THÈSE

Thomas Karaouzene

31 octobre 2017

# Delete line 6 if you only have one advisor

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

## Introduction

En 2011, les bases moléculaires d'environ 3700 pathologies à transmission Mendélienne avaient été élucidées. Cependant, pour une quantité équivalente de pathologies Mendéliennes (ou suspectées de l'être) cette cause reste un mystère (Amberger, Bocchini, & Hamosh, [2011](#ref-Amberger2011)). Avec plusieurs centaines de pathologies caractérisées depuis 2010 (Ng et al., n.d.), les séquençages WGS et WES ont, depuis leur émergence, révolutionnés les méthodes de recherche dans le cadre d'étude phénotype-génotype en permettant de manière rapide et à moindre coup le séquençage de la quasi-totalité des gènes humains. Dès lors, le défis de ces analyses n'est plus le séquençage de l'ADN mais l'interprétation des données massives produites. En effet, l'un des plus grands challenges des analyses phénotype-génotype réalisées par WES réside dans l'analyse de l'importante quantité de variant portés par chaque individu s'élevant à plusieurs dizaines de milliers lorsque l'on compare avec le génome de référence. Même après avoir retiré les variants retrouvés fréquemment dans la population générale, des méthodes additionnelles sont nécessaires pour prédire, parmi les variants restant, lesquels induisent des conséquences fonctionnelles sérieuses afin de les prioriser (Pelak et al., [2010](#ref-Pelak2010)). De nombreux logiciels tel que Variant Effect Predictor (McLaren et al., [2016](#ref-McLaren2016)), SnpEff (Cingolani et al., [2012](#ref-Cingolani2012)) ou encore ANNOVAR (K. Wang, Li, & Hakonarson, [2010](#ref-Wang2010)) permettent d'identifier quels sont les variants qui ont un effet tronquant sur la protéine. Cependant, avec en moyenne 165 variants homozygotes ayant un effet tronquant retrouvés dans chaque exomes (Pelak et al., [2010](#ref-Pelak2010)) ces méthodes, bien qu'efficaces sont souvent insuffisantes.

D'autres logiciels tel que Exomiser (P. N. Robinson et al., [2014](#ref-Robinson2014)) vont, à partir d'une liste de variants **déjà** appelés effectuer les étapes d'annotation, de filtrage et de priorisation. Malgré l'efficacité de ces logiciels, aucun d'entre eux ne couvrent l'ensemble des étapes allant de l'alignement des *reads* à la priorisation des variants. La plupart ayant pour point de départ une liste de variants appelés en amont. Ils ne contrôlent donc en aucune manière les étapes d'alignement et d'appel des variants. Or, comme il a été dit plus tôt, ces deux étapes constituent la base de l'analyse [].

Ce chapitre décrit à la fois la constitution d’un pipeline d’analyse des données de séquençage exomique recouvrant l'ensemble des étapes allant de l'allignement des séquences à la priorisation des variants ainsi que son utilisation dans le cadre de la recherche de mutations entrainant différents phénotypes d’infertilité d'une part de cas familiaux composés de duos ou trio et, pour finir, d'une large cohorte d'individus non apparentés présentant tous le même phénotype.

## Résultats

### Etude d’une large cohorte de patients MMAF

#### Description de la cohorte

Après avoir mis en évidence l'implication du gène *DNAH1* dans le phénotype MMAF notre équipe s'est en partie spécialisé dans la caractérisation ce syndrome. Ainsi, entre 2012 et 2015, notre équipe a effectué le séquençage de 62 individus présentant tous ce phénotype afin d'en établir la cause génétique. Ces séquençages ont été effectué dans 3 centres différents que sont Genoscope, MountSinai et Strasbourg et sur une seule plateforme de séquençage, le Illumina HiSeq2000.

Table 1 Liste des différents projets de séquençages effectués

|  |  |  |  |
| --- | --- | --- | --- |
| Place | Year | Platform | Nb of individuals |
| MountSinai | 2012 | Illumina Hiseq2000 | 2 |
| Strasbourg | 2012 | Illumina Hiseq2000 | 13 |
| Genoscope | 2013 | Illumina Hiseq2000 | 13 |
| Genoscope | 2014 | Illumina Hiseq2000 | 31 |
| Genoscope | 2015 | Illumina Hiseq2000 | 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 :** 1) (**Figure :** 1 - **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 :** 1 - **B** et **C**).

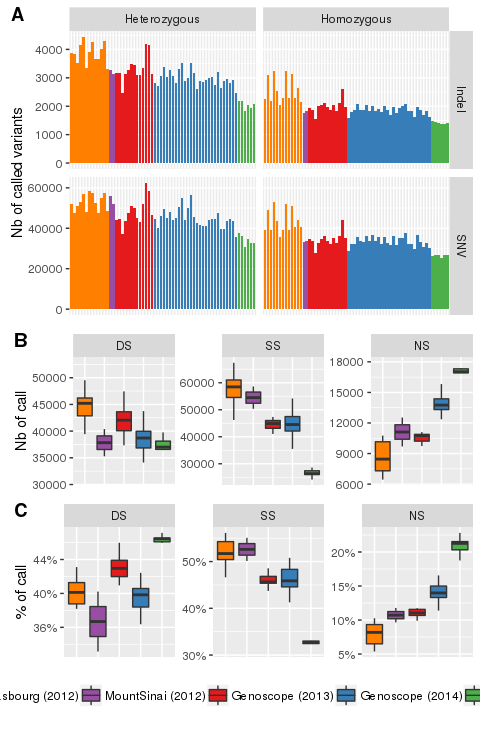


Figure 1 Résultats de l'appel des variants par individus et par projet de séquençage : Chaque couleur définit un projet de séquençage caractérisé par un centre de séquençage et une année. **A** : Quantification pour chaque individus (représentés par les barres) du nombre de variants (SNVs et Indels) 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"](#filterdescription) 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 NA et NA gènes impactés par au moins un variants homozygote (**Figure :** 2 - **A** et **B**).

