UNIVERSITÉ GRENOBLE-ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE-ALPES

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : ?

Présentée par

Thomas Karaouzene

Thèse dirigée par **Pierre Ray**

Thèse co-dirigée par **Nicolas Thierry-Mieg**

préparée au sein du laboratoire
et de l'école doctorale **"Ingénierie de la Santé, de la Cognition et Environnement" (EDISCE)**

Écrire le titre de la thèse ici

Thèse soutenue publiquement le 31 octobre 2017,
devant le jury composé de :

Université Grenoble Alpes

Préface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Delete line 6 if you only have one advisor

Remerciements

Résumé

Chapitre 2

Introduction

Chapitre 3

Investigation génétique et physiologique de la globozoospermie

Chapitre 4

Mise en place d'une stratégie pour l'analyse des données exomiques – application en recherche clinique

4.1 Intro

4.2 Résultats

Dans cette partie, nous allons détailler les résultats de l'analyse des données de WES de 75 patients tous atteins d'un phénotype d'infertilité. Ces études seront séparées en deux parties distinctes, la première se concentrera sur l'études de ... familles incluant 13 de ces patients. Le seconde portera sur l'analyse des 62 patients restant étant tous non-apparentés et présentant un phénotype MMAF.

4.2.1 Description de la pipeline

Après avoir été séquencés les données receuillies pour ces patients sont procéssés au sein de la même pipeline d'analyse qui comprend quatre étapes allant de l'alignement de *reads* au filtrage des variants :

1. **L'alignement** : L'alignement des *reads* le long du génome de référence (hg19 / GHRC37) est effectué par le logiciel MAGIC (Su et al., 2014). Afin d'écarté toute ambiguité au moment de l'interprétation de l'alignement, l'intégralité des *reads* dupliqués et / ou s'alignant à plusieurs zones du génome seront filtrés et ne seront donc pas pris en compte pour l'ensemble des analyses en aval. Suite à celà„ MAGIC va produire quatre comptages pour chaque position couverte du génome : R+, V+, R- et V- :

   a. **R+ et R-** : Ces deux comptages correspondent au nombre de *reads forward* (+) et *reverse* (-) sur lesquels est observé l'allèle de **référence** (R) à une position donnée.

   b. **V+ et V-** : À l'inverse de R+ et R-, ces comptages correspondent au nombre de *reads forward* et *reverse* sur lesquels est observé un allèle de **variant** (V) à une position donnée.

2. **L'appel des variants** : Comme nous l'avons vu plus tôt, il est fortement conseillé d'effectuer l'appel des variants en tenant compte de l'aligneur choisi (Nielsen, Paul, Albrechtsen, & Song, 2011, M. A. DePristo et al. (2011), Lunter & Goodson (2011)). C'est pourquoi, nous avons conçu notre propre algorithme d'appel des variants spécialement conçu pour l'analyse des données de MAGIC. Ainsi, l'appel des variants sera directement basé sur les quatre comptages vus précédemment. Tout d'abord, les positions ayant une couverture < 10 sur l'un des deux *strands* sera considérée comme de faible qualité, celles ayant une couverture < 10 sur les deux *strands* seront exclus. Ensuite pour chaque variant, des appels indépendants seront effectués pour chaque *strand*. L'appel final sera une synthèse de ces deux appels où seul les cas où ces deux appels sont concordants seront considérés comme de bonne qualité.

3. **L'annotation** : Chaque variant retenu sera ensuite annoté tout d'abord par le logiciel *variant effect predictor* (VEP) (W. McLaren et al., 2016) qui nous indiquera pour chaque variant la conséquence que celui-ci aura sur la séquence codante de l'ensemble des transcrits Ensembl qu'il chevauche (**Figure :** 4.1) (**Table :** 4.1). Ensite, nous ajoutons récuperons pour chaque gène son expression tissulaire en nous basant sur le projet Illumina BodyMap [TODO ref] qui recanse les données RNAseq des gènes humains pour 16 tissus différents. Suite à cela nous ajoutons, lorsque celle-ci est disponible, la fréquence du variant dans les bases de données ExAC (Lek et al., 2016), ESP600 [TODO] et 1000Genomes [TODO] donnant ainsi une estimation de sa fréquence dans la population générale. De même, la particularité de ce pipeline est qu'elle conserve l'ensemble des variants identifiés dans les études effectuées précédemment permettant d'ajouter aux annotations la fréquence d'un variant chez les individus déjà séquencé et donc la fréquence d'un variant dans chaque phénotype étudié créant ainsi une base de données interne qui pourra servir de contrôle dans les études ultérieur.

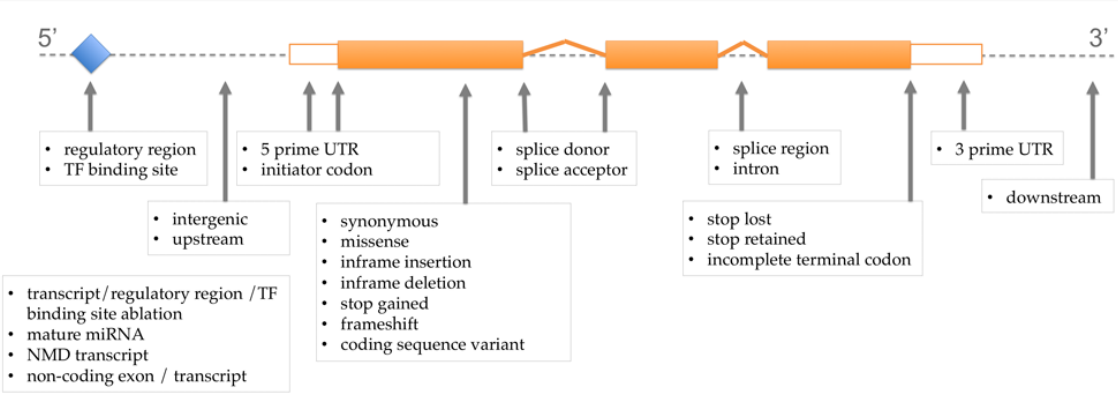


Figure 4.1 – Listes des différentes conséquences prédites par VEP et leur positionnement sur le transcrit d'après [VEP site](http ://www.ensembl.org/info/genome/variation/consequences.jpg)

4. **Le filtrage des variants** : L'étape de filtrage est extrêmement importante si l'on souhaite analyser de manière efficace les données provenant de WES. C'est pourquoi elle occupe une place importante dans notre pipeline. L'intégralité des paramètres de cette étape peuvent être modifier par l'utilisateur de sorte à faire correspondre les critères de filtre aux besoins de l'étude. Afin de rendre son utilisation le plus efficace possible, nous avons souhaité définir des paramètres par défauts pertinent dans la plupart des études de séquençage exomique de sorte que à moins que le contraire ne soit spécifié, seul les variants impactant les transcrits codant pour une protéine sont conservés. De même les variants synonymes ou affectant les séquences UTRs sont filtrés ainsi que les variants ayant une fréquence $\geq 1\%$ dans les bases dans l'une des bases données (ExAC, ESP6500 ou 1KH). Aussi, pour un phénotype donné, l'ensemble des variants observés chez les individus étudiés présentant un phénotype différent sont de même enlevés de la liste finale.

Table 4.1 – Liste simplifiée des conséquences prédites par VEP avec leur description et impact associée

VEP consequence	VEP impact	Description
Splice acceptor / donor	HIGH	A splice variant that changes the 2 base region at the 3' / 5' end of an intron
Stop gained	HIGH	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript
Frameshift	HIGH	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three
Stop lost	HIGH	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
Start lost	HIGH	A codon variant that changes at least one base of the canonical start codo
Inframe insertion / deletion	MODERATE	An inframe non synonymous variant that inserts / deletes bases into in the coding sequenc
Missense	MODERATE	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved
Splice region	LOW	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron
Stop retained	LOW	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains
Synonymous	LOW	A sequence variant where there is no resulting change to the encoded amino acid
5 / 3 prime UTR	MODIFIER	A UTR variant of the 5' / 3' UTR
Intron	MODIFIER	A transcript variant occurring within an intron
NMD transcript	MODIFIER	A variant in a transcript that is the target of NMD
Non coding transcript	MODIFIER	A transcript variant of a non coding RNA gene

4.2.2 Utilisation du pipeline dans des cas familiaux :

Description des familles

Dans cette partie, je me concentre sur l'analyse bioinformatique des résultats des séquençages exomiques effectués entre 2012 et 2014 de 13 individus infertiles provenant de 6 familles différentes. Parmi celles-ci, 3 phénotypes différents ont été observés :

1. **L'Azoospermie :** Comme nous avons pu le voir, l'azoospermie est un phénotype d'infertilité masculine caractérisé par l'absence de spermatozoïde dans l'éjaculât

2. **Échec de fécondation :** Ce phénotype d'infertilité se caractérise par l'incapacité des spermatozoïdes à féconder l'ovocyte.

3. **MMAF** : Le syndrome MMAF (*multiple morphological abnormalities of the sperm flagella*) caractérise comme son nom l'indique les patients présentant une majorité de spermatozoïdes atteins par une mosaïque d'anomalie morphologique du flagelle.

Parmi ces 6 chacune composée de 2 à 3 frères, 3 d'entre elles présentent un historique de consanguinité, les parents étant soit cousins germains, pour les familles ... et ..., soit cousins au second degré, pour la famille ... . La consanguinité favorisant la transmission de variants à l'état homozygote, nous avons décidé, dans un premiers temps de concentrer nos analyses uniquement sur les variants (SNVs et indels) homozygotes pour l'ensemble des familles. Pour les 3 familles n'ayant pas d'historique de consanguinité, ce choix nous permet de réduire la liste des variants candidats de sorte à faciliter les analyses. L'études des variants hétérozygotes sera effectuée *a posteriori* pour les familles dont la cause génétique du phénotype n'a pas pu être identifiée en se limitant aux variants homozygotes. Un récapitulatif des familles et de leur phénotype est disponible dans la table 4.2.

Table 4.2 – Tableau récapitulatif des familles séquencées et de leur phénotype

Family	Consanguinity	Individuals	Phenotype	Year	Place
AZ	Yes	AZ1, AZ2	Azoospermia	2012	Mount Sinai Institut
FF	Yes	FF1, FF2	Fertilization failure	2014	Genoscope (Evry)
MMAF1	No	MMAF1.1, MMAF1.2	MMAF	2014	Genoscope (Evry)
MMAF2	Yes	MMAF2.1, MMAF2.2	MMAF	2014	Genoscope (Evry)
MMAF3	No	MMAF3.1, MMAF3.2	MMAF	2014	Genoscope (Evry)
MMAF4	No	MMAF4.1, MMAF4.2, MMAF4.3	MMAF	2014	Genoscope (Evry)

Résultats des exomes

Résultat de l'alignement Pour rappel, l'alignement consiste à repositionner l'ensemble des *reads* générés au cours de l'étape de séquençage le long d'un génome de référence.

La quantité de *reads* composant les exomes de chaque individu peut varier en fonction de plusieurs paramètres et n'est donc pas égale pour chaque patient bien que l'ordre de grandeur reste le même avec une médiane de 91438630 *reads*. Seuls les deux frères AZ1 et AZ2 se distinguent avec près de 3 fois plus de *reads* que les autres patients. Cette différence peut être expliqué car ces deux patients sont les deux seuls à voir été séquencé au Mount Sinaï Institut or leur protocole d'amplification précédent le séquençage contient un nombre de cycles de PCR supérieur à ceux appliqué au Génopole d'Évry où ont été séquencé les autres patients. Il faut noter que ce nombre plus important de *reads* n'est en rien le reflet d'une meilleure qualité. En effet, ce nombre important de *reads* est causé par une grande quantité de *reads* dupliqués qui seront pour la plupart filtrés au cours des analyses ulterieure (**Table :** 4.2, **Figure :** 4.2 - **A**).

L'ensemble de nos exomes ayant été réalisés en *paired-end*, les deux extrémités de chaque fragment sont séquencées. Chaque *end* d'un même *read* peut donc être considéré comme un *read* à part entière qui sont alignées **indépendamment** le long du génome de référence. L'information fournit par le *paired-end* n'étant utilisé qu'à *posteriori* en tant que critère qualité. La première étape du contrôle qualité des *reads* consiste filtrer les *reads* ne s'étant pas aligné sur le génome. Ces *reads* sont extrêmement minoritaires puisqu'ils ne représentent qu'entre 1.2 et 5.5 % des *reads* de nos individus (**Figure :** 4.2 - **B**).

Une fois cela fait, nous vérifions la "compatibilité" des deux *ends* composant chacun des *reads* s'étant correctement alignés. Un *reads* est dit compatible lorsque les deux *ends* qui le composent s'alignent face à face (une sur le *strand* + et l'autre sur le *strand* -) et couvrent une zone ne faisant pas plus de 3 fois la taille médiane de l'insert. Les *reads* dont les deux *ends* se sont alignées mais ne remplissant pas ces conditions seront dit "Non compatible", ceux dont une seule des deux *ends* s'est alignés seront appelés "orphelins". Dans nos analyses, seuls les *reads* compatibles sont conservés, c'est à dire environs 89.5 % des *reads* s'étant correctement alignés. (**Figure :** 4.2 - **C**).

La dernière étape de ce contrôle-qualité consiste à analyser le nombre de site auxquels se sont alignés les *reads*. En effet, certaine zone du génome étant dupliqué, l'une des problématiques des *short-reads* est qu'il est possible que ceux-ci s'alignent à plusieurs régions différentes du génome. Afin d'éviter toute ambiguïté, seul ceux s'étant aligné sur un site unique sont conservés pour la suite des analyses. Ces *reads* représente entre 92.3 et 96.9 % des *reads* ayant passé les précédents filtres (**Figure :** 4.2 - **C**).

Les *reads* ayant passé l'ensemble des critères qualité mentionnés précédemment seront ensuite utilisés pour effectuer l'appel des variants.

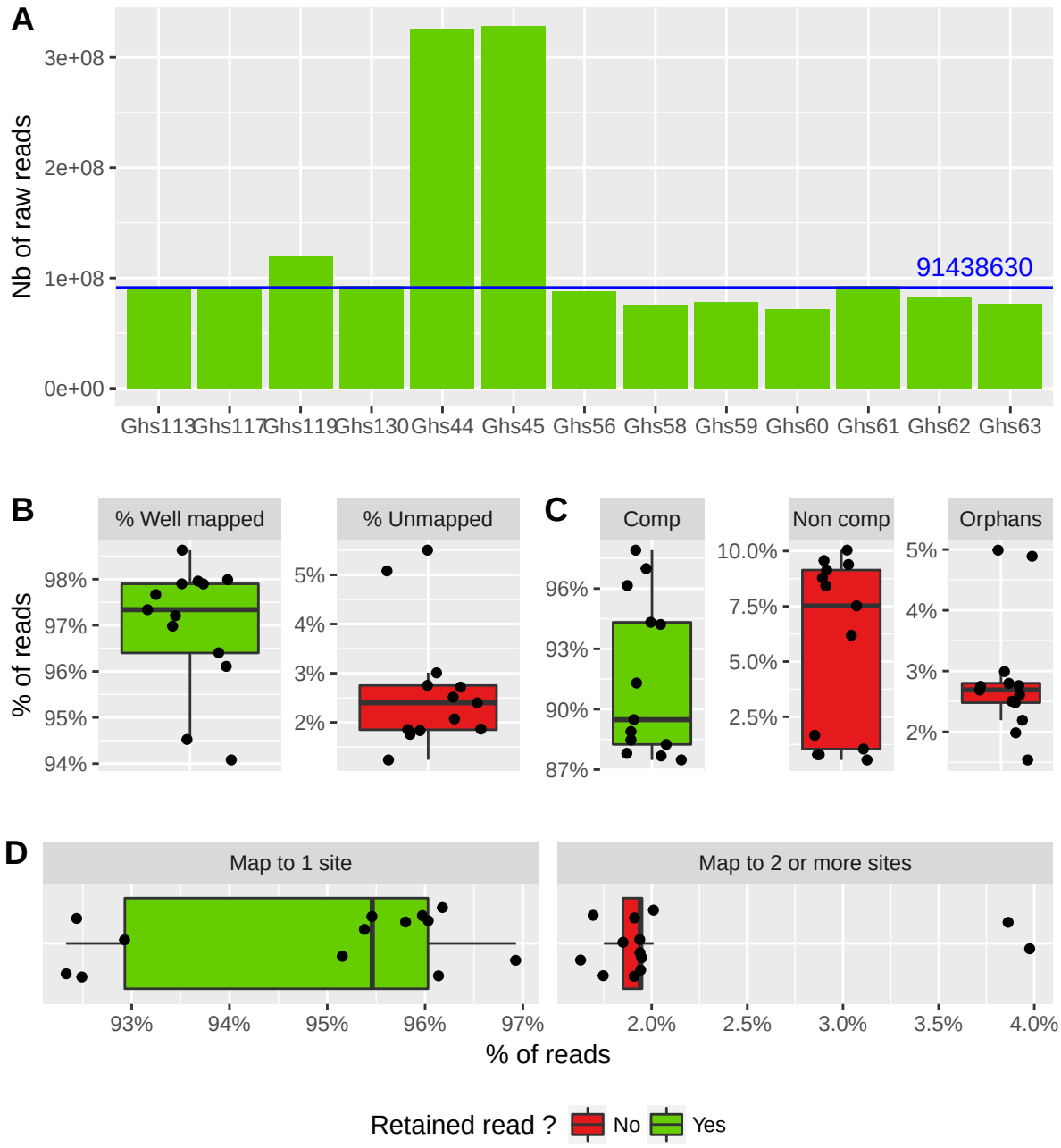


Figure 4.2 – Processus simplifié du contrôle qualité des *reads* : Pour chacun des graphiques, les *reads* représentés en vert sont conservés tandis que ceux en rouge sont filtrés. **A** : Quantité de *reads* bruts générés pour chaque patient au cours de l'étape de séquençage. La médiane des *reads* est représentée en bleue. **B** : Pourcentage pour chaque individu de *reads* s'étant aligné correctement et ne s'étant pas alignés sur le génome de référence. **C** : Distribution pour chaque patient des *reads* compatibles (Comp), non compatibles (Non comp) et orphelins (Orphans). **D** : Présentation pour chaque *reads* du nombre de site auxquels ils s'alignent

Résultat de l'appel des variants Comme dit précédemment, l'appel des variants fait suite à l'alignement et consiste à comparer la séquence d'un individu avec celle d'un génome de référence afin d'en relever les différences. La particularité de notre algorithme d'appel est d'effectuer pour chaque position deux appels indépendants Le premier sera effectué en utilisant uniquement les *reads forward* et le second le *reads reverse*. Encore une fois, plusieurs filtres sont appliqués de sorte à conserver uniquement les variants les plus qualitatifs.

Tout d'abord, nos appels sont classés en trois catégories :

1. **Les appels *double strand* (DS) :** Qualifie les positions ayant une couverture ≥ 10 sur **les deux** strands. Ces appels sont ceux sont ceux ayant la meilleure qualité
2. **Les appels *single strand* (SS) :** Ces appels définissent les positions pour lesquels **un des deux** *strands* présentent une couverture ≤ 10. Dans ce cas, ce *strand* est ignoré et l'appel est effectué uniquement en utilisant le second *strand*.

3. **Les appels *non strand* (NS) :** Les positions NS sont celles pour lesquelles la couverture est ≤ 10 sur **les deux** strands. Aucun appel n'est effectué à ces positions.

Dans nos données, les appels SS sont majoritaires et représentent environ 48.1 % de nos appels (contre 35.6 % d'appels DS). Au vus de l'importance de ces appels, nous avons fait le choix de les conserver afin de ne pas filtrer une quantité trop importante de données. Ces appels seront cependant considérés comme étant de faible qualité, de fait, leurs analyses et interprétation seront plus précautionneuses En revanche, au vus de la trop grande incertitude de l'appel des variants NS, ceux-ci sont systématiquement filtrés éliminant ainsi entre 10.3 et 18.7 % des positions appelées (**Figure :** 4.3 - **A**).

Un second filtre est appliqué aux variants ayant été précédemment appelés DS. Celui-ci consiste à comparer les appels effectués indépendamment sur chacune des deux *ends* et à vérifier leur concordance, c'est à dire que les deux appels soit identique. Les appels discordant et ambigus sont filtrés, soit environ 86.3 % des variants DS. Il est intéressant de noter que bien que les variants *single strand* (SS) soient conservés, on peut s'attendre à ce qu'environ 13.7 % de ceux-ci soient aberrants, ceux-ci n'ayant pu subir le même contrôle que les SS (**Figure :** 4.3 - **B**).

