing, they must be filtered to only retain those potentially involved in a disease. We use annotators that predict the impact of variants. These predictions are done for each variant taken independently without considering cis variants in the same gene. However, neutral variants can become deleterious when associated together. I have developed the bioinformatics tool GEMPROT, which makes it possible to visualize the effect of genetic variants on the protein sequence and to highlight combinations of variants affecting the same functional domain.

I then studied the impact of two variants associated with p.Phe508del (508del) on CFTR protein function. The variant p.Val470M is present on all carrying deletion haplotypes but not on the reference sequence, which is generally used for the construction of plasmids. We have shown differences in the function of the mutated CFTR protein 508del according to the amino acid at position 470. The function is increased with a Valine and it is therefore necessary to ensure, when constructing plasmids, that the haplotype context of the studied variants is well respected. The variant p.Ile1027Thr leads to a degradation of the function of the 508del protein. This variant is present only on a portion of the 508del haplotypes and could therefore have a modifying effect on deletion expression. In conclusion, we show the importance of considering haplotype contexts in the diseases studies and propose a bioinformatics tool to do so.