Warning: Removed 3 rows containing non-finite values (stat\_boxplot).

Warning: Removed 3 rows containing missing values (geom\_point).

Warning: Removed 3 rows containing missing values (position\_stack).

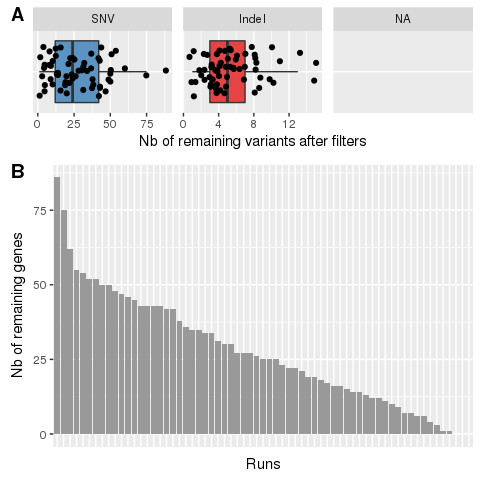


Figure 2 Résultats de l'étape de filtrage : **A** : Quantification du nombre de SNVs et indels ayant passé l'ensemble des filtres pour chaque patient. **B** : Nombre de gènes impactés par au moins un variant ayant passé les filtres pour chaque individu représenté par les barres. **C** : Présentation

Afin de 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 le cilliome humain (Ivliev, ’t Hoen, Roon-Mom, Peters, & Sergeeva, [2012](#ref-Ivliev2012)) en se basnat entre autre sur les gènes fournis par la base de données CilDB (Arnaiz, Cohen, Tassin, & Koll, [2015](#ref-Arnaiz2015)) mais aussi des analyses *in silico* permettant d'identifier des gènes jusqu'à présent jamais associés au ciliome. 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, 33 de nos 1432 gènes retenus faisaient partis de cette liste dont 22 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 :** 5 - **A**), or on connait désormais son implication dans le phénotype MMAF (**Table :** ??, **Figure :** 3 - **A**).

Parmi l'ensemble des gènes retrouvés mutés, on peut constater que 1212 d'entre eux, soit 84.6%, ne sont observés muté, à l'état homozygote, uniquement chez un seul de nos patients tandis que 220 sont retrouvés muté à l'état homozygote chez au moins 2 patients. Ainsi, analyser en priorité ces 220 gènes permet à la fois de réduire considérablement notre liste, et donc de faciliter les analyses, mais aussi de pouvoir identifier en priorité les gènes affectant le plus grand nombre de patient et donc probablement les principaux acteurs du impliqués dans le phénotype MMAF (**Figure :** 3 - **B**).

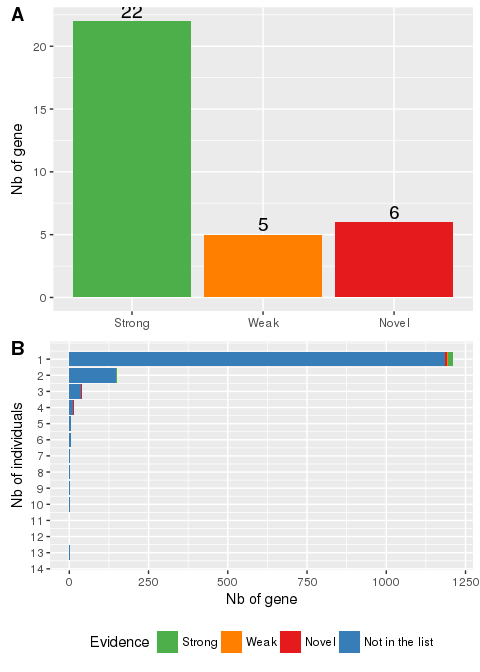


Figure 3 Répartition des gènes retrouvés mutés chez nos patients dans les différentes classes de la liste du cilliome : Chaque couleur définit une classe de la liste des gènes du cilliome décrit dans (Ivliev et al., [2012](#ref-Ivliev2012)). Vert = *Strong evidence from previous studies* (*Strong*), 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** : Quantification pour chacun des gènes du nombre d'individus présentant un variant homozygotes ayant passé l'ensemble des filtres

Nous avons ensuite *designer* 5 analyses fonctionnant de manière pyramidale. C'est à dire que la première analyse permet d'identifier les candidats les plus évidents tandis que la cinquième ceux pour lesquels il y a moins d'indices. L'interêt de ce fonctionnement est qu'à chacune des analyses, l'ensemble des variants portés par les patients pour lesquels un candidat a pu être déterminer avec **certitudes** seront retirés dans analyses suivantes. Ce procédé nous permettra ainsi après chacune de ces 5 analyses d'alléger notre liste de variants, et donc de gènes, rendant l'analyse suivante plus simple :

1. **Analyse n°1** : 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 malgré son expression non spécifique au testicule (**Figure :** 5 - **A**). C'est pourquoi nous avons recherché en priorité des patients portant des variants tronquants homozygotes sur ce gène.
2. **Analyse n°2** : Dans second temps, nous avons sélectionné 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**.
3. **Analyse n°3** : Ensuite, nous avons étudié les gènes **absents dans la liste cilliome** mais sur lesquels on trouvait toujours **au moins deux** de nos patients présentant au moins 1 variant tronquant à l'état **homozygote**.
4. **Analyse n°4** : 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**.
5. **Analyse n°5** : 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**.