Pour l'ensemble des variants ayant passé les filtres énoncés ci-dessus, c'est à dire les variants SS et les variants DS avec appels concordants, le génotype est déterminé en fonction du pourcentage de *reads* portant le variant à cette position. Par exemple, si à une position donnée, 0% des *reads* portent un variant, l'individu sera appelé "Homozygote référence", si 50% des *reads* sont porteurs d'un variant, l'appel sera "hétérozygote" et si 100% des *reads* portent un variant, l'appel sera "Homozygote variant". Ainsi, pour chaque individu nous avons pu établir une liste de SNVs et d'indels avec leur génotype associé. Pour chacun de nos 13 patients les ordres de

grandeur du nombre de variants appelés sont identique. Ainsi pour chaque patient nous avons appelés environ 43670 variants hétérozygotes (41044 SNVs et 2626 indels) et 65040 variants homozygotes (32520 SNVs et 1809 indels) (**Figure :** 4.3 - **C**).

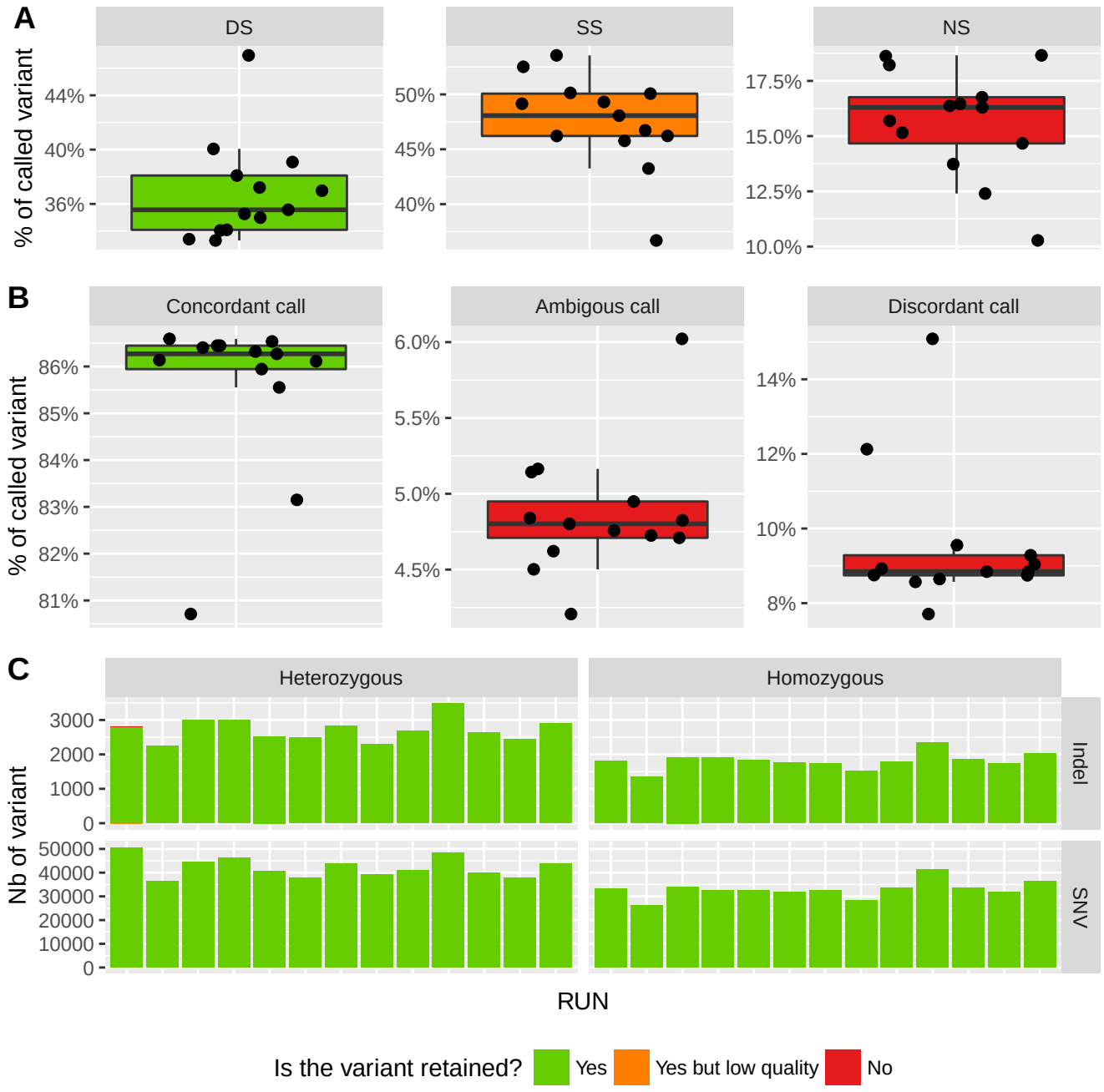


Figure 4.3 – Contrôle qualité des variants appelés : Pour chacun des graphiques, les variants représentés en vert et en orange sont conservés tandis que ceux en rouge sont filtrés. **A** : Distribution du *stranding* des appels pour chaque patient. **B** : Comparaison des appels entre les deux *ends* des variants appelés DS. **C** : Distribution des SNVs et indels en fonction de leur génotype pour chaque patients (représentés par une barre

Résultats de l'annotation L'annotation des variants appelés consiste à ajouter un maximum d'informations sur les variants. Ces informations seront ensuite utilisées afin de filtrer et / ou prioriser les variants. Dans ces analyses nous avons utilisé le logiciel *Variant Effect Predictor* (VEP) (W. McLaren et al., 2016) qui va à la fois prédire l'effet qu'auront ces variants sur l'ensemble des transcrits (et gènes) qu'ils chevauchent, ajouter, lorsqu'elle est disponible, la fréquence de chacun de ces variants dans les bases de données ExAC, 1000Genomes (1KG) et ESP6500. Pour finir VEP nous permettra de connaitre les prédictions de pathogénicités fournies par SIFT et PolyPhen pour les variants faux-sens.

Après avoir annoté nos variants par VEP, nous avons pu constater que pour chaque patient 24975 gènes sont en moyenne affecté par au moins un variant pour en moyenne 122735 transcrits (soit environ 5 transcrits par gènes). Il faut noter que parmi ces gènes se trouvent à la fois des gènes codant pour des protéine **et** d'autres non codant (**Figure :** 4.4 - **A**).

Chaque variant affectera l'ensemble des transcrits qu'il chevauche, ainsi un même variant pourra impacter plusieurs transcrits. Ces impacts sont ensuite classés par VEP en quatre catégories qui sont, de la plus délétère à la moins délétère : HIGH, MODERATE, LOW, MODIFIER (**Table :**4.1). Comme attendu, les variants ayant un impact tronquant se retrouvent être les moins fréquent chez chacun de nos patients. Ceci est d'autant plus flagrant pour l'impact HIGH qui regroupe, entre autres, les variants créant un codon stop ou encore ceux causant un décalage du cadre de lecture (**Table :**4.1), se retrouvent en quantité extrêmement faible puisqu'ils ne représentent en moyenne que 0.15 % des variants, soit une moyenne de 466 hétérozygotes et 370 homozygotes par patient) (**Figure :** 4.4 - **B**).

Parmi ces variants, certains étaient déjà recensés dans une des trois base donnée (ExAC, ESP et 1KG). Ainsi, on peut observer qu'entre 38.6 et 55.5 % de nos variant étaient listés dans ExAC et entre 33.1 et 43.8 % dans ESP. En revanche environ 87.1 % d'entre eux sont recensés dans 1KG (**Figure :** 4.4 - **C**) (À discuter !!!!!).

(À discuter !!!!!) (**Figure :** 4.4 - **D**)

LES FIGURES SUR LA FRÉQUENCE SONT À DISCUTER CAR LEUR INTERPRÉTATION ME LAISSE PERPLEX (SURTOUT LA PROPORTION DE NOS VARIANTS PRÉSENTS DANS 1KG)

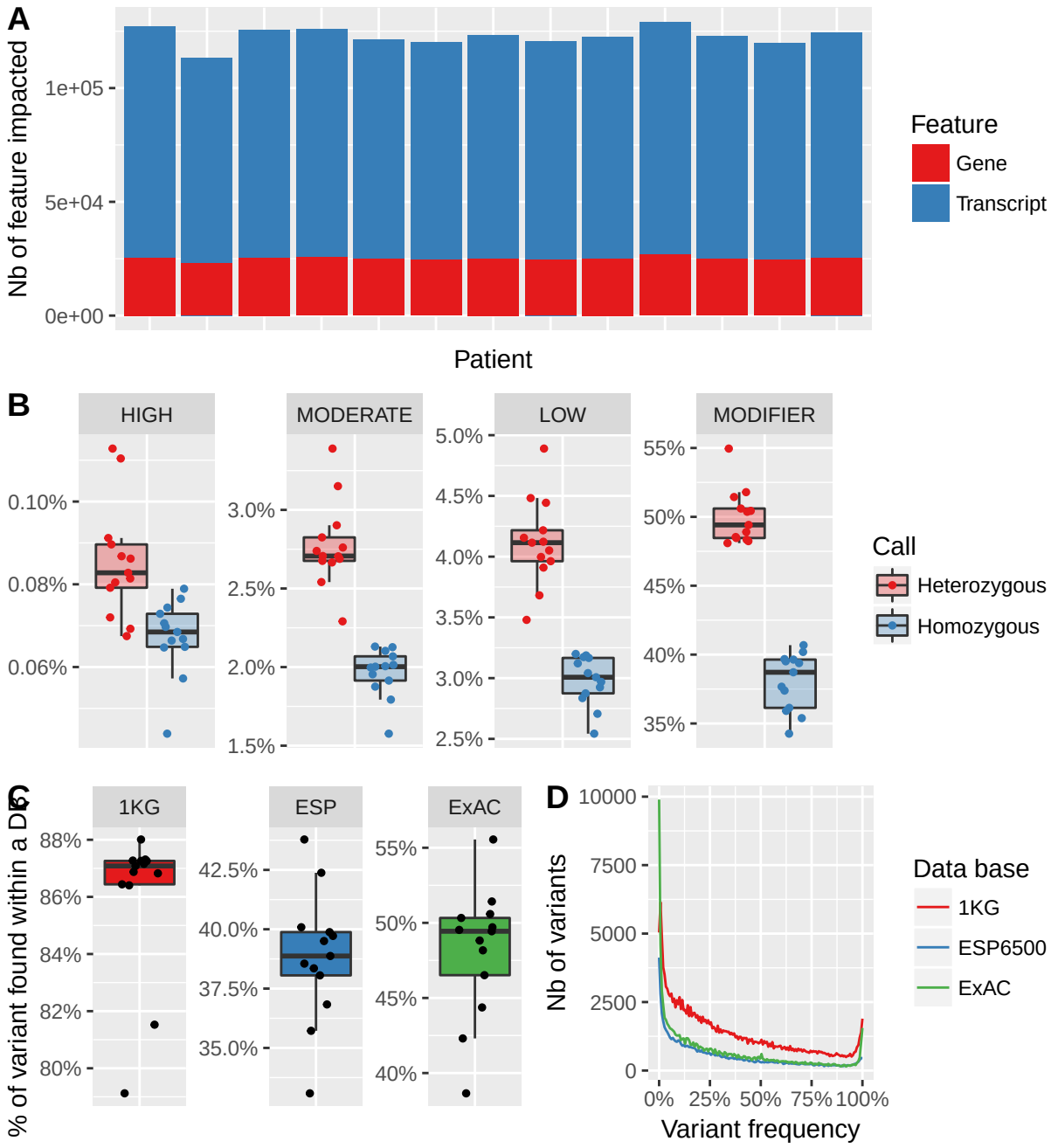


Figure 4.4 – Annotation des variants par VEP : **A** : Quantification du nombre de gènes (en bleu) / transcrits (en rose) impactés par au moins un variant pour chaque patient chacun représentés par une barre. **B** : Distribution des impacts HIGH MODERATE LOW et MODIFIER en fonction des patients et du génotype du variant. **C** : Pourcentage des variants retrouvés au sein des trois bases de données : ExAC, ESP et 1KG. **D** : Distribution des fréquences de nos variants au sein des trois bases de données : ExAC, ESP et 1KG

Résultats du filtrage Les étapes précédentes nous ont permis de mettre en évidence pour chaque patient une liste de variants passant l'ensemble de nos critères qualités. Ces variants ont dès lors pu être annotés nous permettant notamment d'avoir connaissance de leurs impacts sur les différents transcrits qu'ils chevauchent ou encore leur fréquence dans la population générale. Désormais, afin de ne conserver que les variants ayant la plus forte probabilité d'être responsable du phénotype de ces patients, nous avons appliqué successivement six filtres basés à la fois sur les différentes annotations que nous avons ajoutées mais aussi sur nos connaissances du mode de transmission du phénotype :

1. **Filtre 1 : L'union des variants :** Dans ces différentes études, nous avons à chaque fois séquencé des frères (deux ou trois) présentant phénotype. Ainsi nous avons pu formuler l'hypothèse d'une cause génétique commune entre les différents patients d'une même famille et donc filtrer l'ensemble des variants qui ne sont pas partagés les deux ou trois frères atteints testés.

2. **Filtre 2 : Génotype des variants :** Dans ces études, nous avons émis l'hypothèse d'une transmission récessive du phénotype. Ainsi, seuls les variants homozygotes ont été conservés. (**Figure : ??**, **??**).

3. **Filtre 3 : Impact du variant :** Afin de ne conserver que les variants ayant un effet potentiellement délétère sur la protéine, nous avons filtré les variants intronique et ceux tombant dans les séquences UTRs. De même les variants synonymes ne sont pas conservés (exceptés ceux se trouvant proches des régions d'épissage) car ceux-ci n'ont aucun effet sur la séquence protéique. Pour les variants faux sens (changement d'un seul aa de la séquence protéique) il est plus difficile de se trancher, nous avons donc utilisé les logiciels SIFT (Kumar, Henikoff, & Ng, 2009) et Polyphen (Adzhubei et al., 2010) et filtré l'ensemble des faux-sens prédits comme *tolerated* par SIFT et *benign* par Polyphen.

4. **Filtre 4 : Les transcrits "non pertinents" :** Au cours de nos analyses nous nous sommes concentré uniquement sur les transcrits codant pour une protéine. Ainsi, l'ensemble des transcrits annotés comme étant non codant furent filtrés. De même le mécanisme NMD (*nonsense-mediated decay*) a pour but de contrôler la qualité des ARNm cellulaires chez les eucaryotes (Y.-F. Chang, Imam, & Wilkinson, 2007) en éliminant les ARNm qui comportent un codon stop prématuré (K. E. Baker & Parker, 2004), pouvant être le résultat d'une erreur de transcription, d'une mutation ou encore d'une erreur d'épissage. Il est donc peu probable que les variants présents sur des transcrits annotés NMD soient responsables du phénotype. Dès lors, ces transcrits ont été également filtrés. Ainsi, nous avons pu retirer de nos listes de variants l'ensemble des mutations impactant **uniquement** des transcrits non codant et / ou annoté NMD. Cette étape de filtre permet à elle seule de filtrer systématiquement entre 36576 et 44581 transcrits différents par patients, soit une moyenne de 3107 variants par individus (**Figure :** 4.5).

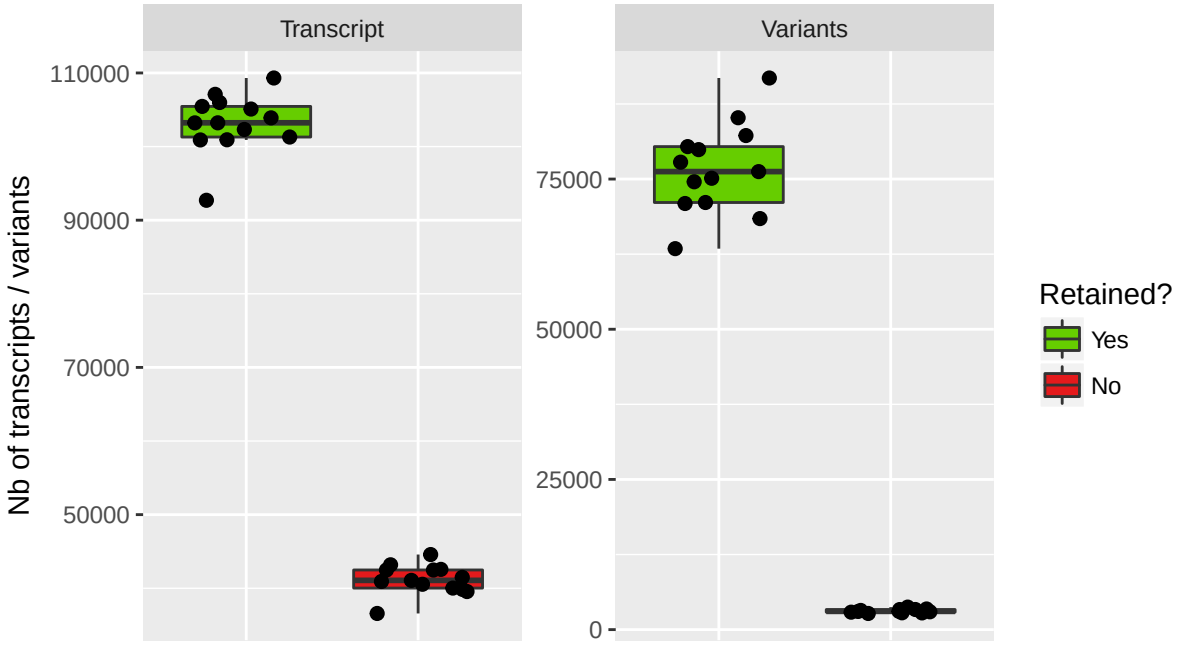


Figure 4.5 – Filtrage des transcrits jugés "non pertinents" et des variants les chevauchant : Pour chaque patients nous avons filtrer les transcrits jugés "non pertinents" pour l'analyse, c'est à dire ceux ne codant pas pour une protéine et ceux annoté NMD. Dès lors, l'intégralité des variants chevauchant uniquement des transcrits non pertinents ont put systématiquement être filtrés (boites rouges). les autres furent conservés (boites vertes)

5. **Fréquence des variants :** La fréquence d'un variant dans la population générale est un moyen rapide d'avoir une prédiction fiable de l'effet délétère ou non de celui-ci. En effet, il est peu probable qu'un variant retrouvé fréquemment dans la population générale soit causal d'une pathologie sévère. Ainsi nous avons filtré pour l'ensemble de nos patients l'ensemble des variants ayant une fréquence ≥ 0.01 dans l'une des trois bases de données que sont ExAC, ESP et 1KG.

6. **Présence des variants dans la cohorte contrôle :** Au cours de nos différentes études, nous avons été amenés à séquencé un total de 134 un des 6 phénotypes étudiés au cours de nos différentes études (**Table : ??**). Ces phénotypes étant très différent, on peut émettre l'hypothèse que leurs causes génétiques soient également différentes. De même, les variants recherchés étant rares, il est peu probable qu'un individu porte les variants de deux phénotypes différents. Ainsi, pour chacune des 6 familles, nous avons pu constituer une cohorte contrôle composée dans l'ensemble des patients précédemment analysés et ne présentant pas le même phénotype que celui étudié dans la famille (**Figure :** 4.6). Dès lors, nous avons pu filtrer l'ensemble des variants retrouvés à la fois chez nos patients et observés à l'état homozygote dans la cohorte contrôle. Cette cohorte contrôle

présente ainsi le même rôle que les bases de données publiques. Sont intéret principale par rapport à celles-ci est que les individus qui la composent ont pour la plupart la même origine ethnico-géographique que nos patients. De plus ceux-ci ont été séquencés en même temps dans les mêmes centres permettant ainsi d'identifier les artefacts dûs aux protocols de séquençage.

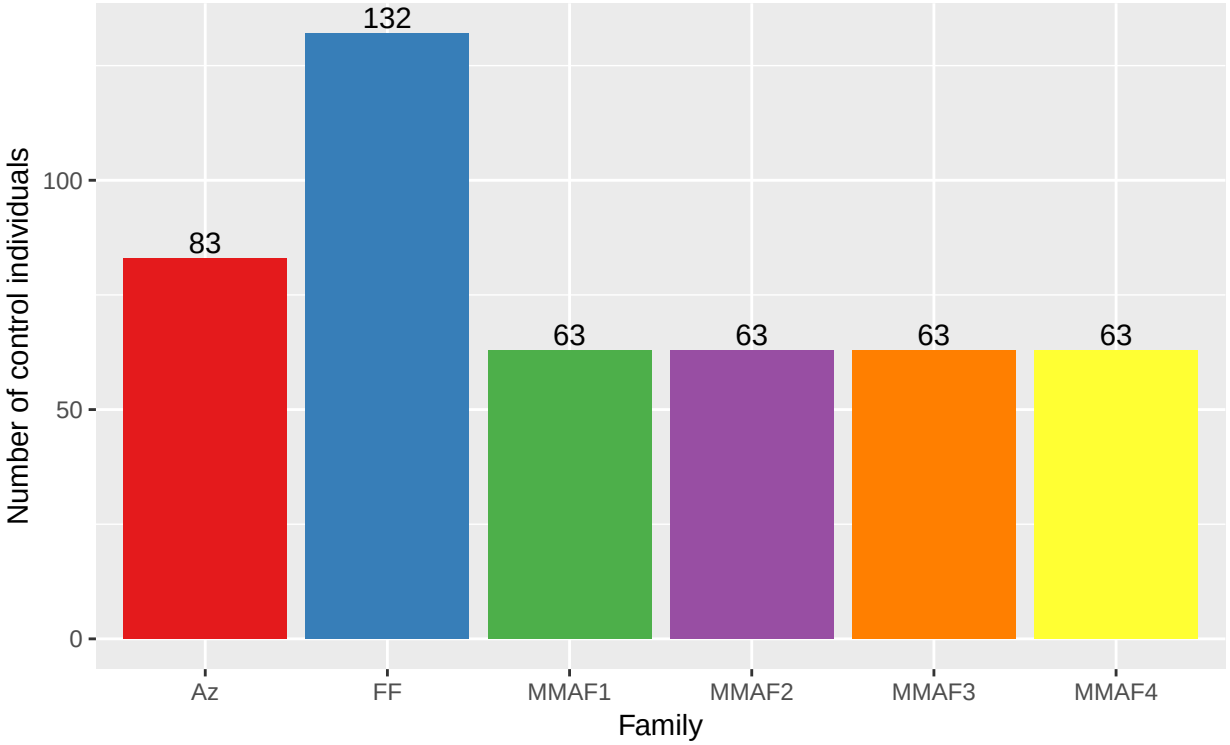


Figure 4.6 – Nombre d'individus composant la cohorte contrôle de chaque famille : Ici, chaque barre représente une famille et sa hauteur est déterminée par le nombre d'individus composant la cohorte contrôle à laquelle elle a été confronté. Chaque individu de la cohorte contrôle a été séquencés en WES par notre équipe. Afin d'être considéré comme "contrôle" et intégrer cette cohorte, un individu doit être sain ou présenter un phénotype d'infertilité différent de la famille étudiée. Par exemple, un individus MMAF pourra servir de contrôle aux familles AZ et FF mais pas aux familles MMAF1-4

Comme on pouvait s'y attendre, ces six filtres ont un pouvoir discriminant extrêmement différent (**Figure :** 4.7). En effet, tandis que le filtre "Transcript relevance" (filtre n°4) éliminer en moyenne 3.9 % des variants de chaque individu tandis que le filtre "Variant impact" (filtre n° 3) élimine jusqu'à 90.1 % de ces mêmes variants (**Figure :** 4.7 - **A**). Cette différence n'est pas surprenante. En effet, comme nous l'avions vu plus tôt, les variants de la catégorie VEP MODIFIER qui regroupe entre autres les variants chevauchant les séquences UTRs et introniques (**Table : ??**) représentent en moyenne ... % des variants de nos patients (**Figure :** 4.4 - **A**). Ceux-ci étant tous filtrés, on s'attendait donc à une valeur aussi élevée. On peut également constater l'importance de la cohorte contrôle qui, je le rappelle, permet de filtrer l'ensemble des variants homozygotes observés en son sein, puisque ce filtre permet retirer entre 76.5 et 88.4% des variants de chaque individus (**Figure :** 4.4 - **A**).

Cependant, regarder uniquement le pourcentage de variants filtrés par chaque filtre révèle une information partielle. En effet, dans ce cas de figure, on observe la quantité de variant éliminé par chaque filtre indépendamment les uns des autres. Ainsi, un même variant peut donc être filtré par plusieurs filtres. Dès lors, il faut également analyser la quantité de variants filtrés **spécifiquement** par chaque filtre. Ainsi, on peut constater que le classement des filtres en fonctions de leur stringance reste quasiment identique (**Figure :** 4.7 - **B**) il est tout de même intéressant de noter que désormais le filtre "Variant impact" apparait moins efficace que les filtres "Ctrl" et "Genotype" en filtrant spécifiquement une moyenne de 253 variants par individu contre 423 pour le filtre génotype et 882 pour le filtre "Ctrl". Ainsi, ce dernier devient celui filtrant spécifiquement le plus de variants avec entre 364 et 1060 variants spécifiquement filtrés par patients confirmant ainsi l'importance de ce filtre dans nos analyses. Aussi, les filtres "Transcript relevance", "Union" et "Frequency" apparaissent désormais comme étant anecdotiques en comparaison aux trois autres filtres puisqu'ils filtrent au maximum 43 variants spécifiques (**Figure :** 4.7 - **B**).

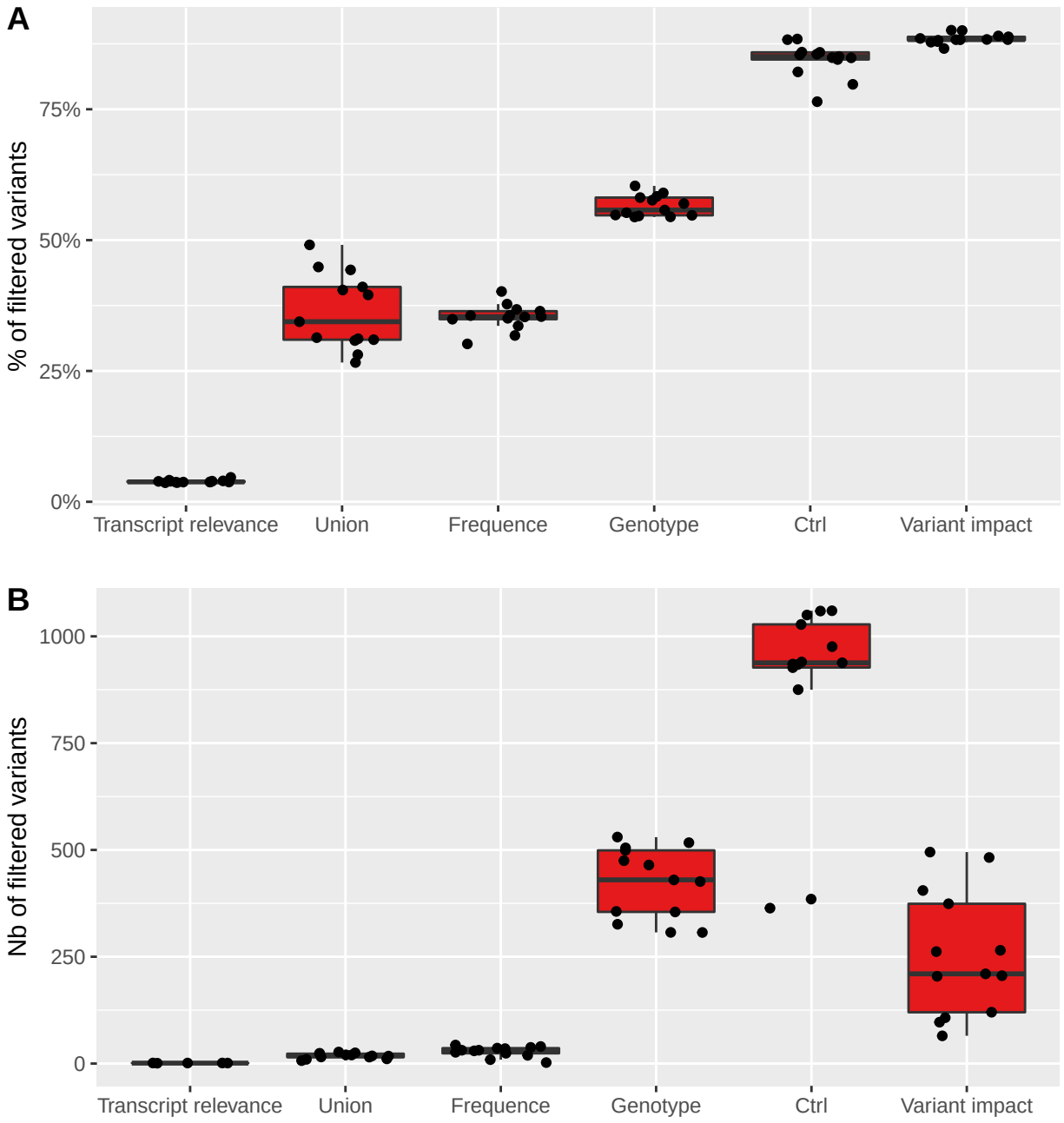


Figure 4.7 – Comparaison de l'efficacité de chacun des six filtres utilisés : **A** : Comparaison du pourcentage de variants filtrés par chacun des six filtres indépendamment les uns des autres pour chaque patient (représenté par les points. Dès lors, un même variant peut être filtré par plusieurs filtres. **B** : Comparaison du nombre de variant filtrés spécifiquement par chacun des filtres. Ici, un variant ne peut-être filtré que par un seul filtre

Après avoir appliqué l'ensemble de ces filtres, seuls quelques variants subsistent nous permettant d'obtenir une liste de gènes restreinte pour chaque famille (**Table :** 4.3) et ainsi de tirer des conclusions quant au variant responsable du phénotype.

1. **Famille AZ** : Parmi les 2 gènes restant pour cette famille, *SPINK2* est apparu comme étant un candidat évident. Notamment son expression étant spécifique au testicule tandis que celle de *GUF1* est ubiquitaire (**Figure :** 4.8). De plus, des mutations du gène *Spink2* chez la souris avait déjà été identifiée comme induisant des défauts de la spermatogenèse (Lee et al., 2011).

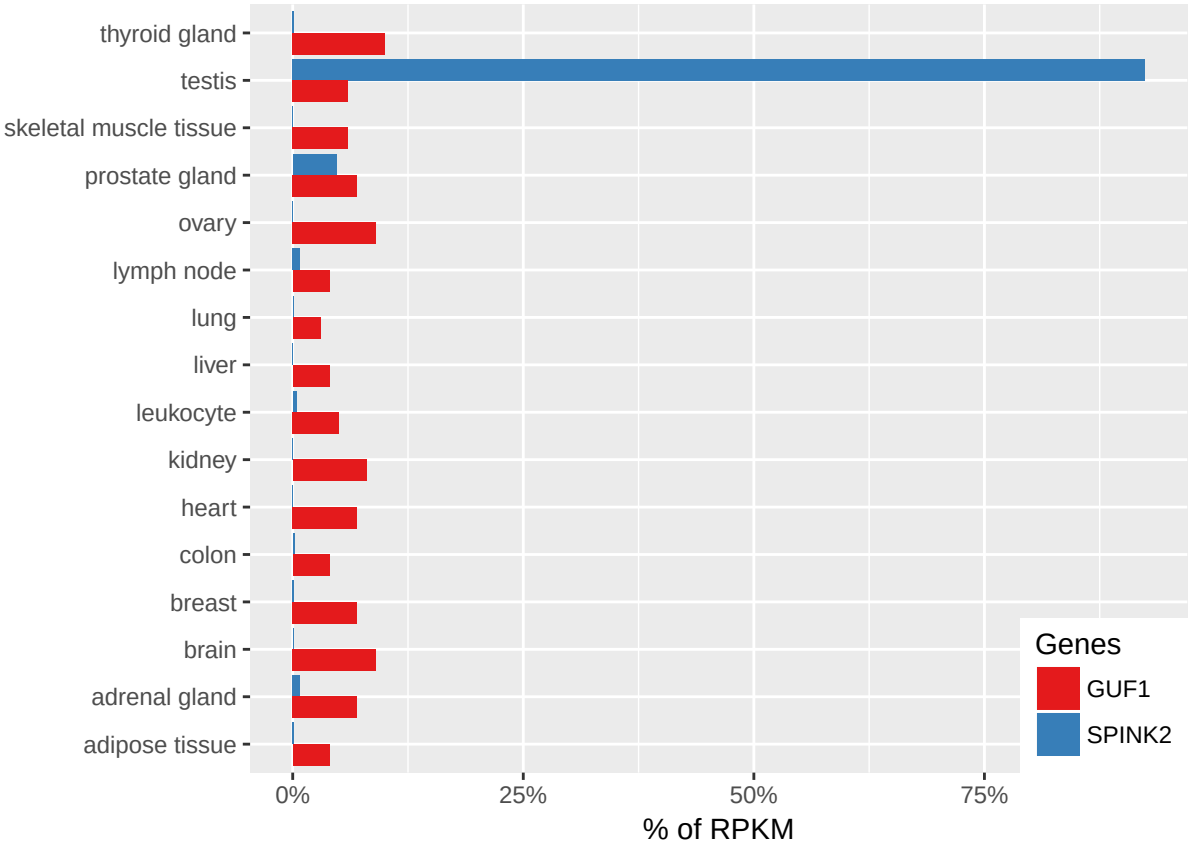


Figure 4.8 – Expression tissulaire des gènes *SPINK2* et *GUF1* : Données provenant du projet de transcriptome Illumina bodyMap

2. **Famille FF** : Pour cette famille, le gène *PLC*ζ1 a passé l'ensemble des filtres. Nos connaissances sur la fonction de se gène et notamment son rôle dans l'activation ovocytaire (TODO : REF) ainsi que sa forte expression testiculaire ont fait de ce gène le candidat idéal pour expliquer le phénotype de ces deux frères (**Figure :** 4.9).

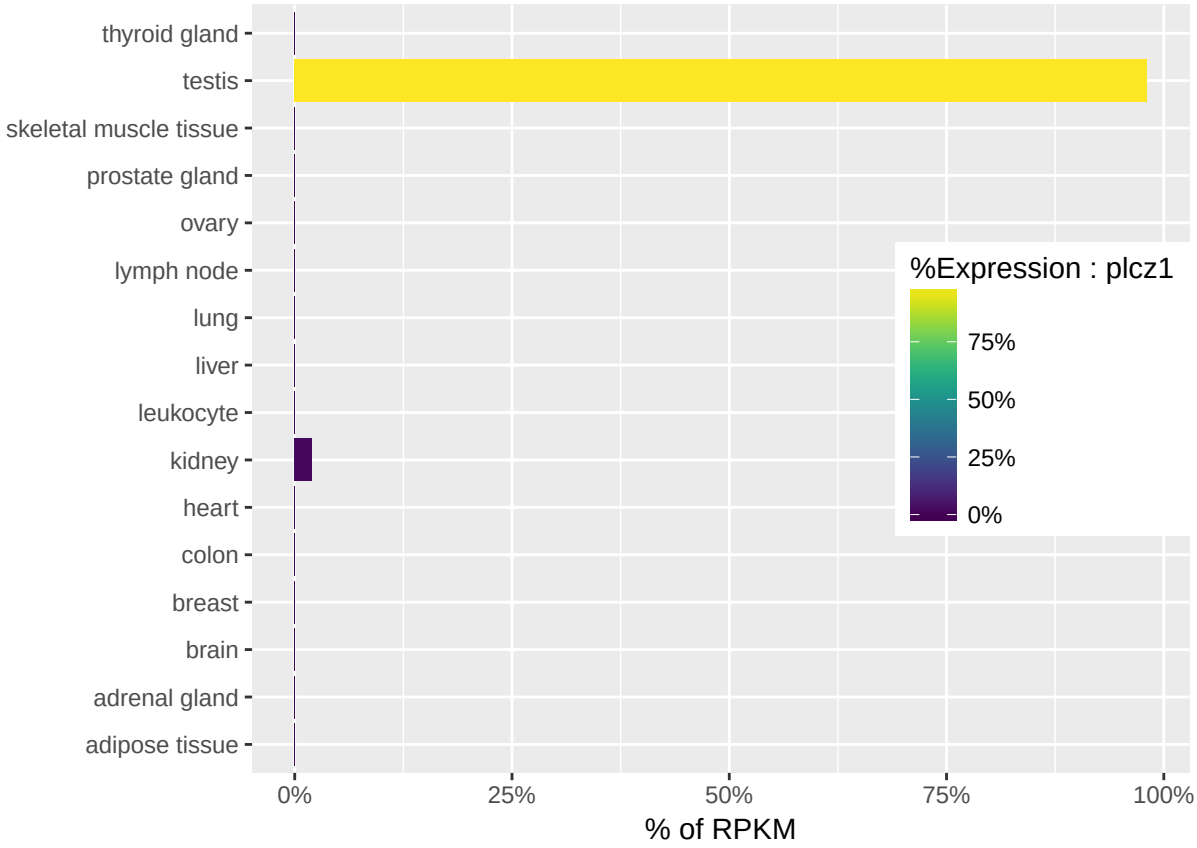


Figure 4.9 – Expression tissulaire du gène *PLCZ1* : D'après les données du Illumina BodyMap

3. **Famille MMAF1** : L'analyse bibliographique des 2 gènes ayant passé l'ensemble des filtres n'a pas pu nous permettre de d'affirmer que l'un de ces gènes étaient responsable du phénotype MMAF de ces 2 frères.

4. **Famille MMAF2** : À l'issue des filtres, 2 gènes ressortaient chez ces deux frères : *MYH11* et *DNAH1*. Or, notre équipe ayant déjà établit le lien entre des mutations du gène *DNAH1* et le syndrome MMAF (Ben Khelifa et al., 2014) ce gène s'est révélé être un candidat idéal pour expliquer le phénotype de ces 2 frères. De plus, l'implication de *MYH11* dans le phénotype de dissection aortique (Imai et al., 2015) l'ont écarté des candidats pour le phénotype MMAF.

5. **Famille MMAF3** : Comme pour les gènes de la famille MMAF2, l'analyse bibliographique des 7 gènes ayant ici passé les filtres de même que l'étude de leurs expressions ne nous a pas permis de conclure que l'un d'entre eux étaient responsable du phénotype MMAF de ces 2 frères.

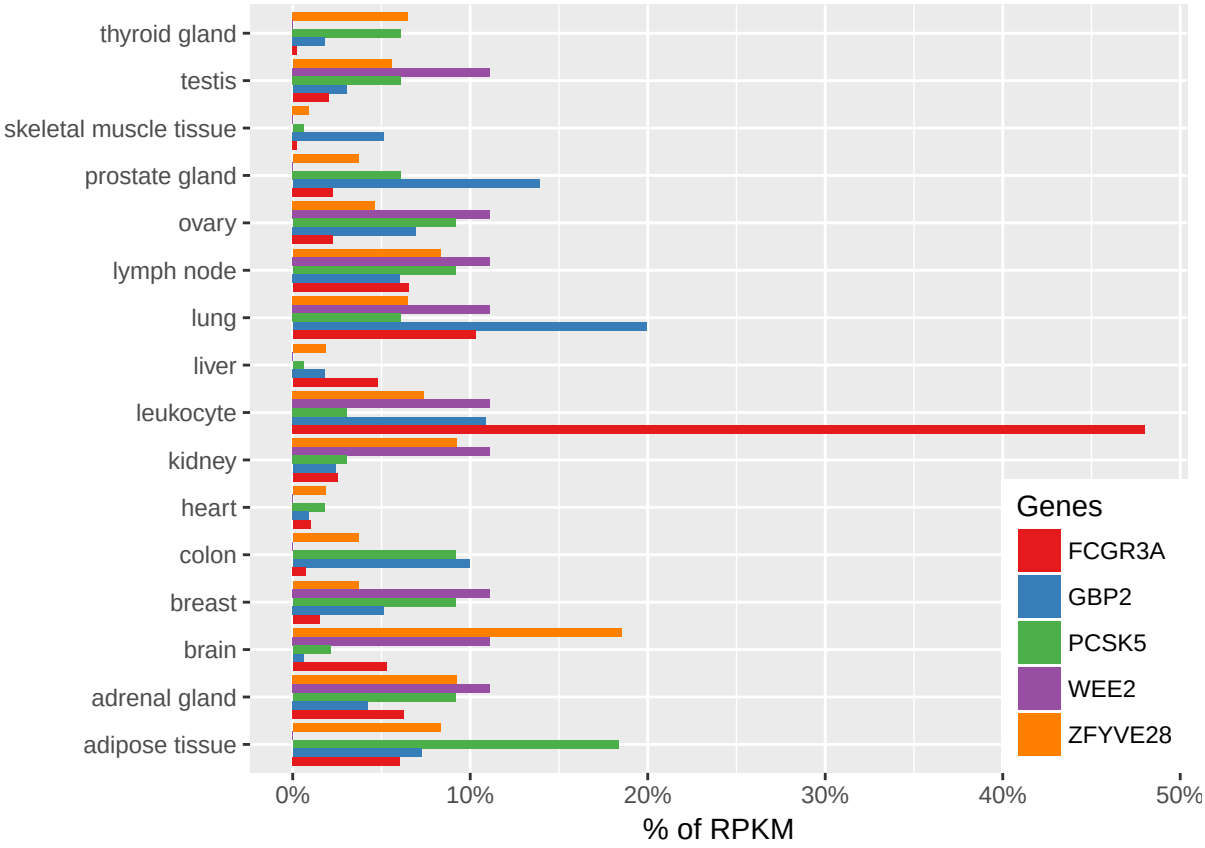


Figure 4.10 – Expression tissulaire des gènes retenus pour la famille MMAF3 : Données provenant du projet de transcriptome Illumina bodyMap

6. **Famille MMAF4** : Seul le gène *TGIF2* a passé l'ensemble des filtres pour la famille MMAF4. L'expression ubiquitaire de ce gène n'en font pas un candidat idéal. Cependant une étude de 2011 effectuée sur le wallaby décrit que la protéine TGIF2 est localisée spécifiquement dans le cytoplasme du spermatide, ainsi que dans le corps résiduel et la pièce intermédiaire du flagelle du spermatozoïde mature (Hu, Yu, Shaw, Renfree, & Pask, 2011). Ces données pourraient corréler avec le phénotype MMAF de ces 3 frères bien que l'expression de ce gène soit ubiquitaire (**Figure :** 4.11).