Chacune de ces analyses s'est ensuite déroulée en 5 étapes :

1. **Étape n°1** : Cette première étape consistait à récupérer l'ensemble des patients répondant aux critères de l'analyse en question, par exemple : l'ensemble des patients portant un variant homozygote sur le gène *DNAH1* dans le cas de l'analyse n°1. Au vu de l'effet délétère des variants sélectionés au cours de cette étape, les preuves génétiques seront considérées comme de forte confiance (*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 effets **non tronquants** 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 régulièrement (Salgado, Bellgard, Desvignes, & B??roud, [2016](#ref-Salgado2016)) (**Figure :** ??). Au vus de la difficulté à déterminer l'effet délétère de ces variants, les preuves génétiques seront considérées comme de confiance modérée (*Moderate trust*).
3. **Étape n°3** : Cette étape consiste à rechercher 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. Malheureusement, dans le cadre des séquençages WES et 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 (**Figure :** 4). Pour cela, des analyses de biologie moléculaire sont nécessaires. C'est pour cette raison que ces preuves génétiques seront considérées comme étant de faible confiance (*Low trust*).

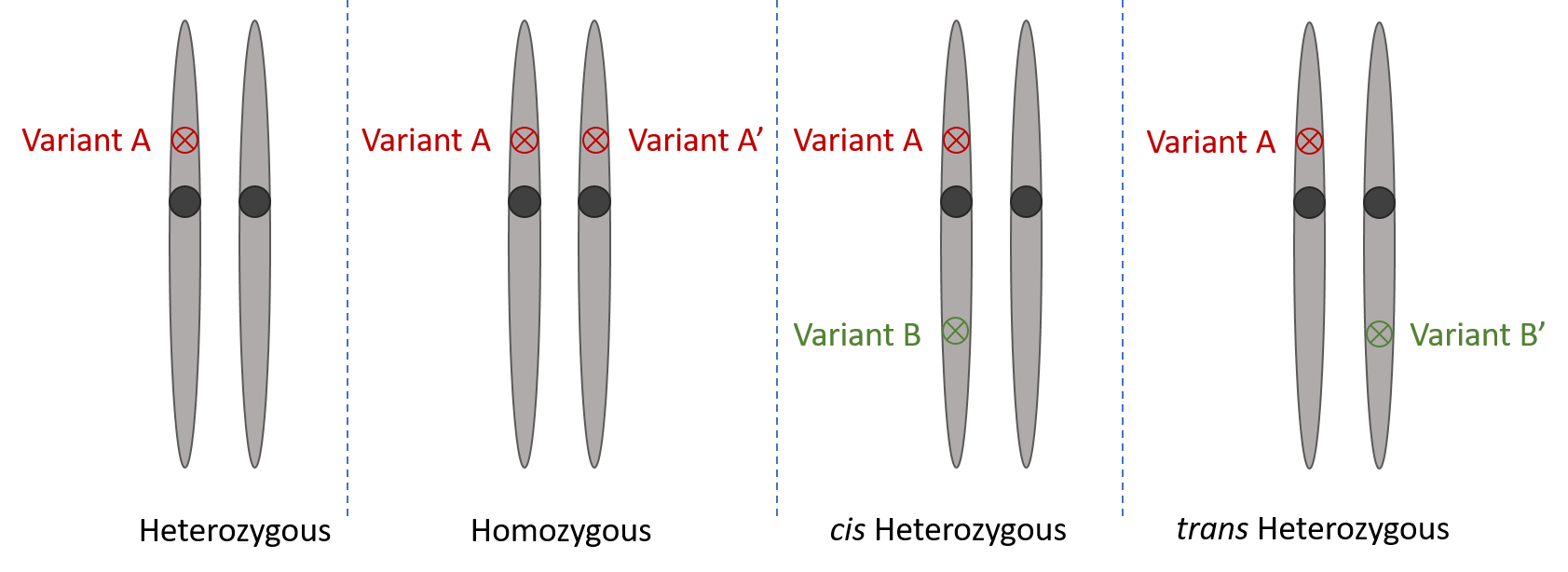


Figure 4 Représentation schématique des ....TODO : Un variant est dit homozygote lorsque le **même** variant est présents sur les deux allèles d'un gène et hétérozygote lorsqu'il est présent sur **un seul** des deux allèles. On parlera d'hétérozygotes *cis* lorsque deux variants hétérozygotes différents seront positionnés sur **le même** allèle et d'hétérozygote *trans* (ou composite) lorsque ces deux variants hétérozygotes seront positionnés sur **deux allèles différents**. En WES et en WGS il est impossible de différentié les hétérozygotes *cis* des hétérozygotes *trans*

1. **Étape n°4** : Au cours de cette étape, nous allons étudier chacun des gènes sélectionnés et déterminer au vus des différentes preuves génétiques, des données d'expressions de ces gènes mais aussi des informations disponibles dans la litterature, si ceux-ci sont de bons candidats pour expliquer le phénotype MMAF de nos patients. Il est à noter que **cette étape est la seul à ne pas être automatisée**.
2. **Étape n°5** : Pour finir, les données des patients pour lesquels un gène candidat a été identifié et dont les preuves génétiques sont considérées comme de forte confiance (c'est à dire les patients identifiés lors de l'étape n°1) sont retirés de notre liste de variants allégeant ainsi celle-ci pour les analyses suivante.

##### Analyse n°1 :

1. ***DNAH1*** : Parmi l'ensemble de nos 62 patients MMAF 1 portait un indel homozygote entrainant un décalage du cadre de lecture sur le gène *DNAH1* et répertorié dans aucune des 3 bases de données qua nous avons utilisées. Le patient et *Ghs90* était porteur de 3 variants homozygotes successifs entrainant tous un faux-sens tandis que le patient et *Ghs95* portait lui 1 autre variant homozygote entrainant un faux-sens différent des 3 portés par et *Ghs90*. On peut noter qu'aucun de ces 4 variants n'est répertorié dans les bases de données, cependant tous sont prédits comme *benign* par PolyPhen tandis que SIFT ne donne aucune prédiction. La recherche de patients potentiellement hétérozygotes composites nous a permis de révéler 6 patients portant tous 2 variants hétérozygotes sur le gène *DNAH1*. On peut alors noter le patients *Ghs36* porteur d'un premier variant créant un codon stop de manière prématuré et d'un second variant entrainant un faux-sens absent 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 candidats pour expliquer le phénotype du patient *Ghs36*. On peut noter également 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 données ExAC avec une très faible fréquence. Pour les 4 autres patients, il est plus difficile de se prononcer à ce stade compte tenus des fréquences parfois élevée des variants ou bien des prédictions fournies par PolyPhen.

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 protéine, seul le variant homozygote porté par le patient et *Ghs122* nous permets d'être certains de la causalité du phénotype MMAF. C'est pourquoi, seul les données de ce patient seront retirées de notre liste de variants réduisant celle-ci à 1661 variants chevauchant 1395 gènes (**Table :** ??). Pour les 8 autres patients, des analyses fonctionnelles complémentaires sont nécessaires.

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% observés dans notre étude précédente (Ben Khelifa et al., [2014](#ref-BenKhelifa2014)), 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.

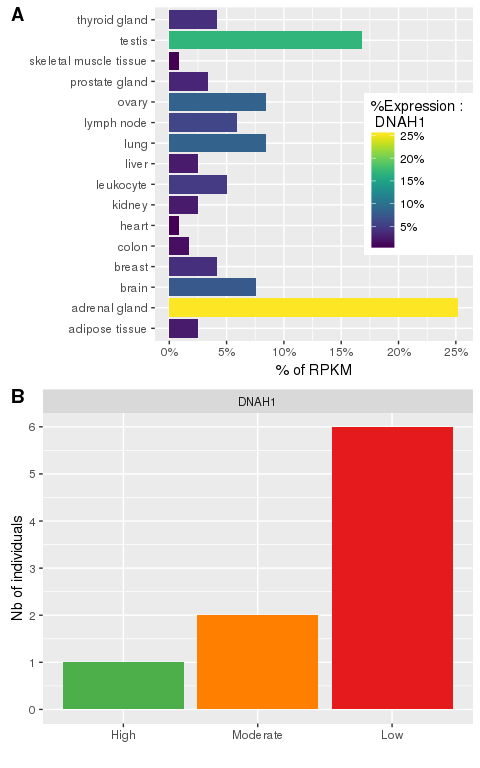


Figure 5 Analyse du gène *DNAH1* : Expression tissulaire du gène *DNAH1* d'après les données du projet Illumina BodyMap. Quantification du nombre de patients portant au moins un variant sur le gène *DNAH1* pour chacun des 3 niveau de confiance

##### Analyse n°2

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 :** ??). 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 :** 6 - **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 *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 tous 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 :** ?? et ??, **Figure :** 6 - **B**).
3. ***CCDC146*** : Sur ce gène, seuls les patients *Ghs32* et *Ghs35* sont retrouvés mutés. Cependant, ces deux patients portent tous deux des variants homozygotes induisant respectivement un variant un codon stop prématuré et un décalage du cadre de lecture. 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 données. On peut également ajouter que ce gène décrit comme faisant partie du cilliome humain avec de fortes preuves présente également 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 (E. N. Firat-Karalar, Sante, Elliott, & Stearns, [2014](#ref-Firat-Karalar2014)) renforçant 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 :** ?? et ??, **Figure :** 6 - **B**).
4. ***TTC29*** : Sur ce gène, les patients *Ghs19* et *Ghs26* portent la même variation retrouvée à très faible fréquence dans les trois bases 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 défauts de la cilliogénèse (Chung et al., [2014](#ref-Chung2014)) faisant de celui-ci un bon candidat pour expliquer le phénotype de ces 2 patients (**Table :** ?? et ??, **Figure :** 6 - **B**).

Cette première analyse au cours de laquelle nous avons sélectionné 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 4 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 restant dans notre cohorte. Pour les 2 autres, des analyses fonctionnelles sont nécessaires 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 leurs 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, elles, conservées afin de voir si un meilleur candidat pourrait expliquer le phénotype de ces patients.

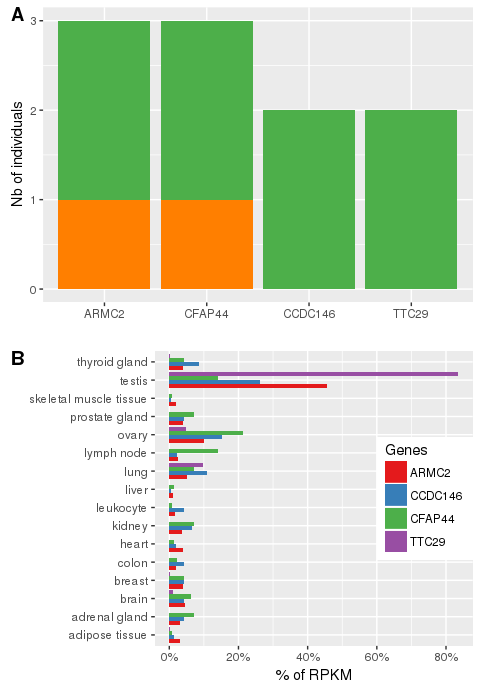


Figure 6 Analyse des gènes sélectionnés dans l'Analyse n°1 : Expression tissulaire des gènes retenus d'après les données du projet de transcriptome Illumina bodyMap. Résumé de l'Analyse 1, quantification du nombre de patients retrouvés mutés sur chacun des gènes retenus ainsi que du degré de confiance accordé à la cause génétique