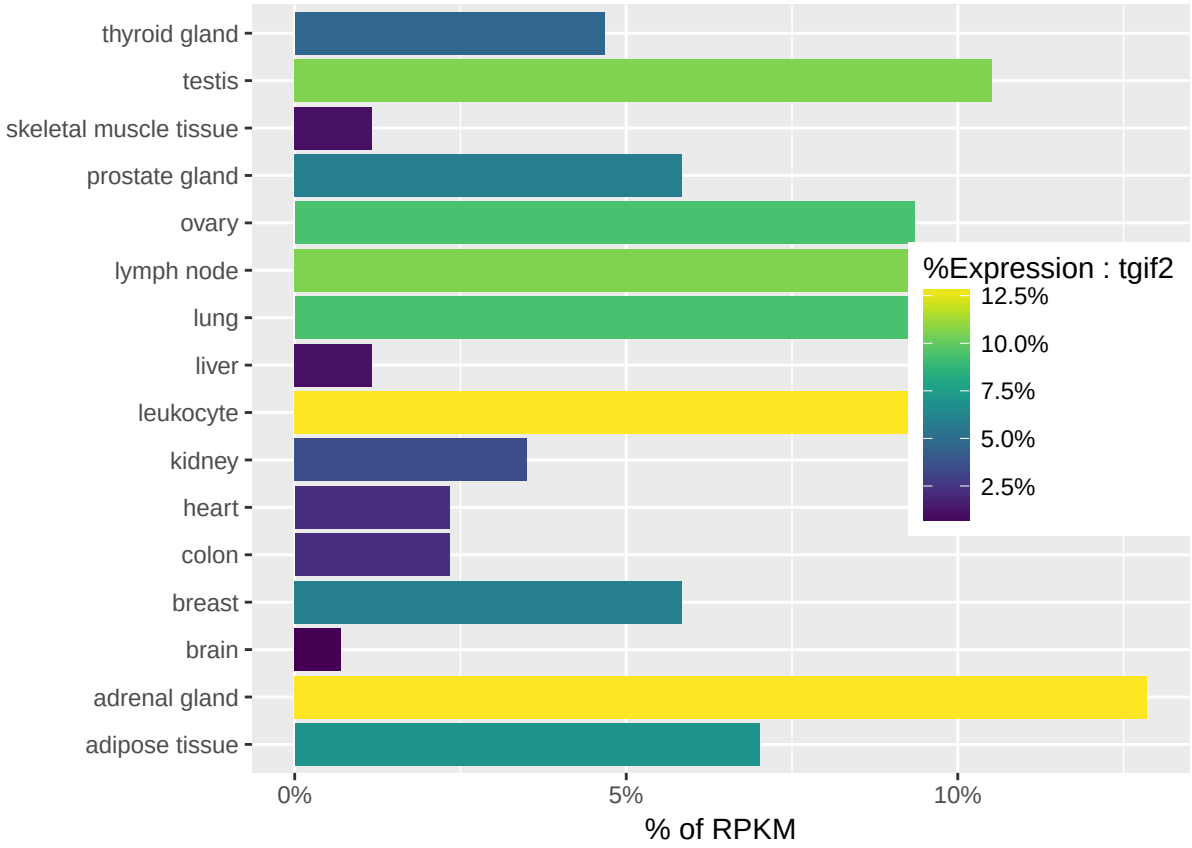


Figure 4.11 – Expression tissulaire du gène *TGIF2* : D'après les données du Illumina BodyMap

Discussion

L'analyse de ces 6 familles nous a permis de mettre en évidence l'efficacité de notre pipeline d'analyse puisque pour 3 d'entre elles (soit 50%) le variant causal a pu être identifié avec certitude (**Figure :** 4.12) et les résultats publiés dans trois revus dont je suis co-auteur :

1. **Famille AZ** : **SPINK2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes** : Dans cet article j'ai effectué non seulement l'intégralité des analyses bioinformatiques des données d'exomes de deux frères infertiles présentant un phénotype d'azoospermie mais j'ai aussi séquencé en Sanger les séquences codantes du gène *SPINK2* pour une partie des 611 individus analysés ainsi que contribué à l'extraction de l'ARN testiculaire des souris pour l'analyse fonctionnelle du gène *Spink2* sur le modèle murin.

2. **Famille FF** : **Homozygous mutation of PLCZ1 leads to defective human oocyte activation and infertility that is not rescued by the WW-binding protein PAWP** : Dans cet article j'ai, effectué l'intégralité des analyses bioinformatiques des données d'exomes effectuées sur deux frères infertiles présentant des échecs de fécondation.

3. **Famille MMAF2** : **Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations** : Dans cet article j'ai, comme précédemment, effectué l'ensemble des analyses bioinformatiques des données d'exomes effectuées sur deux frères infertiles présentant des échecs de fécondation.

Pour une d'entre elle, un candidat potentiel a pu être mis en évidence avec le gène *TGIF2* et notre équipe travaille actuellement sur la caractérisation de ce gène afin de savoir s'il peut effectivement expliquer le phénotype MMAF de cette famille (**Figure :** 4.12).

TODO : Il faut aller plus loin dans l'analyse et les arguments pour convaincre qu'il s'agit d'un bon candidat : quel type de mutation, ce gène est-il bien conservé, son expression n'est pas spécifique au testicule et ce gène serait impliqué dans un phénotype d'holoproencephaly...

Pour les 2 familles restantes, aucun variant n'a pu pour l'instant expliquer leur phénotype. L'explication la plus vraisemblable est que le variant ait été filtré par l'un de nos six filtres, probablement celui consistant à filtrer l'ensemble des variants hétérozygotes. En effet, l'hypothèse d'un variant causal homozygote était extrêmement crédible pour les familles AZ, FF et MMAF2 étant donné l'historique consanguin de ces 3 familles dont les parents sont à chaque fois apparentés. En revanche rien ne laisse supposer une telle chose pour les familles restantes. Cependant, le filtre des variants hétérozygotes pour l'ensemble des patients de ces 3 familles a été maintenu en

première intention afin de faciliter les analyses en réduisant au maximum le nombre de variant. Au vus des résultats il apparait clair que les variants responsables de leur phénotype aient été filtrés pour au moins 2 de ces familles. Dès lors, l'ensemble des analyses effectuées lors de l'étape de filtrage doivent être refaites en changeant les paramètres de filtrage. Cette fois-ci, les variants hétérozygotes seront conservés et les gènes sur lesquels au moins deux variants hétérozygotes seront recensés seront analysés en priorité. En effet, bien que les analyses exomiques nous fournissent en l'état pas d'informations suffisante pour savoir si ces deux variants sont présent sur le même allèle ou bien sur deux allèles différents, cela pourrait-être la signature de variants hétérozygotes composites. C'est donc sur ces analyses que se concentre actuellement notre équipe.

Table 4.3 – Récapitulatif des variants ayant passé l'ensemble des filtres pour chaque famille

Family	Gene	Impact
MMAF3	WEE2	missense
MMAF3	FCGR3A	missense
MMAF3	FCGR3A	missense
MMAF3	FCGR3A	missense
MMAF3	GBP2	missense
MMAF2	MYH11	missense
MMAF2	DNAH1	splice acceptor
MMAF3	PCSK5	missense
AZ	GUF1	missense
MMAF1	PLA2G4B	splice region
MMAF1	JMJD7-PLA2G4B	splice region
MMAF3	ZFYVE28	missense
MMAF4	TGIF2	missense
AZ	SPINK2	splice region
FF	PLCZ1	missense
FF	PLCZ1	missense
FF	PLCZ1	missense

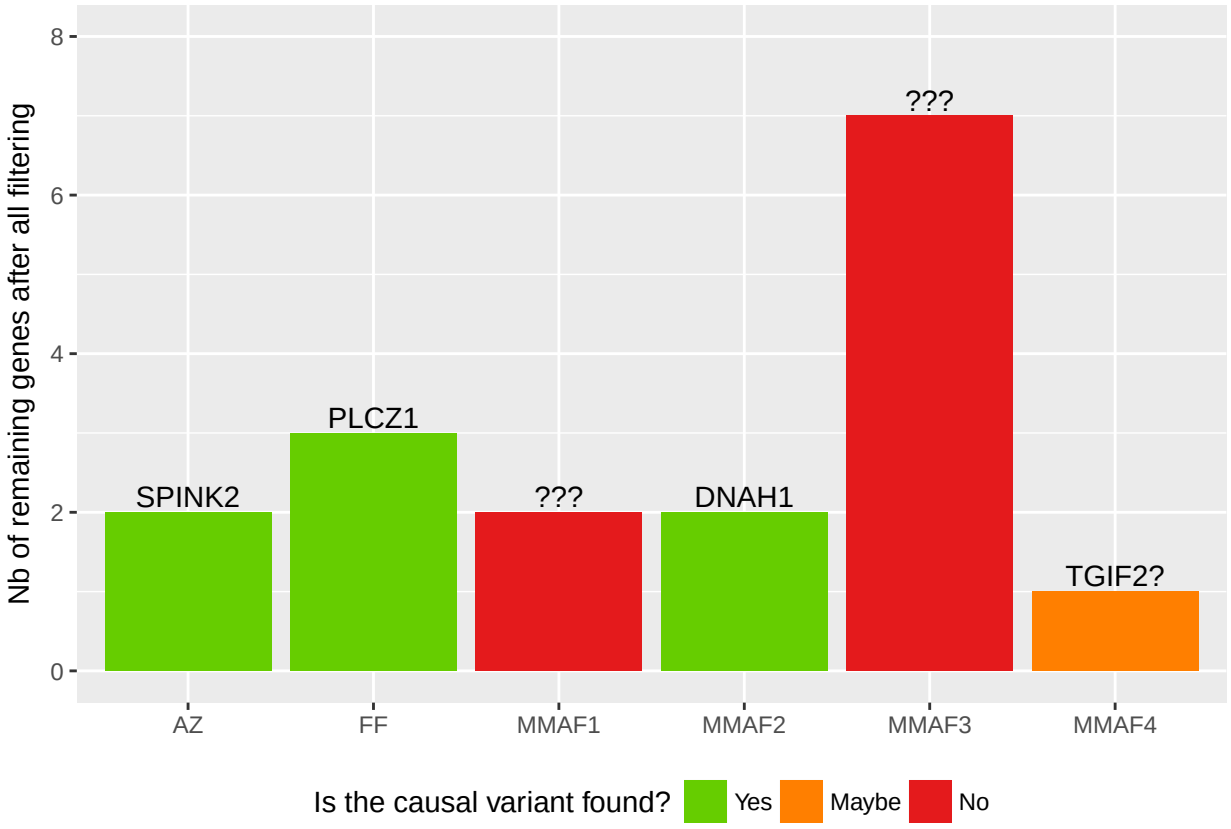


Figure 4.12 – Nombre de gènes passant l'ensemble des filtres par famille : Chaque barre représente une des familles analysées. La hauteur de cette barre correspond au nombre de gènes ayant passé l'ensemble des filtres pour chaque famille. Les barres vertes caractérisent les familles pour lesquelles le gène responsable de la pathologie a été identifié parmi la liste de gène (dans ce cas le symbole du gène est écrit au-dessus de la barre). La barre orange caractérise la famille pour laquelle un candidat potentiel a été identifié (le symbole du gène est écrit au-dessus suivit d'un " ?"). Les barres rouges indiquent qu'aucun des gènes ayant passé les filtres pour ne semble expliquer le phénotype (dans ce cas il est écrit " ? ? ?" au-dessus de la barre)

4.2.3 Etude d'une large cohorte de patients MMAF

Description de la cohorte

Historique : après avoir mis en évidence DNAH1 -> MMAF notre équipe s'est en partie spécialisé dans ce syndrome.

ainsi, entre (année) et année, notre équipe a effectué le séquençage de ... individus présentant ce phénotype afin d'en établir la cause génétique. parmi ces patients, la majorité provenait d'Afrique du Nord, cependant ... vfenaient de et de ... ces séquençage ont été effectué dans ... centres diférents que sont (listes des centre de séquençage) et sur ... plateforme : liste des plateformes

Table 4.4 – Liste des différents projets de séquençages effectués

Place	Year	Nb of sequenced individuals
MountSinai	2012	2
Strasbourg	2012	13
Genoscope	2013	13
Genoscope	2014	28
Genoscope	2015	6

Application de la pipeline - Résultats

Après avoir appelé les variants de nos 62 patients, nous avons obtenu un total de 4484558 variants différents comprenant 4160274 SNVs et 324284 indels. Ces variants étant répartit entre chaque patient qui portaient environs chacun 81618 SNV et 5148 indels dont 42.8 % étaient homozygote. Comme on peut le voir, la proportion de chaque appel est relativement homogène lorsque l'on compare les patients ayant été séquencés dans le même centre la même année. Cependant, il est possible de noter de grandes disparités lorsque l'on compare les données provenant de différents centres ou bien du même centre avec plusieurs années de différences. Ces écarts peuvent-être causés par plusieurs facteur, tel que les différents kits de capture d'exons qui on put être utilisés puisque ... (**todo lister les différents kits de capture dans une table**) en revanche nous pouvons écarter un effet dus à la plateforme de séquençage ou encore le modèle de séquenceur puisque tous ces projets ont été réalisés sur des Illumina HiSeq2000 (**Table :** 4.4) (**Figure :** 4.13 - **A**).

Le même constat peut être effectué lorsque l'on compare la qualité des appels puisque plus les projets de séquençage s'avèrent être récent, plus la proportion d'appel *Single Strand* s'avère être faible tandis que la proportion d'appel *Double Strand* (DS) est élevée. Ceci est une bonne chose, car, bien que ces deux appels soient conservés dans les analyses ultérieures, les appels DS sont de meilleure qualité que les appels SS. Cette augmentation des appels DS au cours du temps pourrait s'expliquer par une amélioration des protocoles de séquençage ainsi que des kits de capture. En revanche cela est à pondérer avec le taux croissant d'appels *No-strand* (NS) au fur et à mesure des années pour atteindre environs 21.3 % en 2015 avec un projet réalisé au Génoscope. Ces derniers appels étant systématiquement filtrés, ils n'altèreront en rien les résultats obtenus en aval hormis le fait qu'ils réduisent la quantité des données utilisées (**Figure :** 4.13 - **B** et **C**).

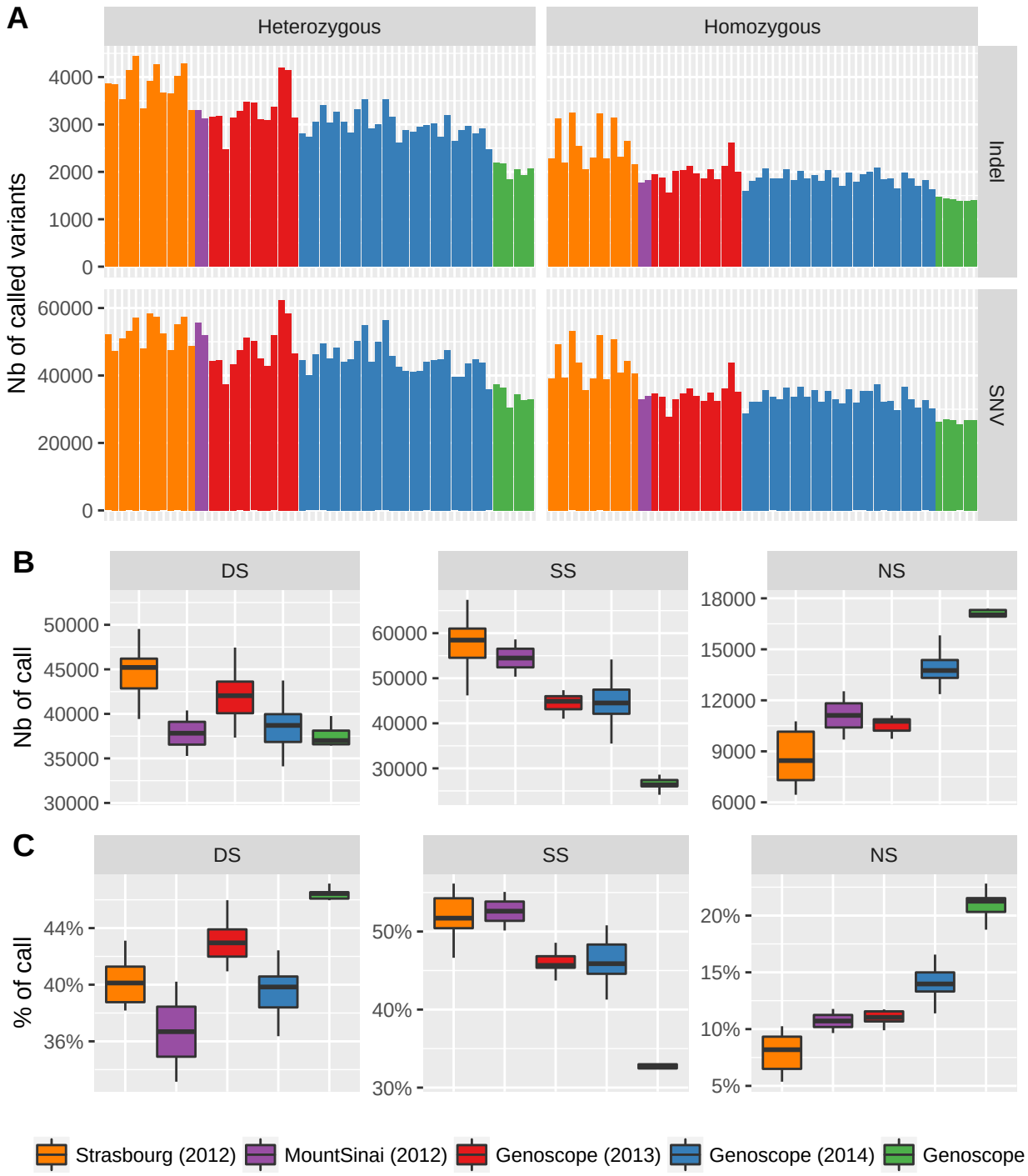


Figure 4.13 – Résultats de l'appel des variants par individus et par projet de séquençage : Chaque couleur définit un projet de séquençage caractérisé par un centre de séquençage et une année. **A** : Quantification pour chaque individus (représentés par les barres) du nombre de variants (SNVs et Indels) appelés homozygotes et hétérozygotes. **B** : Quantification des appels *Double Strand* (DS), *Single Strand* (SS) et *No strand* (NS) pour chaque projet de séquençage. **C** : Même chose en pourcentage

Analyse des listes de gènes

Après avoir appliqué les mêmes filtres que ceux décrit précédemment à l'exception du filtre n°1 "Union" puisqu'ici nous avons uniquement des individus non apparentés, nous avons pu obtenir une liste de 1711 variants différents composés de 1470 SNVs et 241 indels et impactant un total de 1432 gènes distincts. Ces variants étant répartis sur l'ensemble de nos 62 patients ceux-ci portaient en moyenne 27 SNVS et 5 indels, de sorte que chacun d'entre eux avaient entre 1 et 86 gènes impactés par au moins un variants homozygote (**Figure :** 4.14 - **A** et **B**).

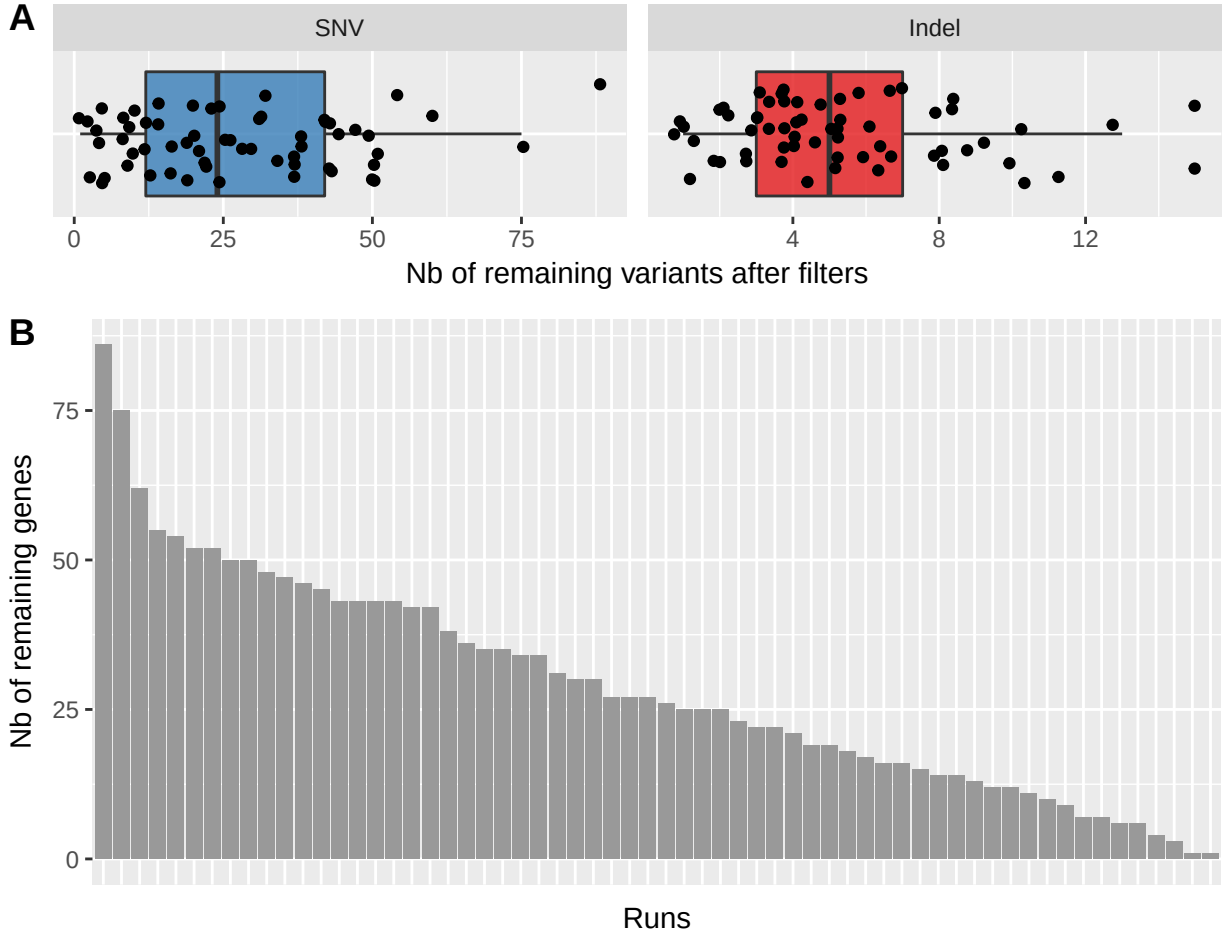


Figure 4.14 – Résultats de l'étape de filtrage : **A** : Quantification du nombre de SNVs et indels ayant passé l'ensemble des filtres pour chaque patient. **B** : Nombre de gènes impactés par au moins un variant ayant passé les filtres pour chaque individu représenté par les barres. **C** : Présentation

Afin de déterminer parmi cette ensemble de gènes ceux responsables du phénotype de nos patients, nous avons procédés en trois étapes :

1. **Étape n°1** : Cette étape consiste à séléctionner uniquement les variants ayant un effet tronquant sur la protéine comme un décalage du cadre de lecture, l'apparition d'un codon stop prématuré ou encore une perturbation des sites accepteurs / donneur d'épissage. Une analyse de l'expression testiculaire des gènes retenus ainsi qu'une étude bibliographique nous permet ensuite de séléctionner tout ou partie de ceux-ci. Du fait des effets extremement délétères de ces variants, les patients ressortant de cette première étape sont considérés comme de confience élevée (*High trust*).

2. **Étape n°2** : Pour l'ensemble des gènes retenus dans l'étape n°1, nous recherchons ensuite des patients portant, toujours à l'état homozygote, des variant aux effet non tronquant tel que des variants faux-sens ou encore des variants intronique situés proches des sites d'épissage. Dans le cas des variants faux-sens, les logiciels SIFT et PolyPhen sont ensuite utilisés afin de nous orienter quant à l'effet délétère du variant, bien que comme nous l'avons déjà vu, ces logiciels son contredisent regulièrement [TODO : ref!!!]. Au vus de la difficulté à déterminer l'effet délétère de ces variants, les patients identifiés au cours de cette étape sont marqués comme de confience modérée (*Moderate trust*).

3. **Étape n°3** : Cette étape consiste à recherchere des patients éventuellement hétérozygotes composites, c'est à dire des patients portant deux variants hétérozygotes différents sur chacun des deux allèles d'un même gène. Malheuresement, dans le cadre de séquençage WES WGS, il est impossible de connaitre le "phasage" des variants, c'est à dire que l'on ne peut déterminer si deux variants hétérozygotes sont situés sur le même allèles ou sur deux allèles différents. Pour cela, des analyses de biologie moléculaire sont nécéssaire. C'est pour cette raison que les patients identifier au cours de cette étape sont labellisés comme étant de faible confience (*Low trust*).
4. **Étape n°4** : Cette étape a pour but d'alleger le nombre de variant et donc de gène restant à analyser. Pour cela, nous retirons les données des patients pour lesquels la cause génétique a pu être déterminée **avec certitude**, c'est à dire ceux portant des variants homozygotes tronquants (c'est à dire ceux identifiés lors de l'étape n°1). En effet, la plupart des variants ressortant de l'étape n°2 sont des variants faux-sens, or, malgré les prédiction de SIFT et PolyPhen, il est impossible d'affirmer avec certitude que ces variants ont un réel impact sur la protéine sans effectuer d'analyses fonctionelles. De même pour les patients identifiés lors de l'étape n°3 pour lesquels il est impossible sans analyses de moléculaire de déterminer si les deux variants hétérozygotes affectent le même allèle ou bien deux allèles différents.

***DNAH1*, un acteur primordial dans le phénotype MMAF** Au moment de nos analyses, le gène *DNAH1* était encore le seul décrit comme responsable du phénotype MMAF faisant de lui un candidat évident pour expliquer le phénotype MMAF de nos patients malgrés on expression non spécifique au testicule (**Figure :** 4.15 - **A**). C'est pourquoi nous avons appliquer les 3 étapes précédement décrites en ciblant spécifiquement les patients ayant des variants chevauchant les gène *DNAH1*.

1. **Étape n°1** : Parmis l'ensemble de nos 62 patients MMAF 1 portait un indel homozygote entrainant un décalage du cadre de lecture sur le gène *DNAH1* (**Table :** 4.8.

2. **Étape n°2** : La recherche de variants homozygote non tronquant sur le gène *DNAH1* nous a permis d'identifier 2 nouveaux patients : ... . On note ainsi que le patient *Ghs90* est porteur de ... variants homozygotes successif, tous entrainant des faux-sens prédits comm *benign* par PolyPhen et pour lesquels SIFT ne proposen aucune prédiction. Le patient *Ghs95*, lui porte égallement un variant faux-sens différent de ... observés chez *Ghs90* mais également prédit comme *benign* par PolyPhen. Il est cependant à ce stade impossible de conclure avec certitudes que ces variants soient effectivement responsables du phénotype MMAF de ces 2 patients compte tenus des prédictions fournis par PolyPhen et du fait qu'il est difficile de conclure quant à l'effet d'un variant-faux-sens sans effectuer d'analyses fonctionelles. Cependant, l'abscence de ces ... dans chacune des 3 bases de donées que nous avons utilisés, c'est à dire ExAC, 1KG et ESP6500, laisse supposer que ceux-ci soient êtremement rare
3. **Étape n°3** : La recherche d'hétérozygote composites nous a permis de révéler 6 patients portant tous 2 variants hétérozygotessur le gène *DNAH1*. On oeut alors noter le patients *Ghs36* portant 2 variant hétérozygotes. Le premier d'entre eux créé un codon stop de manière prématuré et n'est retrouvé ni dans ESP6500 ni dan 1KG tandis que sa fréquence dans la base de donnée ExAC est de .... Le second est un variant faux-sens abscent des 3 bases de données et prédit comme *probably damaging* par PolyPhen. Ainsi, s'ils étaient situés sur deux allèles différents, ces deux variants pourraient être des bons candidat pour expliquer le phénotype du patient *Ghs36*. On peut noter égallement le patient *Ghs129* portant deux faux-sens hétérozygote tout deux prédits comme *probably damaging* par PolyPhen et dont un seul est retrouvé dans la base de donnée exac avec une fréquence de ... . Pour les ... autres patients, il est plus dificile de se prononcer à ce stade compte tenus des fréquence parfois élevée des variants ou bien des prédictions fournies par PolyPhen.

4. **Étape n°4** : Au vus des résultats, seuls les données du patient *Ghs122* ont été retirées de notre liste de variants, celui-ci étant le seul à porter un variant homozygote tronquant. Cela a ainsi permis de réduire notre liste à 1661 variants distincts cheveauchant 1395 gènes différents.

Ainsi, cette première analyse nous a permis de révéler que 9 des 62 patients de notre cohorte portaient au moins 1 variant sur le gène *DNAH1* et que pour 3 d'entre eux ce(s) variants étaient présents à l'état homozygote. Cependant, il faut noter que du fait de son effet tronquant sur la proétine, seul le variant homozygote porté par le patient *Ghs122* nous permets d'être certains de la causalité du phénotype MMAF. Pour les 8 autres patient, des analyses fonctionelles complémentaires sont nécéssaires (**Figure :** 4.15 - **B**).

Ainsi, les mutations du gène *DNAH1* seraient ainsi responsables de manière certaine de 2 % des phénotype MMAF de notre cohorte. Ce pourcentage monte jusqu'à 15 % si l'on considère l'ensemble des patients identifiés dans cette analyse.

Bien que ce pourcentage soit en deçà des 40% (TODO : à confirmer !) observés dans notre étude précédente (Ben Khelifa et al., 2014), ces résultats confirment néanmoins le rôle primordial de la protéine DNAH1 dans la structure du flagelle et l'implication majeure du gène *DNAH1* dans le phénotype MMAF.

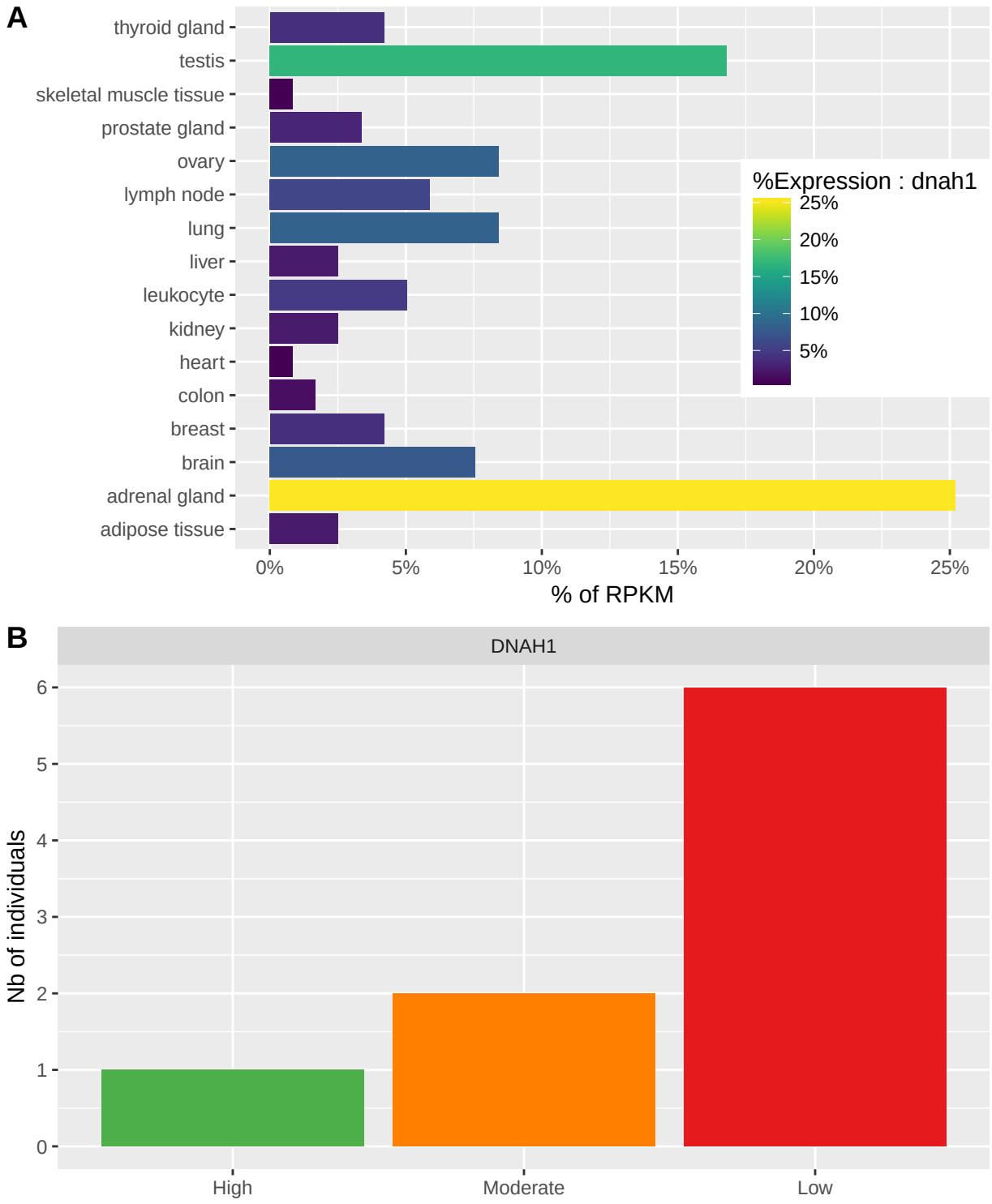


Figure 4.15 – Analyse du gène *DNAH1* : Expression tissulaire du gène *DNAH1* d'après les données du projet Illumina BodyMap. Quantification du nombre de patients portant au moins un variant sur le gène *DNAH1* pour chacun des 3 niveau de confiance

Table 4.5 – Liste des patients portant au moins un variant homozygote tronquant sur le gène *DNAH1*

Patient	Gene	Evidence	Consequence	ESP	1KG	ExAC
Ghs122	DNAH1	Not in the list	frameshift	none	none	none

Table 4.6 – Liste des patients portant au moins un variant homozygote non tronquant sur le gène *DNAH1*

Patient	Gene	Consequence	SIFT	PolyPhen	ESP	1KG	ExAC
Ghs90	DNAH1	missense	none	benign	none	none	none
Ghs95	DNAH1	missense	none	benign	none	none	none

Table 4.7 – Liste des patients portant au moins deux variant hétérozygotes sur le gène *DNAH1*

Patient	Gene	Consequence	SIFT	PolyPhen	ESP	1KG	ExAC
Ghs28	DNAH1	missense	none	benign	1e-04	none	1.65e-05
Ghs28	DNAH1	missense	none	benign	0.0027	0.0019	0.00233
Ghs36	DNAH1	missense	none	probably damaging	none	none	none
Ghs36	DNAH1	stop gained	none	none	none	none	8.29e-06
Ghs42	DNAH1	missense	none	probably damaging	none	none	none
Ghs42	DNAH1	splice region	none	none	none	none	none
Ghs87	DNAH1	missense	none	benign	none	none	0.00024
Ghs87	DNAH1	missense	none	probably damaging	7e-04	5e-04	0.000457
Ghs88	DNAH1	missense	none	benign	1e-04	none	0.000115
Ghs88	DNAH1	missense	none	benign	1e-04	0.0019	0.000149
Ghs129	DNAH1	missense	none	probably damaging	none	none	none
Ghs129	DNAH1	missense	none	probably damaging	none	none	8.26e-06

Les nouveaux acteurs Comme dit à l'instant, la recherche de variants sur le gène *DNAH1* nous permettrait d'expliquer jusqu'à 15 % des phénotypes des patients de notre cohorte. Cela signifie donc que pour au moins 85 % d'entre eux, la cause génétique du phénotype MMAF est donc expliquée par d'autres facteurs.

Afin de nous orienter dans nos recherches, nous nous sommes basés sur une étude de 2012 qui établissait une liste des gènes humains pouvant être impliqués dans cilliome, c'est à dire (todo def cilliome) (Ivliev, 't Hoen, Roon-Mom, Peters, & Sergeeva, 2012). La constitution de cette liste se basait à la fois sur les données de CilDB [ref ?] et de MEDLINE [ref ?] mais aussi des analyse *in silico* permettant d'effectuer des prédiction. Ainsi, chaque gène était classé dans l'une des 3 catégories suivantes en fonction des preuves déjà existante (au moment de l'étude) permettant de lier un gène au cilliome humain : **Strong evidence from previous studies** (Strong), **Weak evidence from previous studies** (Weak) et **No evidence from previous studies** (Novel). L'utilisation de cette liste nous a permis d'ajouter une nouvelle annotation pertinente à nos gènes. En effet, le spermatozoïde humain est une cellule ciliée, et le flagelle en est le cil. Nous pouvons donc attendre à ce qu'une partie des gènes responsables du phénotype MMAF soit présents dans cette liste de 371 gènes.

Ainsi, 32 de nos 1395 gènes retenus faisaient partis de cette liste dont 21 présentaient des preuves fortes de leur appartenance au cilliome. Il faut tout de même noter que bien que cette liste soit un bon outil pour orienter les recherches et prioriser certains gènes, elle ne peut constituer un critère suffisant pour filtrer les gènes n'en faisant pas partie. Par exemple le gène *DNAH1*, de par son expression ubiquitaire n'a pas été intégré à cette liste (**Figure : ??**, or on connait désormais son implication dans le phénotype MMAF (**Figure : ?? - A**).

Suite à cela, afin de nous concentrer en priorité sur les gènes entrainant un phénotype MMAF chez le plus grand nombre d'individus, nous avons sélectionnés ceux sur lesquels plusieurs patients portaient au moins un variant ayant passé l'ensemble des filtres nous permettant alors d'obtenir une liste de 212 gènes dont 145 (soit 68 %) étaient retrouvés variants chez uniquement 2 patients (**Figure : ?? - B**).

Ces deux informations nous ont ensuite permis de *designer* 4 analyses que nous avons effectué de manière successives nous permettant ainsi de séléctionner dans un premier temps les gènes candidats les plus évidents nous permettant ainsi d'épurer progressivement notre liste de variants rendant ainsi plus facile l'identification des candidats les moins évidents.

1. **Analyse n°1** : Dans un permier temps, nous avons étudié uniquement les gènes **présents dans la liste cilliome** sur lesquels **au moins deux** de nos patients présentaient au moins 1 variant tronquant à l'état **homozygote**.
2. **Analyse n°2** : Ensuite, nous avons étudié les gènes **absents dans la liste cilliome** mais sur lesquls on trouvait toujours **au moins deux** de nos patients présentant au moins 1 variant tronquant à l'état **homozygote**.
3. **Analyse n°3** : Dans un troisième temps nous sommes revenus à étudier les gènes **présents dans la liste cilliome** en considérant cette fois-ci les gènes sur

lesquels **un seul** de nos patients présentaient au moins 1 variant tronquant à l'état **homozygote**.