##### Analyse n°3

Pour rappel, au cours de cette analyse, nous avons étudié les gènes **absents dans la liste cilliome** mais sur lesquels 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*. Parmi 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'épissage, 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 intéressant, puisque celui-ci porte deux variant hétérozygotes sur le gène *CFAP43* parmi 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 :** 7 - **B**), le nombre important de patients portant des variant sur celui-ci, notamment 5 portant 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* portent tous 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 également 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 sa forte 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](#ref-Brown2003)) font de ce gène un excellent candidat pour expliquer le phénotype d'au moins 2 patients. Pour les autres, des analyses fonctionnelles seront nécessaires.
3. ***SART3*** : Un total de 4 patients portait **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 protocole 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 donné le fait que ces patients ne sont pas apparentés et n'ont pas été séquencés 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](#ref-Mally1996), Stassi, Schloot, & Pietropaolo ([1997](#ref-Stassi1997))) et que celui-ci serait à la fois lié au diabète de type 1 (Martin et al., [1995](#ref-Martin1995), R. Gaedigk, Duncan, Miyazaki, Robinson, & Dosch ([1994](#ref-Gaedigk1994)), Martin, Lampasona, Dosch, & Pietropaolo ([1996](#ref-Martin1996))) et au syndrome de Gougerot-Sjögren (S. Winer et al., [2002](#ref-Winer2002)). 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 tous deux le même indel à l'état homozygote. Celui-ci entrainant un décalage du cadre de lecture induit la formation d'une protéine aberrente. L'absence de cette indel dans les trois bases de données laisse supposer 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* *Baz1a*^{-/-} présentaient de nombreux défauts tel que des spermatozoïdes non motiles, présentant une morphologie de la tête et du flagelle aberrante (Dowdle et al., [2013](#ref-Dowdle2013)) 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 répertorié dans les bases de données, causant un décalage du cadre de lecture. On peut également 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 partagent 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 porteurs respectifs de variant homozygote tronquant sur les gène *CCDC146* et *ARMC2* portaient également 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 d'ailleurs noter que le patient *Ghs41* était lui aussi porteur du variant sur *SART3* renforçant ainsi l'hypothèse de l'erreur de séquençage. Pour ces raisons, ce gène n'a pas été retenu en tant que candidat.

Dans cette analyse, 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 suffisamment 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 notamment 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 suffisamment fortes, l'ensemble de leurs variant furent ainsi retirés de notre liste contenant désormais 1093 variants et 962 gènes différents.

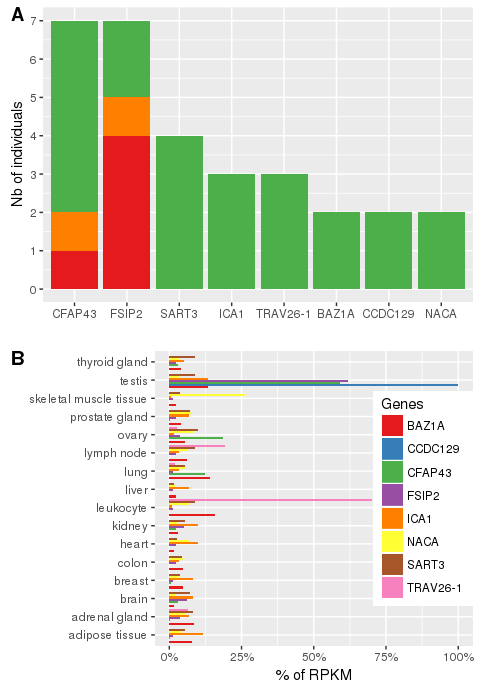


Figure 7 Analyse des gènes sélectionnés dans l'Analyse n°2 : Expression tissulaire des gènes retenus d'après les données du projet de transcriptome Illumina bodyMap. Résumé de l'Analyse 2, quantification du nombre de patients retrouvés mutés sur chacun des gènes retenus ainsi que du degré de confiance accordé à la cause génétique

##### Analyse n°4

Dans cette troisième analyse, nous avons sélectionné à nouveau les gènes **présents dans la liste cilliome** en conservant cette fois-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 appartenance au ciliome humain.

1. ***C6orf118*** : Ce gène a été retrouvé muté 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 également 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 difficile de se prononcer quant à l’effet délétère de ces deux variants sans effectuer d'analyses fonctionnelles. Il faut noter que *C6orf118* présente une forte expression dans le poumon. De plus, ce gène a récemment été décrit comme étant associé au phénotype de tuberculose pulmonaire (E. P. Hong, Go, Kim, & Park, [2017](#ref-Hong2017)). Cependant cela n'est en rien contradictoire avec le phénotype MMAF de ces 2 patients, le poumon comprenant de nombreuses cellules ciliées, notamment au niveau de l'épithélium respiratoire, il n'est donc pas surprenant que des gènes du flagelle aient également une fonction au sein d'autres organes ciliés.
2. ***C21orf59* et *CCDC65*** : Les patients *Ghs127* et *Ghs88* porte tout deux des indels induisant un décalage du cadre de lecture, le premier sur le gène *C21orf59*, le second sur *CCDC65* deux gènes déjà connu pour être impliqué dans la formations des cils. En effet, la protéine NYD-SP28 (ancien nom de CCDC65) avait déjà caractérisé comme faisant partie du flagelle spermatique (Zheng et al., [2006](#ref-Zheng2006)) alors que *C21orf59* contrôlerait à la fois la mobilité et la polarisation des cils chez *zebrafish* (Jaffe et al., [2016](#ref-Jaffe2016)). On note aussi que ces deux gènes ont été associés à des pathologies ciliaires chez le *zebrafish*, *Chlamydomonas* ainsi que chez l'humain, pouvant entrainer un phénotype de dyskinésie ciliaire primaire, de même, la protéine FBB18 du gène orthologue de C21orf59 régulerait la motilité du flagelle chez *Chlamydomonas* (Austin-Tse et al., [2013](#ref-Austin-Tse2013)). Ainsi, les arguments génétiques associés aux implications déjà avérées de ces 2 gènes dans la ciliogénèse font de *CCDC65* et *c21orf59* des excellents candidats pour expliquer les phénotype MMAF de nos 2 patients malgré l'expression non spécifique au testicule du gène *C21orf59*.
3. ***SPEF2*** : Ce gène est retrouvé muté à l'état homozygote chez le patient et *Ghs131* qui porte un indel créant un décalage du cadre de lecture. Malgré son expression non spécifique au testicule, plusieurs études ont démontré le rôle important de la protéine SPEF2 chez la souris dans la formation et la structure du flagelle spermatique de même que l'absence de cette protéine entrainait la formation d'un flagelle court et désorganisé (Sironen et al., [2010](#ref-Sironen2010), Sironen et al. ([2011](#ref-Sironen2011))) faisant ainsi de ce gène un excellent candidat pour le phénotype MMAF du patient et *Ghs131*.