4. **Analyse n°4** : Pour finir nous avons étudié les gènes **absents dans la liste cilliome** sur lesquels **un seul** de nos patients présentaient au moins 1 variant tronquant à l'état **homozygote**.

Il faut noter qu'au sein des chacun de ces 4 analyses nous avons appliquer les étapes n°1-4 décrites précédemment.

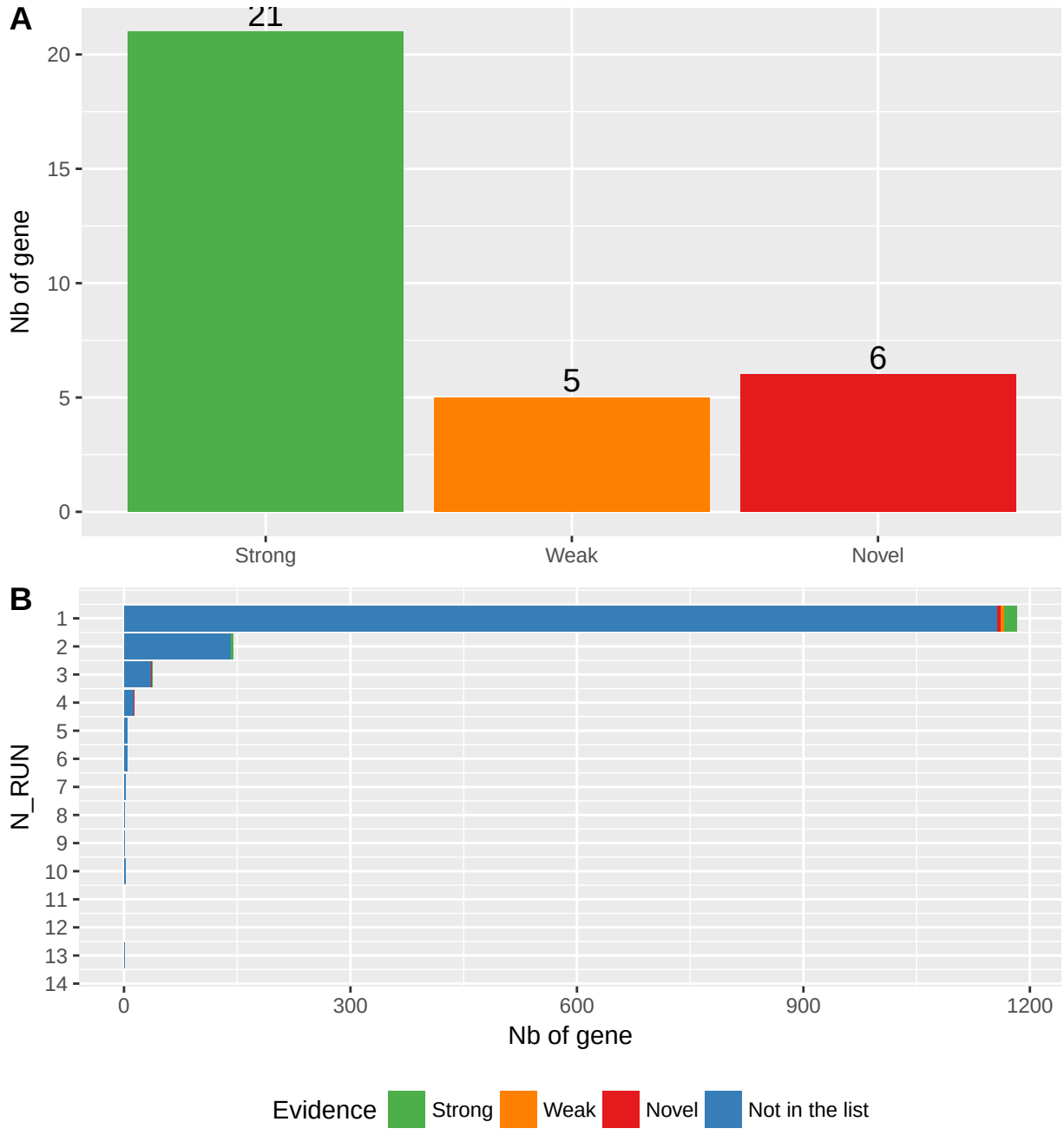


Figure 4.16 – TODOOOOOOOOOOOOOOOOOO : Chaque couleur définit une classe de la liste des gènes du cilliome décrit dans [@Ivliev2012]. Vert = *Strong evidence from previous studies* (*Strong*), Orange = *Weak evidence from previous studies* (*Weak*), rouge = *No evidence from previous studies* (*Novel*), bleu = Non présent dans la liste. **A** : Quantification du nombre de gène ayant passé les filtres au sein des 3 classes de la liste des gènes du cilliome. **B** : TODOOOOOO : Besoin d'aide pour le nom de l'axe des Y !!!!!

Analyse n°1 Comme dit précédemment, dans cette analyse nous nous sommes concentrés sur les gènes **présents dans la liste cilliome** sur lesquels **au moins deux** de nos patients présentaient au moins 1 variant **homozygote** ayant un effet tronquant sur la protéine. Nous avons ainsi pu identifier les 4 gènes suivants : ARMC2, CCDC146, CFAP44 et TTC29 :

1. ***ARMC2*** : Sur ce gène, 2 patients portent un variant homozygote tronquant : *Ghs37* et *Ghs93*. Le patient et *Ghs37* portent un indel créant un décalage du cadre de lecture tandis que et *Ghs93* porte un variant créant un codon stop prématuré (**Table :** 4.8). Le patient et *Ghs107* porte quant à lui un variant faux-sens prédit comme *deleterious* par SIFT et *probably damaging* par PolyPhen. Nous pouvons également noter qu'aucun des variants portés par ces 3 sont absents des 3 bases de données. Les arguments génétiques mentionnés couplés à la forte expression testiculaire de ce gène (**Figure :** 4.17 - **B**) font de celui-ci un très bon candidat pour expliquer le phénotype de ces 3 bien que des analyses fonctionnelles soient nécéssaires pour 1 d'entre eux.
2. ***CFAP44*** : Comme *ARMC2*, 2 patients portent un variant homozygote tronquant et 1 porte un variant homozygote non tronquant. et *Ghs22* porte un variant stop et et *Ghs34* un variant affectant un site donneur d'épissage. Le fait qu'aucun de ces variant ne soir répertorié dans aucune des bases de données laisse supposer qu'ils sont tout deux très rares. Le patient et *Ghs89* porte un variant faux-sens prédit comme *tolerated* par SIFT et *possibly damaging* par PolyPhen. Les preuves fortes impliquant ce gène dans le cilliome humain ainsi que les effets délétères retrouvé chez 2 des 3 portant des mutations sur ce gène font de *CFAP44* un autre très bon candidat malgré son expression non spécifique au testicule (**Table :** 4.8 et 4.9, **Figure :** 4.17 - **B**).
3. ***CCDC146*** : Sur ce gène, seuls les patients *Ghs32* et *Ghs35* sont retrouvés mutés. Cependant, ces deux patients portent respectivement un variant induisant un codon stop prématuré et un décalage du cadre de lecture entrainant tout deux la production d'une protéine tronquée. Le premier de ces variants est retrouvé à une fréquence très faible à la fois dans ESP et ExaC, tandis que le second est totalement absent de l'ensemble des bases de donnés. On peut égallement ajouter que ce gène décrit comme faisant partie du cilliome humain avec de fortes preuves présente égallement une expression testiculaire relativement élevée. Pour finir, la protéine CCDC146 codée par le gène *CCDC146* avait déjà été décrite comme composant du centrosome spermatique, un organite ayant un rôle dans l'orientation des cellules et étant à l'origine des cils et des flagelles (Firat-Karalar, Sante, Elliott, & Stearns, 2014) reforçant ainsi les arguments de l'implication de ce gène dans le phénotype MMAF nous permettent ainsi d'affirmer que ces variants sont responsables du phénotype MMAF de ces 2 patients (**Table :** 4.8 et 4.9, **Figure :** 4.17 - **B**).
4. ***TTC29*** : Sur ce gène, les patients *Ghs19* et *Ghs26* portent la même variation retrouvée à très faible fréquence dans les trois base de données et impactant un site donneur d'épissage du transcrit induisant la production d'une protéine

aberrante. Ce gène à très forte expression testiculaire avait déjà été décrit en 2014 comme localisant au niveau de l'axonème du flagelle et qu'un *knock-down* entrainait des defauts de la cilliogénèse (Chung et al., 2014) (**Table :** 4.8 et 4.9, **Figure :** 4.17 - **B**).

Cette première analyse au cours de laquelle nous avons sélectioné les gènes retrouvés présents dans la liste cilliome et sur lesquels au moins deux de nos patients présentaient au moins 1 variant homozygote ayant un effet tronquant sur la protéine nous a permis de mettre en évidence `rn_gene_grp1_high` nouveaux gènes candidats : *ARMC2*, *CCDC146*, *CFAP44* et *TTC29* retrouvés mutés à l'état homozygote chez 10 de nos patients dont 8 avec des variants tronquants soit 13.1 % des patients resant dans notre cohorte. Pour les 2 autres, des analyses fonctionelles sont nécéssaires afin de pouvoir être sûr ce leurs variants sont bien responsables de leur phénotype. La cause génétique responsable du phénotype des patients *Ghs19*, *Ghs22*, *Ghs26*, *Ghs32*, *Ghs34*, *Ghs35*, *Ghs37* et *Ghs93* ayant été identifiées avec certitude, l'ensemble de leur données de variants sont retirées de nos liste réduisant ainsi celle-ci à 1331 chevauchant 1164 et répartis sur 53. Les données des patients Ghs107 et Ghs89 sont elle conservées afin de voir si un meilleur candidat pourrait expliquer le phénotype de ces patients.

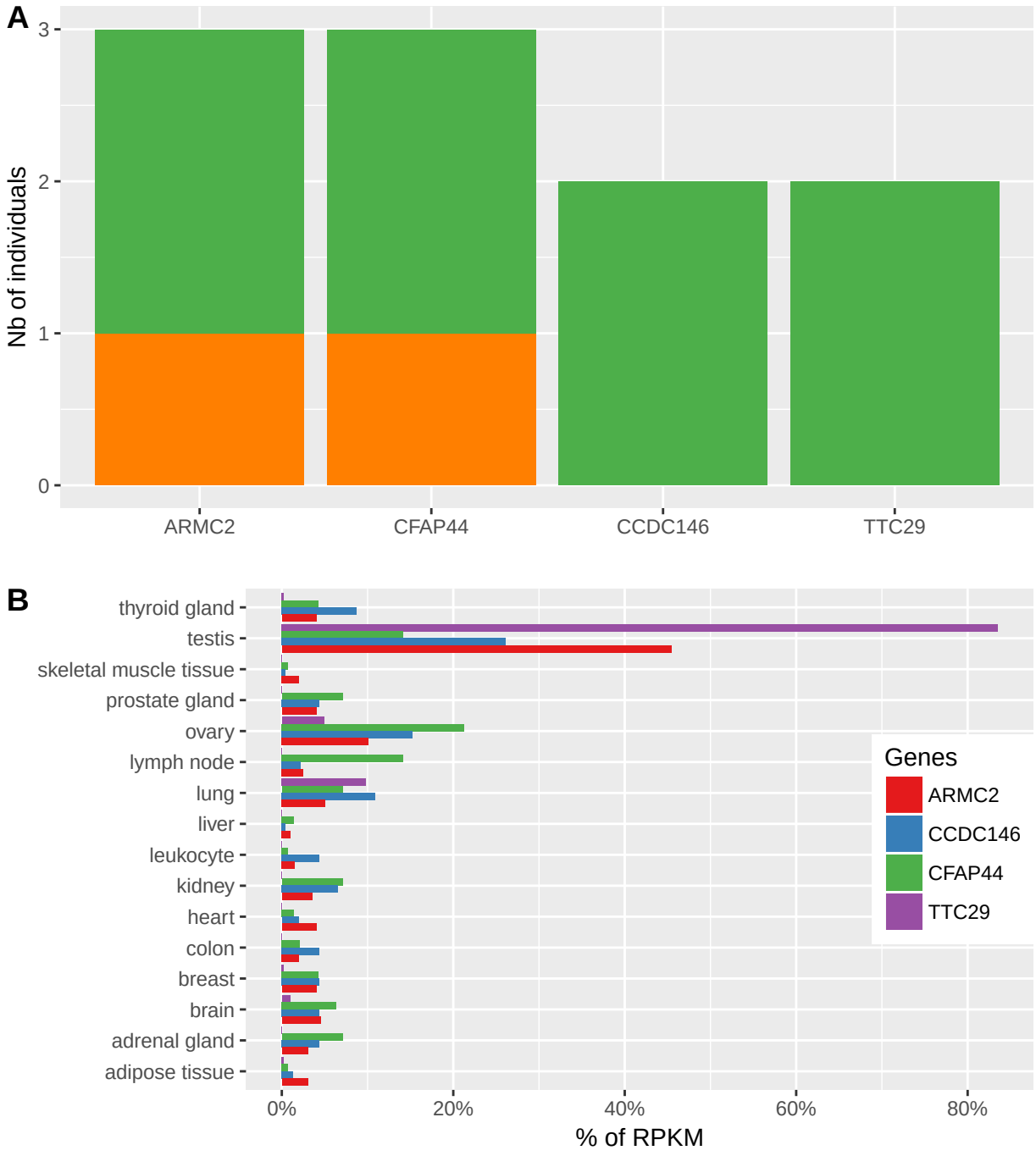


Figure 4.17 – Analyse des gènes séléctionnés dans l'Analyse n°1 : Expression tissulaire des gènes retenus d'après les données du projet de transcriptome Illumina bodyMap. Résumé de l'Analyse 1, quantification du nombre de patients retrouvés mutés sur chacun des gènes retenus ainsi que du degré de confiance accordé à la cause génétique

Table 4.8 – List des gènes présents dans la liste ciliome sur lesquels au moins deux patients portent une mutation tronquante homozygote

Patient	Gene	Evidence	Consequence	ESP	1KG	ExAC
Ghs37	ARMC2	Novel	frameshift	none	none	none
Ghs93	ARMC2	Novel	splice donor	none	none	none
Ghs32	CCDC146	Strong	stop gained	1e-04	none	2.47e-05
Ghs35	CCDC146	Strong	frameshift	none	none	none
Ghs22	CFAP44	Strong	stop gained	none	none	none
Ghs34	CFAP44	Strong	splice donor	none	none	none
Ghs19	TTC29	Strong	splice donor	0.0012	5e-04	0.000158
Ghs26	TTC29	Strong	splice donor	0.0012	5e-04	0.000158

Table 4.9 – Liste des patients portant un variant non troquant homozygote sur un des gènes suivant : ARMC2, CCDC146, CFAP44 et TTC29

Patient	Gene	Consequence	SIFT	PolyPhen	ESP	1KG	ExAC
Ghs89	CFAP44	missense	tolerated	possibly damaging	0.0012	0.0014	0.000692
Ghs89	CFAP44	missense	tolerated	probably damaging	0.0012	0.0014	0.000692
Ghs107	ARMC2	missense	deleterious	probably damaging	none	none	none

Analyse n° 2 Pour rappel, au cours de cette analyse, nous avons étudié les gènes **absents dans la liste cilliome** mais sur lesqules on trouvait toujours **au moins deux** de nos patients présentant au moins 1 variant tronquant à l'état **homozygote**. 17 patients différents portaient ainsi au moins un variant homozygote tronquant sur l'un des 8 gènes suivants : *BAZ1A*, *CCDC129*, *CFAP43*, *FSIP2*, *ICA1*, *NACA*, *SART3* et *TRAV26-1*.

1. ***CFAP43*** : 7 patients portent au moins 1 variant sur le gène *CFAP43*. Parmis ceux-ci, les patients *Ghs102*, *Ghs105*, *Ghs126*, *Ghs17* et *Ghs41* portent une mutation tronquante à l'état homozygote soit absente des bases de données soit présentent avec une très faible fréquence. Le patient et *Ghs25* lui porte un variant homozygote intronique au situé au sein de la région d'épissage (et non sur un site donneur ou accepteur). Bien que ce type de variants puissent effectivement avoir un impact sur l'epissage, il pourrait également être sans effet, or, il est difficile de le prédire à ce stade. (TODO : : faire tourner un algo de prédiction). Le patient et *Ghs132* en revanche semble plus intéréssant, puisque celui-ci porte deux variant hétérozygotes sur le gène *CFAP43* parmis lesquels un est un indel entrainant un décalage du cadre de lecture tandis que l'autre est un faux-sens prédit comme *possibly damaging* par Polyphen, bien qu'il soit annoté *tolerated* par SIFT. Malgré une expression ubiquitaire (**Figure :** 4.18 - **B**), le nombre important de patients portant des variant sur celui-ci, notamment 5 potant des variants tronquants homozygote font de ce gène un bon candidat.
2. ***FSIP2*** : Comme pour le gène *CFAP43*, 7 de nos patients portaient 1 variant sur le gène *FSIP2* cependant pour seulement 2 d'entre eux ce variant était tronquant à l'état homozygote. En effet, les patients *Ghs20* et *Ghs21* porte tout deux un indel entrainant un décalage du cadre de lecture dont aucun n'est répertorié dans les bases de données. Le patient et *Ghs131* porte lui un faux-sens homozygote prédit comme *benign* par PolyPhen. 4 autres patients portent au moins deux variants hétérozygote sur ce gène, cependant la plupart sont des faux-sens prédit égallement comme *benign* par PolyPhen. Bien que l'effet sur la protéine des variants portés par 5 des 7 patients portant au moins un variant sur ce gène soit incertains, les variants tronquant portés par les patients *Ghs20* et *Ghs21* ainsi que saforte expression testiculaire et le fait que son implication dans la structure de la gaine fibreuse du flagelle spermatique ait été montrée en 2003 (Brown, Miki, Harper, & Eddy, 2003) font de ce gène un excellent candidat pour expliquer le phénotype d'au moins 2 patients. Pour les autres, des analyses fonctionelles seront nécéssaires.
3. ***SART3*** : Un total de 4 patients portaient **le même** variants homozygote. Ces patients étant tous issus du même projet de séquençage (Strasbourg 2012) ce variant pourrait parfaitement être un artefact dû au protocol de séquençage. Cette hypothèse est d'autant plus probable que l'ensemble des patients séquencés dans ce projet présentaient un phénotype MMAF, ainsi, aucun des individus de notre cohorte contrôle n'a donc pu servir à filtrer les variants artefactuels issus de ce projet. De plus, on peut constater que l'expression tissulaire de ce gène

est ubiquitaire. Pour ces différentes raisons, ce gène n'a pas été retenu comme candidat dans nos analyses.

4. ***ICA1* et *TRAV26-1*** : Ces deux gènes sont chacun retrouvés présentant des variants homozygotes tronquant chez 6 de nos patients. Pour chaque gène on peut noter que c'est à chaque fois le même variant qui est partagé par les trois patients ce qui est asses étonnant étant doné le fait que ces patients ne sont pas apparentés et n'on pas été séquencé lors du même projet de séquençage. De plus, on peut observer que le patient et *Ghs40* porte un variant sur chacun de ces deux gènes, de même qu'il en portait également un sur le gène *SART3*. L'expression testiculaire de ces deux gènes apparait elle aussi très faible. De plus, la bibliographie du gène *ICA1* révèle que celui-ci est principalement exprimé dans le pancréas (Mally, Cirulli, Hayek, & Otonkoski, 1996, Stassi, Schloot, & Pietropaolo (1997)) et que celui-ci serait à la fois lié au diabète de type 1 (Martin et al., 1995, R. Gaedigk, Duncan, Miyazaki, Robinson, & Dosch (1994), Martin, Lampasona, Dosch, & Pietropaolo (1996)) et au syndrome de Gougerot-Sjögren (S. Winer et al., 2002). Malgré le nombre important de patients présentant des variants homozygotes tronquant sur ces deux gènes, ceux-ci n'ont pas été retenus comme candidat pour expliquer le phénotype MMAF de ces 3 patients.