Analyser les gènes de la liste ciliome sur lesquels **un seul** patient portait un variant tronquant à l'état homozygote nous a permis d'identifier 4 nouveaux gènes : *C21orf59*, *C6orf118*, *CCDC65* et *SPEF2*. Ainsi, 4 patients portaient un variant homozygote tronquant sur l'un d'eux tandis qu'un autre patient portait deux variant faux-sens hétérozygotes sur le gène *C6orf118*. Parmi ces 4, seuls *CCDC65* et *C6orf118* présentaient une forte expression testiculaire, cependant tous avaient déjà été associés à la formation du cil chez l'humain ou chez une autre espèce animale. Bien que des analyses supplémentaires soient nécessaires afin de pouvoir affirmer que les deux variants hétérozygotes portés par le patients et *Ghs27* sont bien responsables de son phénotype, les arguments génétiques ainsi que les informations fournies par la littérature nous ont permis d'affirmer que ces gènes sont responsables des phénotypes des patients *Ghs127*, *Ghs131*, *Ghs40* et *Ghs88*. Les données de variants de ces derniers ont ainsi été retirés réduisant à nouveau notre liste à 934 variants et 821 gènes différents.

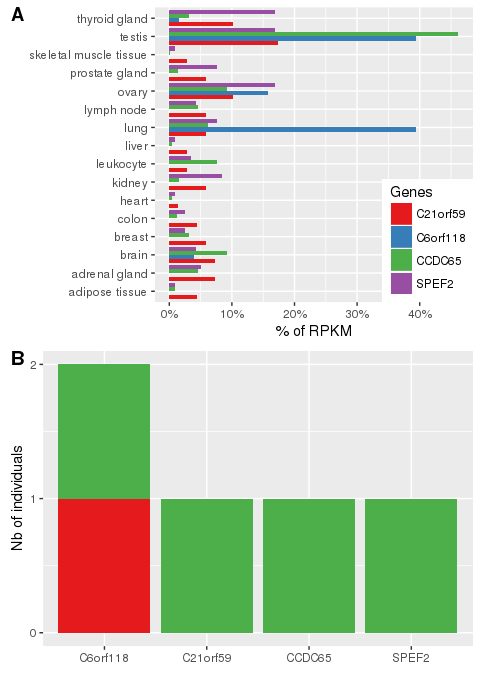


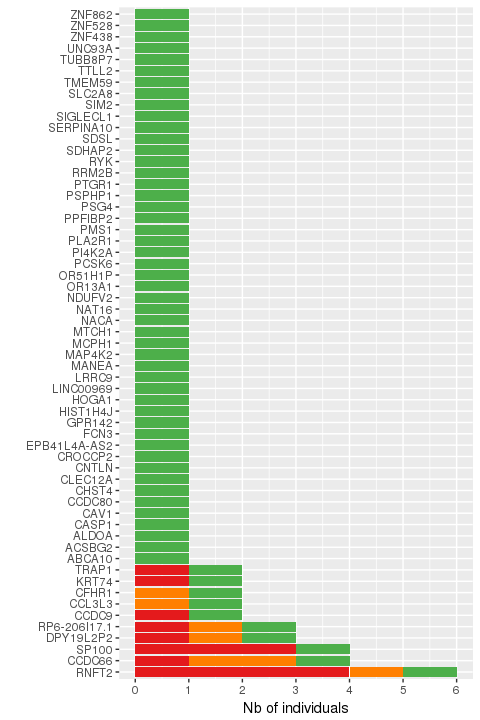
Figure 8 Analyse des gènes sélectionnés dans l'Analyse n°4 : **A** : Expression tissulaire des gènes retenus dans cette analyse d'après les données du projet de transcriptome Illumina bodyMap. **B** : Résumé de l'Analyse 3, quantification du nombre de patients retrouvés mutés sur chacun des gènes retenus ainsi que du degré de confiance accordé à la cause génétique

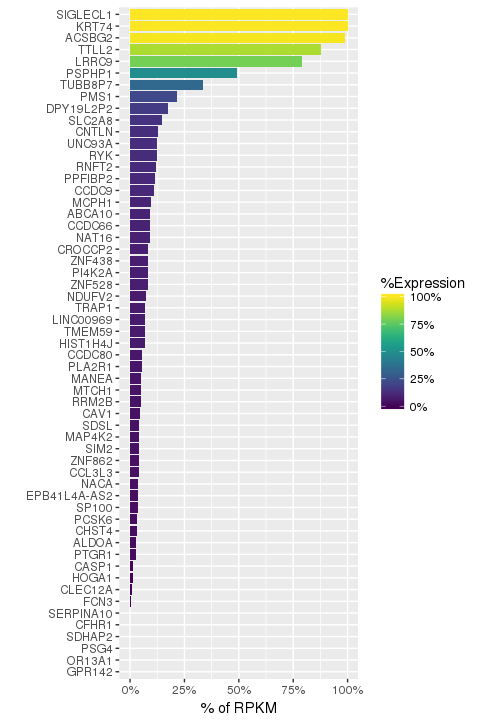
##### Analyse n°5

Pour finir nous avons dans cette analyse sélectionné 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 59 gènes différents retrouvés mutés chez 30 de nos patients.

En raison de la grande quantité de gène, nous nous sommes tout d'abord concentré sur ceux retrouvés mutés à l'état homozygote chez au moins deux patients. Ainsi nous avons obtenus une liste de 6 gènes : *CCDC66*, *CCL3L3*, *CFHR1*, *DPY19L2P2*, *RNFT2* et *RP6-206I17.1*

1. ***CCDC66*** : Le gène *CCDC66* a été retrouvé muté à l'état homozygote chez ... patients parmi lesquels ... portait un variant tronquant entrainant un décalage du cadre de lecture et retrouvé dans aucune des bases de données. Malgré une expression testiculaire faible, l'implication récente de ce gène dans la ciliogénèse (Conkar et al., [2017](#ref-Conkar2017)) fait de ce gène un bon candidat bien que certaines études aient déjà démontré l'implication de celui-ci dans des pathologies rétiniennes à la fois chez le chien (Dekomien et al., [2010](#ref-Dekomien2010)) mais aussi chez l'humain (Gerding et al., [2011](#ref-Gerding2011), Khan et al. ([2017](#ref-Khan2017))). En effet, ces anomalies rétiniennes.





#### 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 9 patients présentant des variants sur ce gène dont 3 à l'état homozygote (dont 1 tronquant). Elle nous a également permis d'identifier 12 nouveaux gène candidats : *ARMC2*, *BAZ1A*, *C21orf59*, *C6orf118*, *CCDC129*, *CCDC146*, *CCDC65*, *CFAP43*, *CFAP44*, *FSIP2*, *SPEF2* et *TTC29*. Ainsi, 27 de nos patients soit 43.5% de la cohorte présentaient au moins 1 variant homozygote sur l'un de ces gènes, et, chez 23 d'entre eux, ce variant induisaient un effet tronquant sur la séquence protéique. Les 4 autres portaint tous au moins deux variants hétérozygotes, tronquants ou non, sur un de ces gènes (**Figure :** 9 - **A**).

Parmi cet ensemble de patients, il faut noter que 5 d'entre eux porte des variants sur plusieurs des gènes candidats que nous avons identifiés (**Table :** 2, **Figure :** 9 - **B**).

Cependant, parmi ces différents variants certains semblent plus probables pour expliquer le phénotype :

1. **Patient *Ghs131* :** Ce patient porte à la fois un variant homozygote sur le gène *FSIP2* et un autre sur le gène *SPEF2*. Cependant, au vu de l'effet tronquant de celui présent sur le gène *SPEF2*, il est plus probable que ce soit celui-ci qui soit responsable du phénotype de ce patient.
2. **Patient *Ghs132* :** Ce patient porte à la fois un variant homozygote causant un décalage du cadre de lecture sur le gène *CCDC129* et deux variants hétérozygotes sur le gène *CFAP43*. Au vu de l'incertitude lié au phasage des variants hétérozygote et de l'impact délétère de celui présent sur le gène *CCDC129*, c'est ce dernier qui sera retenu pour ce patient dans un premier temps.
3. **Patient *Ghs40* :** Ce patient est porteur de 3 variants faux-sens hétérozygotes sur le gène *FSIP2* et d'un variant créant un codon stop prématuré homozygote sur *C6orf118*. En tenant compte du génotype et de l'efet délétère du variant, il est plus probable que celui chevauchant le gène *C6orf118* soit responsable du phénotype.
4. **Patient *Ghs88* :** Comme le patient*Ghs40*, celui-ci porte à la fois deux variants hétérozygotes dur le gène *DNAH1* et un variant tronquant sur *C21orf59*. Pour les même raison que celles précédement évoquées, ces ce dernier qui sera retenu.
5. **Patient *Ghs95* :** Pour les même raison, c'est pour ce patient le faux-sens homozygote présant sur le gène *DNAH1* qui sera ici conservé.

Ainsi, cette analyse révèle l'efficacité de notre pipeline puisqu'elle a permis d'identifier un gène candidat pour 38 des 62 patients de notre cohorte, soit 61.3% de nos patients. Pour les 24 patients pour lesquels aucun candidat n'a été identifié, des analyses individuelles complémentaires sont nécessaires.

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**](#famdnah1) : 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

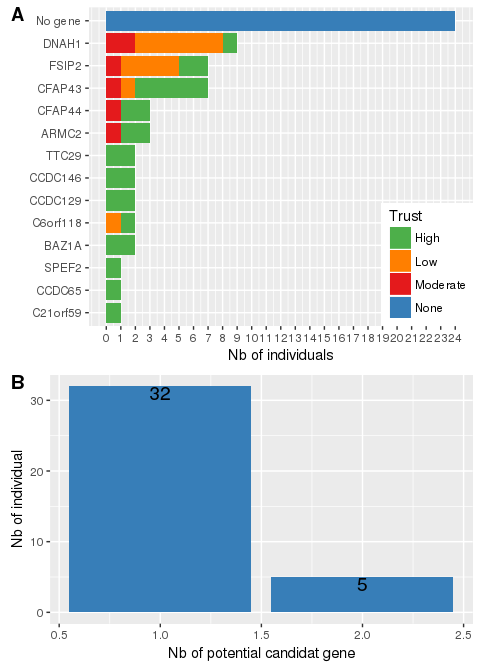


Figure 9 Conclusion des analyses WES de notre large cohorte MMAF, liste des gènes candidats : **A** : Quantification du nombre de patient portant un ou plusieurs variants sur un des gène candidat. La couleur des barres dépend du type et du génotype du ou des variants portés par chaque patient. La barre bleue indique les patients pour lesquels aucun candidat n'a été identifié. **B** : Nombre de candidat potentiel par patient (parmi ceux pour lesquels au moins un gène candidat a été identifié)