5. ***BAZ1A*** : Sur ce gène, les patient *Ghs18* et *Ghs94* portent tout deux le même indel à l'état homozygote. Celui-ci entrainant un décalage du cadre de lecture induit la formation d'une protéine abérente. l'absence de cette indel dans les trois bases de données laisse supposer que une faible fréquence de ce variant dans la population générale. Bien que ce gène présente une faible expression testiculaire, une étude de 2013 effectuée chez la souris a démontré son rôle majeur dans la spermatogénèse, les souris *KO Baz1aˆ{-/-}* présentaient de nombreeux defauts tel que des des spermatozoïdes non motiles, présentant une morphologie de la tête et du flagelle abérante (Dowdle et al., 2013) faisant ainsi de ce gène un bon candidat pour expliquer le phénotype de ces 2 patients.

6. ***CCDC129*** : Ce gène est retrouvé muté à l'état homozygote chez 2 patients porteur du même indel, non repertorié dans les bases de données, causant un décalage du cadre de lecture. On peut égallement constate que ce gène possède une expression testiculaire forte et exclusive faisant ainsi de ce gène un bon candidat malgré une littérature pauvre à son sujet.

7. ***NACA*** : 2 patients paratagent le même variant homozygote causant un décalage du cadre de lecture sur le gène *NACA*. On peut noter que les patient *Ghs37* et *Ghs37* portaient porteur respectifs de variant homozygote tronquant sur les gène *CCDC146* et *ARMC2* portaient égallement ce même variant. L'ensemble de ces patients provenant du même projet de séquençage (Strasbourg 2012), cela laisse supposer que comme pour *SART3*, ce variant est artefactuel. On peut dailleur noter que le patient *Ghs41* était lui aussi porteur du variant sur *SART3* renforçant ainsi l'hypothèse de l'erreur de séquençage. Pour ces raison, ce gène n'a pas été retenu en tant que candidat.

Dans cette analys, 8 gène ont dans un premier temps été identifiés. Cependant, une analyse plus approfondie a fait que seuls les gènes *BAZ1A*, *CCDC129*, *CFAP43* et *FSIP2* ont été retenus. Les variants présent sur les gènes *SART3* et *NACA* étant probablement des artefacts dus aux erreurs de séquençage. Dès lors, seuls 4 des 8 gènes identifiés présentaient des arguments suffisemment convainquant pour être considéré comme responsable du phénotype MMAF. Ainsi, 18 de nos patients sont porteurs de variants sur l'un de ces gènes dont 13 à l'état homozygote avec nottamment les patients *Ghs102*, *Ghs105*, *Ghs126*, *Ghs132*, *Ghs17*, *Ghs18*, *Ghs20*, *Ghs21*, *Ghs41*, *Ghs91* et *Ghs94* porteur de variants homozygotes tronquants. Les gènes *BAZ1A*, *CCDC129*, *CFAP43* et *FSIP2* étant de bons candidats et les données génétiques des 11 patients précédemment cités étant suffisement fortes, l'ensemble de leurs variant furent ainsi retirés de notre liste contenant désormais 1093 variants et 962 gènes différents.

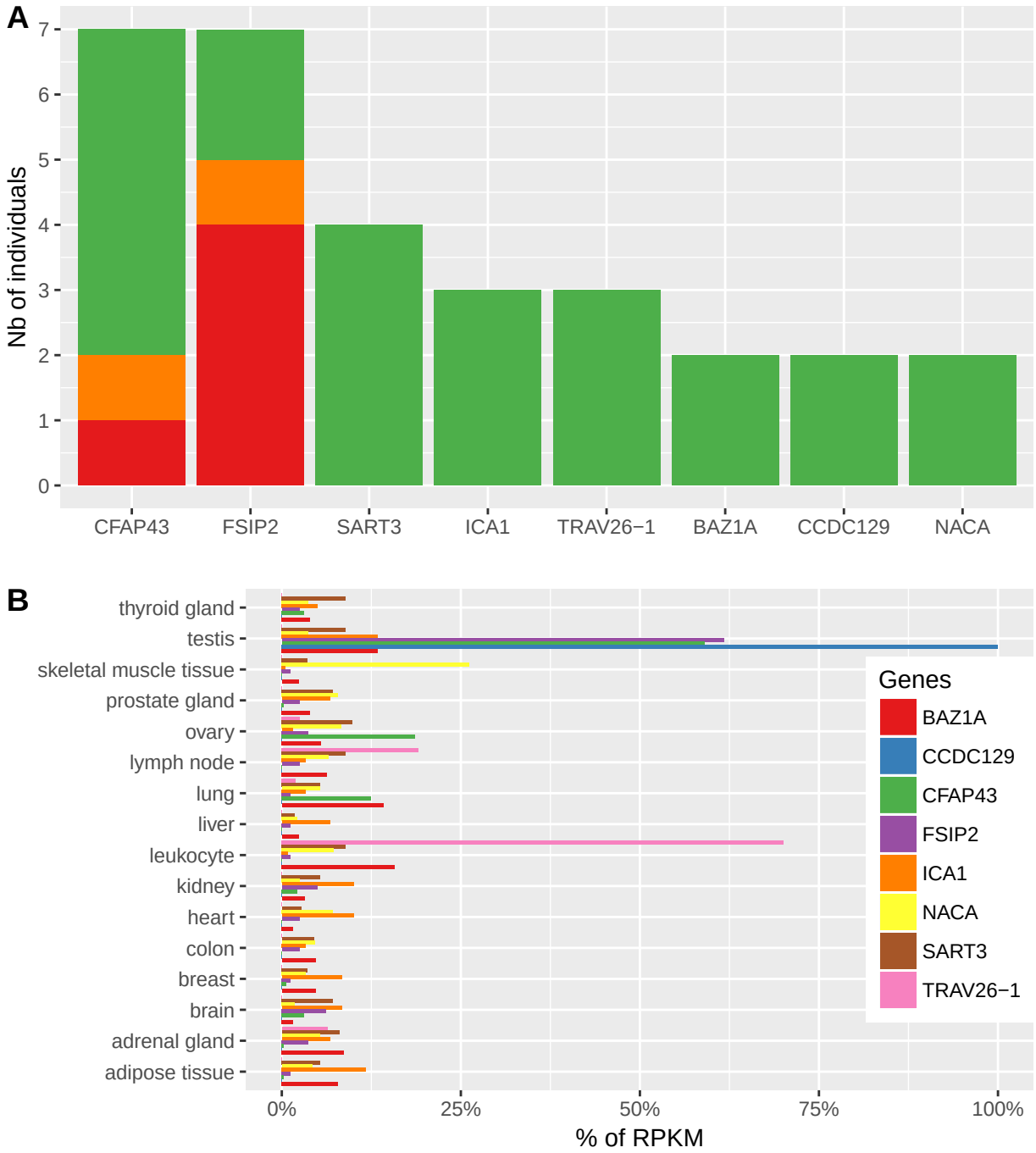


Figure 4.18 – Analyse des gènes séléctionnés dans l'Analyse n°2 : Expression tissulaire des gènes retenus d'après les données du projet de transcriptome Illumina bodyMap. Résumé de l'Analyse 2, quantification du nombre de patients retrouvés mutés sur chacun des gènes retenus ainsi que du degré de confiance accordé à la cause génétique

Table 4.10 – List des gènes sur lesquels au moins deux patients portent une mutation tronquante non présents dans la liste ciliome

Patient	Gene	Consequence	ESP	1KG	ExAC
Ghs18	BAZ1A	frameshift	none	none	none
Ghs94	BAZ1A	frameshift	none	none	none
Ghs91	CCDC129	frameshift	none	none	none
Ghs132	CCDC129	frameshift	none	none	none
Ghs17	CFAP43	stop gained	2e-04	none	9.88e-05
Ghs41	CFAP43	stop gained	none	none	8.24e-06
Ghs102	CFAP43	frameshift	none	none	none
Ghs105	CFAP43	splice acceptor	none	none	none
Ghs126	CFAP43	stop gained	none	none	3.29e-05
Ghs20	FSIP2	frameshift	none	none	none
Ghs21	FSIP2	frameshift	none	none	none
Ghs17	ICA1	frameshift	none	none	none
Ghs40	ICA1	frameshift	none	none	none
Ghs105	ICA1	frameshift	none	none	none
Ghs39	NACA	frameshift	none	none	none
Ghs41	NACA	frameshift	none	none	none
Ghs31	SART3	splice donor	none	none	none
Ghs38	SART3	splice donor	none	none	none
Ghs40	SART3	splice donor	none	none	none
Ghs41	SART3	splice donor	none	none	none
Ghs40	TRAV26-1	frameshift	none	none	none
Ghs43	TRAV26-1	frameshift	none	none	none
Ghs55	TRAV26-1	frameshift	none	none	none

Table 4.11 – Liste des patients portant un variant non troquant homozygote sur un des gènes suivant : TODOOOOOO

Patient	Gene	Consequence	SIFT	PolyPhen	ESP	1KG	ExAC
Ghs25	CFAP43	splice region	none	none	none	none	none
Ghs131	FSIP2	missense	none	benign	none	none	0.00121

Table 4.12 – Liste des patients portant un variant non troquant homozygote sur un des gènes suivant : TODOOOOOO

Patient	Gene	Consequence	SIFT	PolyPhen	ESP	1KG	ExAC
Ghs40	FSIP2	missense	none	benign	none	0.0056	0.00157
Ghs40	FSIP2	missense	none	probably damaging	none	none	none
Ghs40	FSIP2	missense	none	benign	none	none	none
Ghs40	FSIP2	missense	none	benign	0.0034	0.0056	0.0019
Ghs92	FSIP2	missense	none	benign	none	0.0056	0.00157
Ghs92	FSIP2	missense	none	benign	0.0034	0.0056	0.0019
Ghs95	FSIP2	missense	none	benign	none	0.0056	0.00157
Ghs95	FSIP2	missense	none	benign	0.0034	0.0056	0.0019
Ghs101	FSIP2	missense	none	unknown	none	none	none
Ghs101	FSIP2	missense	none	benign	none	none	none
Ghs132	CFAP43	frameshift	none	none	none	none	none
Ghs132	CFAP43	missense	deleterious	probably damaging	2e-04	none	7.41e-05
Ghs132	CFAP43	missense	none	benign	2e-04	none	7.41e-05
Ghs132	CFAP43	missense	tolerated	possibly damaging	2e-04	none	7.41e-05

Analyse n°3 Dans cette un troisième analyse, nous avons séléctionné à nouveau les gènes **présents dans la liste cilliome** en conservanrt cett efois-ci ceux sur lesquels **un seul** de nos patients présentaient au moins 1 variant tronquant à l'état **homozygote**. Les 4 gènes suivant correspondaient à ces critères : *C21orf59*, *C6orf118*, *CCDC65* et *SPEF2* tous présentant de fortes preuves indiquant leur appartenancez au ciliome humain.

1. ***C6orf118*** : Ce gène a été retrouvémuté à l'état homozygote chez le patient et *Ghs40* qui portait une substitution entrainant la formation d'un codon stop prématuré. Il faut noter que ce même patient portait égallement des variants homozygotes tronquant sur les gènes *SART3*, *TRAV26-1*, et *ICA1* cependant la forte expression testiculaire de ce gène en fait un meilleur candidat. Le patient et *Ghs27* quant à lui porte deux variants faux-sens hétérozygote Le premier étant prédit probably *damaging* par PolyPhen et *tolerated low confidence* par SIFT tandis que le second est prédit *possibly damaging* et *tolerated*, il est dificile de se prononcer quant à l'effet délétère de ces deux variants sans effectuer d'analyses fonctionelles. Il faut noter que *C6orf118* présente une forte expression dans le poumoun. De plus, ce gène a récemment été décrit comme étant associé au phénotype de tuberculose pulmonaire (E. P. Hong, Go, Kim, & Park, 2017). Cependant cela n'est en rien contradictoire avec le phénotype MMAF de ces 2 patients, le poumon comprenant de nombreuse cellules ciliées, notamment au niveau de l'épithélium respiratoire, il n'est donc pas surprenant que des gènes du flagelle aient égallement une fonction au sein d'autres organes ciliés.
2. ***C21orf59*** :
3. ***CCDC65*** : En effet, *CCDC65* présente une forte et quasi exclusive expression testiculaire de plus, la protéine NYD-SP28 (ancien nom de CCDC65) avait déjà caractérisé comme faisant partie du flagelle spermatique (Y. Zheng et al., 2006)
4. ***SPEF2*** :
5. **Étape n°1** : La recherche de variants homozygote tronquant nous a permis d'identifier 4 patients d'un variant sur un des 4 gène suivant : C6orf118, C21orf59, CCDC65, SPEF2. Cependant, l'analyse de l'expression testiculaire de ces gènes nous a insitée à n'en garder que CCDC65, C6orf118 d'entre eux : CCDC65, C6orf118 (**Table :** 4.13, **Figure :** 4.19 - **A**). . En revanche .
6. **Étape n°3** : La recherche de potentiels hétérozygotes composites s'est elle révélée plus fructueuse puisque 1 de nos patients portait deux variants faux-sens sur le gene *C6orf118* à l'état hétérozygote. Le premier de ces variants étant prédit *probably damaging* par PolyPhen et *tolerated low confidence* par SIFT tandis que le second est prédit *possibly damaging* et *tolerated*, il est dificile de se prononcer quant à l'effet délétère de ces deux variants (**Table :** 4.14).
7. **Étape n°4** : Comme pour l'analyse n°2, nous avons ici décidé de ne retirer de notre liste de variant que ceux présents chez les patients identifiés au cours de

l'étape n°1. Cela nous a ainsi permis de réduire à nouveau cette liste arrivant ainsi à 1011 variants cheveauchant 891 gènes différents.

Analyser les gènes de la liste ciliome sur lesquels *un seul* patient portait un variant tronquant à l'état homozygote nous a permis d'identifié 4 parmis lesquels 2 présentaient une forte expression testiculaire. L'analyse de ces 2 derniers gènes nous a permis d'identifier 2 patients chacun portant un variant tronquant homozygote sur l'un de ces 2 gènes. Étendre notre recherche à la fois aux variants tronquant ainsi qu'aux hétérozygote composites nous a permis d'identifier 1 patient pour qui deux faux-sens hétérozygotes sur le gène *C6orf118* pourraient expliquer le phénotype (**Figure :** 4.19 - **B**).

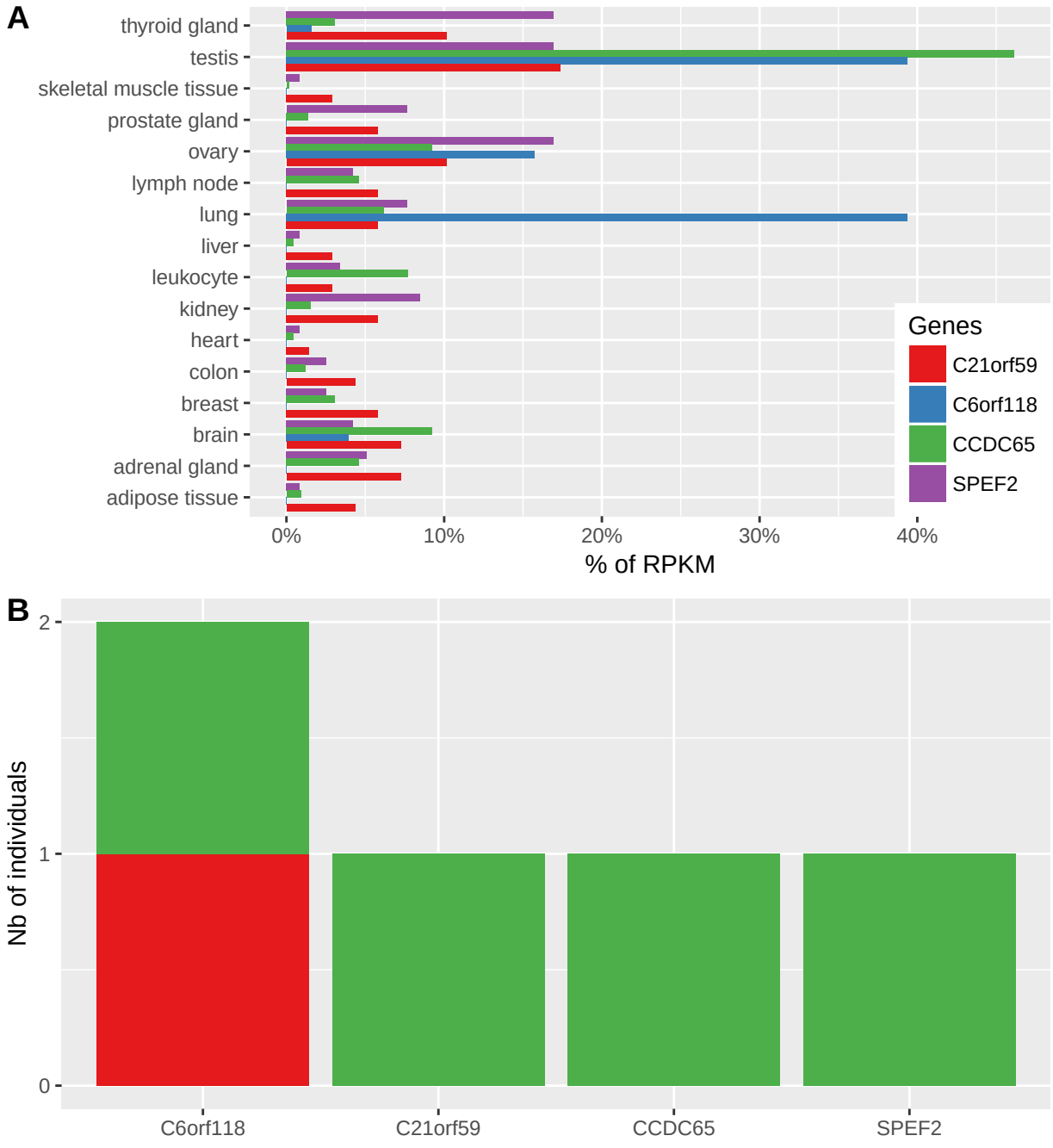


Figure 4.19 – Analyse des gènes séléctionnés dans l'Analyse n°3 : **A** : Expression tissulaire des gènes retenus dans cette analyse d'après les données du projet de transcriptome Illumina bodyMap. **B** : Résumé de l'Analyse 3, quantification du nombre de patients retrouvés mutés sur chacun des gènes retenus ainsi que du degré de confiance accordé à la cause génétique

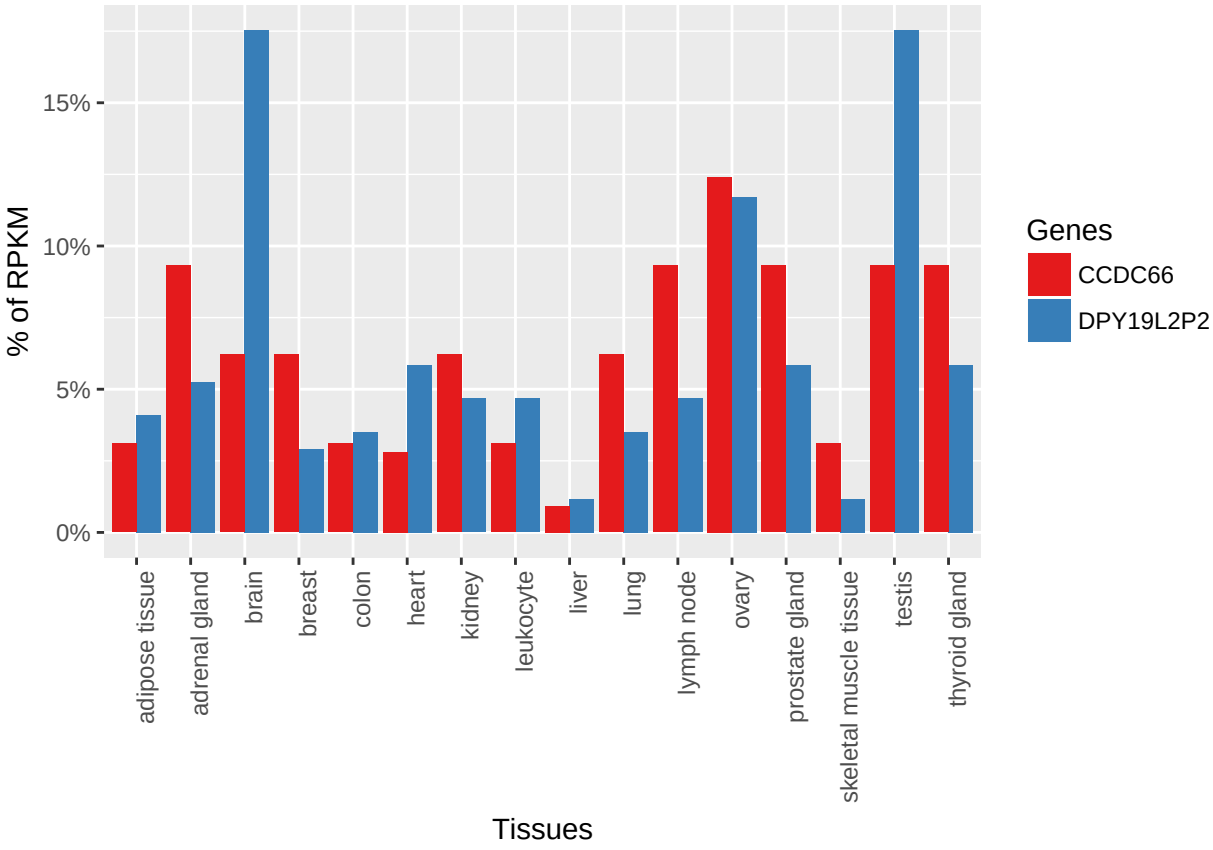
Table 4.13 – Analyse n°3 : List des gènes présents dans la liste ciliome sur lesquels un seul patient portent une mutation homozygote tronquante