Table 2 Liste des patients portant au moins 1 variant sur deux gènes candidats

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Patient | Gene | HGVSc, HGVSp | Consequence | Genotype |
| Ghs131 | FSIP2 | . ; p.Ala86Val | missense | Homozygous |
| Ghs131 | SPEF2 | c.3240delG ; p.Phe1080LeufsTer2 | frameshift | Homozygous |
| Ghs132 | CCDC129 | . ; . | frameshift | Homozygous |
| Ghs132 | CFAP43 | c.1300\_1301insT ; p.Leu435SerfsTer26 | frameshift | Heterozygous |
| Ghs132 | CFAP43 | c.1040G>C ; p.Val347Ala | missense | Heterozygous |
| Ghs40 | C6orf118 | c.1114C>T ; p.Arg372Ter | stop gained | Homozygous |
| Ghs40 | FSIP2 | . ; p.Asn495Ile | missense | Heterozygous |
| Ghs40 | FSIP2 | . ; p.Ile2960Met | missense | Heterozygous |
| Ghs40 | FSIP2 | . ; p.Lys5822Ile | missense | Heterozygous |
| Ghs88 | C21orf59 | c.79delA ; p.Phe28LeufsTer47 | frameshift | Homozygous |
| Ghs88 | DNAH1 | c.2209C>A ; p.Val737Met | missense | Heterozygous |
| Ghs88 | DNAH1 | c.3877T>A ; p.Asp1293Asn | missense | Heterozygous |
| Ghs95 | DNAH1 | c.9278C>G ; p.Ala3093Gly | missense | Homozygous |
| Ghs95 | FSIP2 | . ; p.Asn495Ile | missense | Heterozygous |
| Ghs95 | FSIP2 | . ; p.Lys5822Ile | missense | Heterozygous |

## 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.

# Tables des analyses d'exomes

# References

Amberger, J., Bocchini, C., & Hamosh, A. (2011). A new face and new challenges for Online Mendelian Inheritance in Man (OMIM). *Human Mutation*, *32*(5), 564–567. <http://doi.org/10.1002/humu.21466>

Arnaiz, O., Cohen, J., Tassin, A., & Koll, F. (2015). Remodeling Cildb, a popular database for cilia and links for ciliopathies. *Cilia*, *4*, P21. <http://doi.org/10.1186/2046-2530-4-S1-P21>

Austin-Tse, C., Halbritter, J., Zariwala, M. A., Gilberti, R. M., Gee, H. Y., Hellman, N., … Hildebrandt, F. (2013). Zebrafish Ciliopathy Screen Plus Human Mutational Analysis Identifies C21orf59 and CCDC65 Defects as Causing Primary Ciliary Dyskinesia. *American Journal of Human Genetics*, *93*(4), 672–86. <http://doi.org/10.1016/j.ajhg.2013.08.015>

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>

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>

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. <http://doi.org/10.4161/fly.19695>

Conkar, D., Culfa, E., Odabasi, E., Rauniyar, N., Yates, J. R., & Firat-Karalar, E. N. (2017). The centriolar satellite protein CCDC66 interacts with CEP290 and functions in cilium formation and trafficking. *Journal of Cell Science*, *130*(8). Retrieved from <http://jcs.biologists.org/content/130/8/1450.long>

Dekomien, G., Vollrath, C., Petrasch-Parwez, E., Boev?, M. H., Akkad, D. A., Gerding, W. M., & Epplen, J. T. (2010). Progressive retinal atrophy in Schapendoes dogs: mutation of the newly identified CCDC66 gene. *Neurogenetics*, *11*(2), 163–174. <http://doi.org/10.1007/s10048-009-0223-z>

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>

Gerding, W. M., Schreiber, S., Schulte-Middelmann, T., de Castro Marques, A., Atorf, J., Akkad, D. A., … Petrasch-Parwez, E. (2011). Ccdc66 null mutation causes retinal degeneration and dysfunction. *Human Molecular Genetics*, *20*(18), 3620–3631. <http://doi.org/10.1093/hmg/ddr282>

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>

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>

Jaffe, K. M., Grimes, D. T., Schottenfeld-Roames, J., Werner, M. E., Ku, T.-S. J., Kim, S. K., … Burdine, R. D. (2016). c21orf59/kurly Controls Both Cilia Motility and Polarization. *Cell Reports*, *14*(8), 1841–9. <http://doi.org/10.1016/j.celrep.2016.01.069>

Khan, A. O., Budde, B. S., Nürnberg, P., Kawalia, A., Lenzner, S., & Bolz, H. J. (2017). Genome-wide linkage and sequence analysis challenge CCDC66 as a human retinal dystrophy candidate gene and support a distinct NMNAT1-related fundus phenotype. *Clinical Genetics*. <http://doi.org/10.1111/cge.13022>

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>

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., … Bamshad, M. J. (n.d.). Exome sequencing identifies the cause of a Mendelian disorder. <http://doi.org/10.1038/ng.499>

Pelak, K., Shianna, K. V., Ge, D., Maia, J. M., Zhu, M., Smith, J. P., … Goldstein, D. B. (2010). The characterization of twenty sequenced human genomes. *PLoS Genetics*, *6*(9), e1001111. <http://doi.org/10.1371/journal.pgen.1001111>

Robinson, P. N., Köhler, S., Oellrich, A., Sanger Mouse Genetics Project, S. M. G., Wang, K., Mungall, C. J., … Smedley, D. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research*, *24*(2), 340–8. <http://doi.org/10.1101/gr.160325.113>

Salgado, D., Bellgard, M. I., Desvignes, J. P., & B??roud, C. (2016). How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era. *Human Mutation*, *37*(12), 1272–1282. <http://doi.org/10.1002/humu.23110>

Sironen, A., Hansen, J., Thomsen, B., Andersson, M., Vilkki, J., Toppari, J., & Kotaja, N. (2010). Expression of SPEF2 During Mouse Spermatogenesis and Identification of IFT20 as an Interacting Protein1. *Biology of Reproduction*, *82*(3), 580–590. <http://doi.org/10.1095/biolreprod.108.074971>

Sironen, A., Kotaja, N., Mulhern, H., Wyatt, T. A., Sisson, J. H., Pavlik, J. A., … Lee, L. (2011). Loss of SPEF2 function in mice results in spermatogenesis defects and primary ciliary dyskinesia. *Biology of Reproduction*, *85*(4), 690–701. <http://doi.org/10.1095/biolreprod.111.091132>

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>

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. <http://doi.org/10.1093/nar/gkq603>

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>