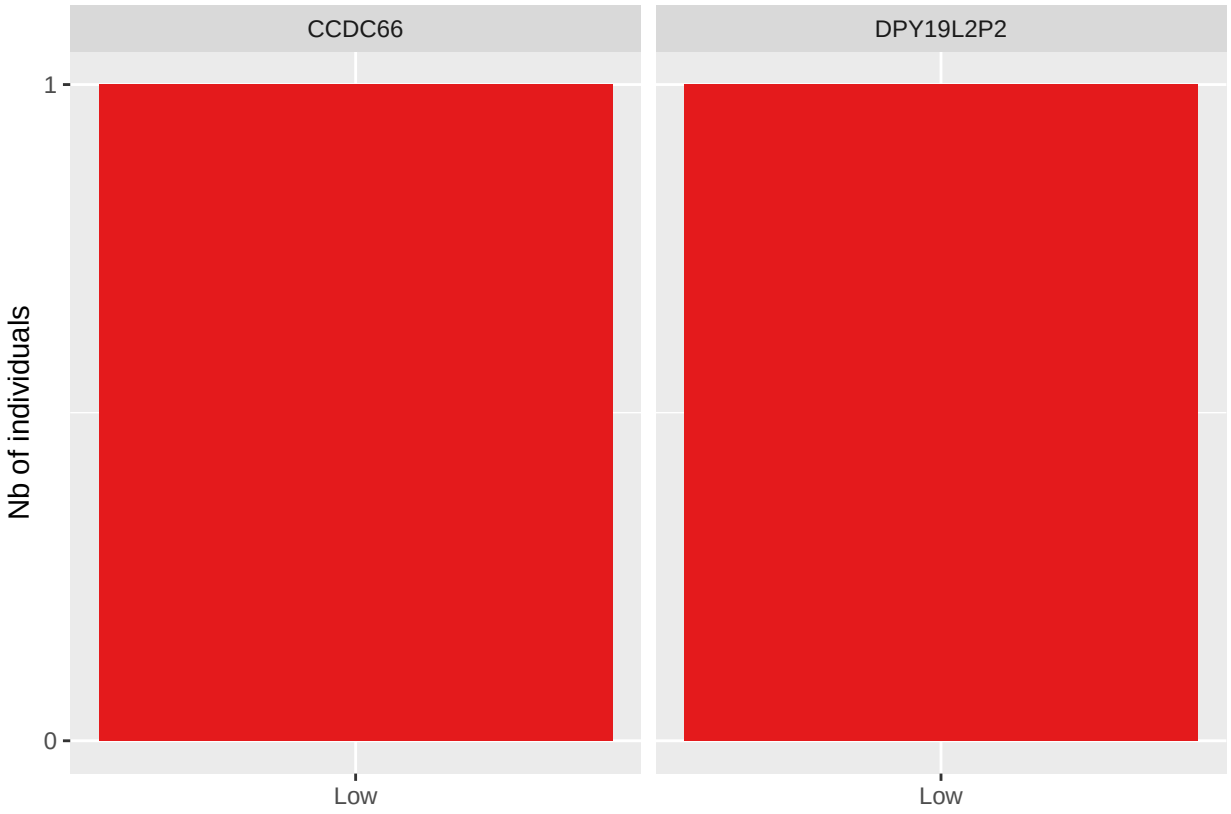
Patient	Gene	Evidence	Consequence	ESP	1KG	ExAC
Ghs88	C21orf59	Strong	frameshift	none	none	none
Ghs40	C6orf118	Strong	stop gained	none	none	8.24e-06
Ghs127	CCDC65	Strong	frameshift	none	none	none
Ghs131	SPEF2	Strong	frameshift	none	none	none

Table 4.14 – Analyse n°3 : Liste des patients portant au moins deux variants hétérozygotes sur un des gènes suivant : *C21orf59*, *C6orf118*, *CCDC65* et *SPEF2*

Patient	Gene	Consequence	SIFT	PolyPhen	ESP	1KG
Ghs27	C6orf118	missense	tolerated low confidence	probably damaging	8e-04	0.004
Ghs27	C6orf118	missense	tolerated	possibly damaging	0.0038	9e-04

Analyse n°4


% of RPKM
15%
10%
5%
0%
adipose tissue
adrenal gland
brain
breast
colon
heart
kidney
leukocyte
liver
lung
lymph node
ovary
prostate gland
skeletal muscle tissue
testis
thyroid gland
Tissues
Genes
CCDC66
DPY19L2P2


CCDC66
DPY19L2P2
1
0
Nb of individuals
Low
Low

Pour finir nous avons dans cette analyse séléctionné l'ensemble des variants chevauchant des gènes **absents dans la liste cilliome** sur lesquels **un seul** de nos patients présentaient au moins 1 variant tronquant à l'état **homozygote**. Cela nous a permis d'obtenir une liste de 0 gènes différents retrouvés mutés chez 0 de nos patients.

1. **Étape n°1** : En raison du grand nombre de gène ayant remplis ces critères

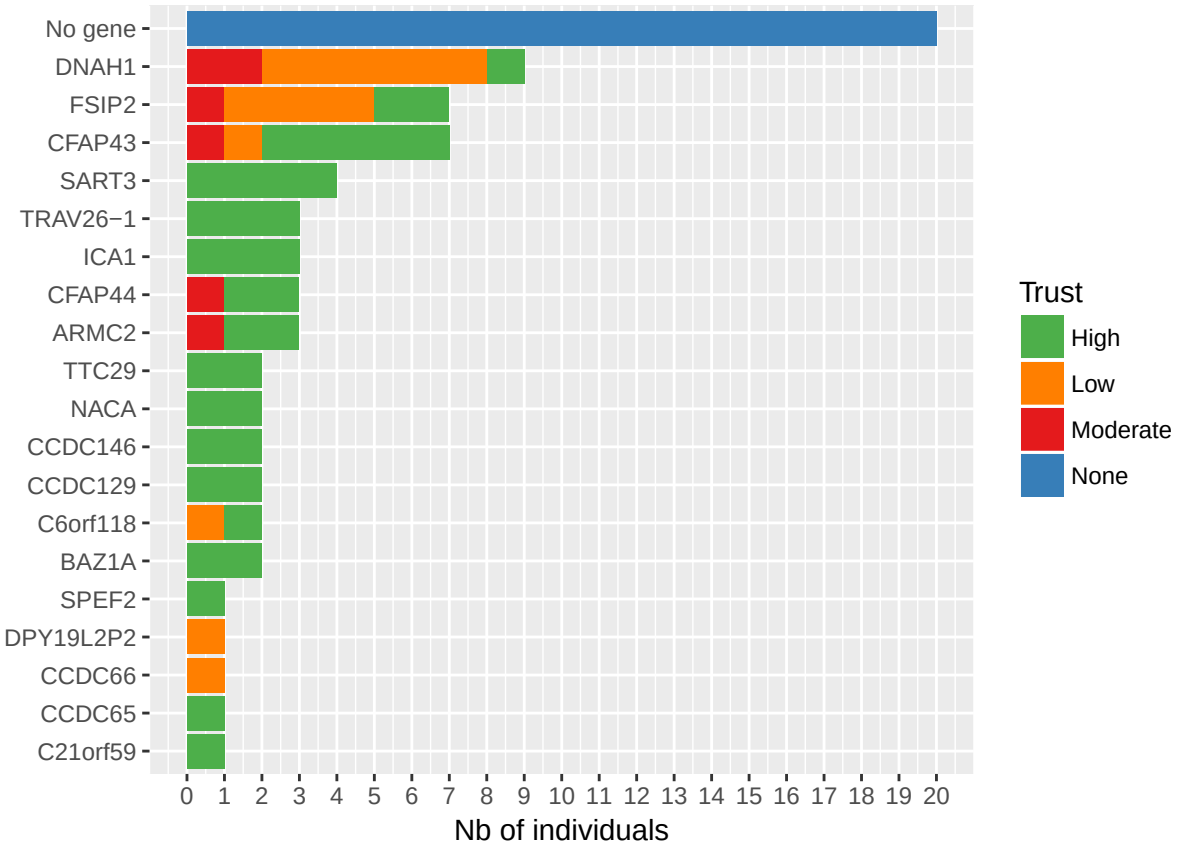
2. **Étape n°2** : Nous avons donc cherché des patients portant au moins 1 variant homozygote non tronquant sur l'un de ces 0. Nous avons ainsi pu identifier 0

Table 4.15 – Liste des patients portant un variant non troquant homozygote sur un des gènes suivant : TODOOOOOO

Gene	Patient	Consequence	SIFT	PolyPhen

```
## Warning in bind_rows_(x, .id): binding factor and character vector,
## coercing into character vector
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```



Discussion

L'analyse de cette cohorte de 62 patients MMAF nous dans un premier temps permis de confirmer l'importance de l'implication du gène *DNAH1* dans ce phénotype grâce à ... patients présentant des variants sur ce gène dont ... à l'état homozygote. Elle nous a également permis d'identifier 7 nouveaux gène candidats pouvant expliquer le phénotype de 29 de nos patients soit 47 % de la cohorte. parmi ceux-ci, ... portaient au moins un variant homozygote sur un de ces gènes. Pour les autres des études sont nécessaires afin de déterminer si les différents variants hétérozygotes qu'ils portent sont situés sur leurs deux allèles différents faisant d'eux des hétérozygotes composites (**Figure : ?? - A**).

Parmi cet ensemble de patients, il faut noter que 5 d'entre eux porte des variants pouvant expliquer leur phénotype sur plusieurs des gènes candidats que nous avons identifiés. En effet, 5 de nos patients portent des variants sur deux de nos gènes candidats et 0 sur 3 d'entre eux (**Figure : ?? - B**).

Cependant, parmi ces différents variants certains semblent plus probables pour expliquer le phénotype (**Table : ??**) avec par exemple :

1. Patient Ghs105 : Ce patient porte à la fois un variant homozygote affectant le site d'épissage du gène *WDR96* ainsi que deux variants hétérozygotes causant tous les deux un faux-sens sur le gène *EFCAB6*. Au vu du génotype homozygote et de l'effet tronquant du variant impactant *WDR96*, il parait plus probable que celui-ci soit responsable du phénotype MMAF au détriment des deux variants hétérozygotes chevauchant *EFCAB6*.

2. Patient Ghs17 : Deux variants homozygotes ont été retenus pour ce patient. L'un causant un faux-sens sur le gène *EFCAB6* l'autre créant un codon stop prématuré sur *WDR96*. Ici aussi, au vu de l'impact délétère du variant chevauchant *WDR96*, il parait plus probable que ce soit celui-ci qui soit la cause du phénotype de ce patient.

3. Patients Ghs32 et Ghs35 : Ces patients portent tous deux à la fois des variants sur le gène *ANKRD20A3* et sur le gène *CCDC146* cependant l'effet tronquant de leur variant impactant ce dernier nous laisse penser que ceux-ci soient la cause du phénotype MMAF de ces deux patients.

En procédant de la même manière pour les autres patients il est, dans la grande majorité des cas, possible de dégager un gène pour lequel l'implication dans le phénotype des patients parait plus évidente bien que des analyses complémentaires soient nécessaires.

Ainsi, cette analyse révèle l'efficacité de notre pipeline puisqu'elle a permis d'identifier au moins un gène candidat pour 47 % de nos patients. Pour les autres des analyses

individuelles complémentaires sont nécessaires afin d'identifier la cause génétique responsable de leur phénotype.

Une partie de ces différents résultats ont déjà été publiés dans deux articles dont je suis co-auteur :

1. **Whole exome cohort study and analysis of mouse and Trypanosoma models demonstrate the importance of WDR proteins in flagellogenesis and male fertility**, *Nat Genet* (soumis) : Cette article présente nos différents résultats dans la caractérisation des gènes *WDR96* et *WDR52* ainsi que les différentes preuves de leur implication dans le phénotype MMAF.
2. **Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations** : En plus des résultats évoqués précédemment pour la famille MMAF2, cet article inclus ceux de ... patients de cette cohorte présentant des variants sur le gène *DNAH1*

Pour les autres, notre équipe travaille actuellement à la caractérisation des différents gènes afin de comprendre les processus moléculaires

4.3 Conclusion

Au cours de ces différentes études nous avons pu identifier les variants pouvant expliquer les phénotypes de ... des différents patients que nous avons analysé que ce soit au sein d'études familiales ou bien au sein de plus large cohorte composés d'individus non apparentés. Bien que ces résultats soient satisfaisant, il faut noter que pour ... patients, soit ... % d'entre eux aucun candidat n'a pu à ce jour être identifié. Pour ces patients, le WES permets désormais de nouvelles approches permettant d'identifier de larges variants structuraux (insertion ou délétions) pouvant eux aussi être responsable du phénotype qui ne sont pas détectés par les analyses classiques. Néanmoins, il semble clair que des avancés soient encore nécessaires afin d'améliorer l'efficacité de ce genre d'étude notamment en créant de nouveaux filtres permettant ainsi d'épurer les listes de variants facilitant ainsi l'identification des gènes candidats.

Chapitre 5

MutaScript

Conclusion

Chapitre 6

The First Appendix

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., ... Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–9. http://doi.org/10.1038/nmeth0410-248

Baker, K. E., & Parker, R. (2004). Nonsense-mediated mRNA decay : terminating erroneous gene expression. *Current Opinion in Cell Biology*, *16*(3), 293–9. http://doi.org/10.1016/j.ceb.2004.03.003

Ben Khelifa, M., Coutton, C., Zouari, R., Karaouzène, T., Rendu, J., Bidart, M., ... Ray, P. F. (2014). Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. *American Journal of Human Genetics*, *94*(1), 95–104. http://doi.org/10.1016/j.ajhg.2013.11.017

Brown, P. R., Miki, K., Harper, D. B., & Eddy, E. M. (2003). A-Kinase Anchoring Protein 4 Binding Proteins in the Fibrous Sheath of the Sperm Flagellum. *Biology of Reproduction*, *68*(6), 2241–2248. http://doi.org/10.1095/biolreprod.102.013466

Chang, Y.-F., Imam, J. S., & Wilkinson, M. F. (2007). The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annual Review of Biochemistry*, *76*(1), 51–74. http://doi.org/10.1146/annurev.biochem.76.050106.093909

Chung, M.-I., Kwon, T., Tu, F., Brooks, E. R., Gupta, R., Meyer, M., ... Wallingford, J. B. (2014). Coordinated genomic control of ciliogenesis and cell movement by RFX2. *ELife*, *3*, e01439. http://doi.org/10.7554/eLife.01439

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., ... Pritchard, E. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. http://doi.org/10.1038/ng.806

Dowdle, J. A., Mehta, M., Kass, E. M., Vuong, B. Q., Inagaki, A., Egli, D., ... Keeney, S. (2013). Mouse BAZ1A (ACF1) is dispensable for double-strand break repair but is essential for averting improper gene expression during spermatogenesis. *PLoS Genetics*, *9*(11), e1003945. http://doi.org/10.1371/journal.pgen.1003945

Firat-Karalar, E. N., Sante, J., Elliott, S., & Stearns, T. (2014). Proteomic analysis of

mammalian sperm cells identifies new components of the centrosome. *Journal of Cell Science*, *127*(Pt 19), 4128–33. http://doi.org/10.1242/jcs.157008

Gaedigk, R., Duncan, A. M., Miyazaki, I., Robinson, B. H., & Dosch, H. M. (1994). ICA1 encoding p69, a protein linked to the development of type 1 diabetes, maps to human chromosome 7p22. *Cytogenetics and Cell Genetics*, *66*(4), 274–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8162706

Hong, E. P., Go, M. J., Kim, H.-L., & Park, J. W. (2017). Risk prediction of pulmonary tuberculosis using genetic and conventional risk factors in adult Korean population. *PloS One*, *12*(3), e0174642. http://doi.org/10.1371/journal.pone.0174642

Hu, Y., Yu, H., Shaw, G., Renfree, M. B., & Pask, A. J. (2011). Differential roles of TGIF family genes in mammalian reproduction. *BMC Developmental Biology*, *11*, 58. http://doi.org/10.1186/1471-213X-11-58

Imai, Y., Morita, H., Takeda, N., Miya, F., Hyodo, H., Fujita, D., ... Komuro, I. (2015). A deletion mutation in myosin heavy chain 11 causing familial thoracic aortic dissection in two Japanese pedigrees. *International Journal of Cardiology*, *195*, 290–292. http://doi.org/10.1016/j.ijcard.2015.05.178

Ivliev, A. E., 't Hoen, P. A. C., Roon-Mom, W. M. C. van, Peters, D. J. M., & Sergeeva, M. G. (2012). Exploring the Transcriptome of Ciliated Cells Using In Silico Dissection of Human Tissues. *PLoS ONE*, *7*(4), e35618. http://doi.org/10.1371/journal.pone.0035618

Kumar, P., Henikoff, S., & 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. http://doi.org/10.1038/nprot.2009.86

Lee, B., Park, I., Jin, S., Choi, H., Kwon, J. T., Kim, J., ... Cho, C. (2011). Impaired spermatogenesis and fertility in mice carrying a mutation in the Spink2 gene expressed predominantly in testes. *The Journal of Biological Chemistry*, *286*(33), 29108–17. http://doi.org/10.1074/jbc.M111.244905

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., ... Exome Aggregation Consortium, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–91. http://doi.org/10.1038/nature19057

Lunter, G., & Goodson, M. (2011). Stampy : A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. http://doi.org/10.1101/gr.111120.110

Mally, M. I., Cirulli, V., Hayek, A., & Otonkoski, T. (1996). ICA69 is expressed equally in the human endocrine and exocrine pancreas. *Diabetologia*, *39*(4), 474–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8777998

Martin, S., Kardorf, J., Schulte, B., Lampeter, E. F., Gries, F. A., Melchers, I., ... Pfützner, A. (1995). Autoantibodies to the islet antigen ICA69 occur in IDDM and in

rheumatoid arthritis. *Diabetologia*, *38*(3), 351–5. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7758883

Martin, S., Lampasona, V., Dosch, M., & Pietropaolo, M. (1996). Islet cell autoantigen 69 antibodies in IDDM. *Diabetologia*, *39*(6), 747. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8781774

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., ... Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. http://doi.org/10.1186/s13059-016-0974-4

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–51. http://doi.org/10.1038/nrg2986

Stassi, G., Schloot, N., & Pietropaolo, M. (1997). Islet cell autoantigen 69 kDa (ICA69) is preferentially expressed in the human islets of Langerhans than exocrine pancreas. *Diabetologia*, *40*(1), 120–2. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9028728

Su, Z., Łabaj, P. P., Li, S. S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., ... Shi, L. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, *32*(9), 903–14. http://doi.org/10.1038/nbt.2957

Winer, S., Astsaturov, I., Cheung, R., Tsui, H., Song, A., Gaedigk, R., ... Dosch, H.-M. (2002). Primary Sj ?gren's syndrome and deficiency of ICA69. *The Lancet*, *360*(9339), 1063–1069. http://doi.org/10.1016/S0140-6736(02)11144-5

Zheng, Y., Zhang, J., Wang, L., Zhou, Z., Xu, M., Li, J., & Sha, J.-H. (2006). Cloning and characterization of a novel sperm tail protein, NYD-SP28. *International Journal of Molecular Medicine*, *18*(6), 1119–25. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17089017