

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

Remerciements	1
Résumé	3
Abstract	5
Chapitre 1 : Introduction	7
1.1 La spermatogénèse	7
1.1.1 Rappels sur le testicule	8
1.1.2 La phase de multiplication	9
1.1.3 La méiose	10
1.1.4 La spermogénèse	15
1.2 Structure et fonction du spermatozoïde	18
1.2.1 Anatomie du spermatozoïde	18
La tête	18
Le flagelle	20
1.2.2 Fonction du spermatozoïde	22
1.3 L'infertilité masculine	23
1.3.1 Les différents phénotypes d'infertilité masculine	23
Anomalies liées à la quantité spermatique	23
Anomalies liées liée à la morphologie	24
Anomalies liées à la mobilité	26
1.3.2 La génétique de l'infertilité	26
Les causes fréquentes	26
Les nouveaux gènes	28
1.4 Les techniques d'analyses génétiques	31
1.4.1 Approche “gènes candidats”	31
1.4.2 Les puces	31
Les puces à SNP, le génotypage... (titre à revoir)	32
Les puces à indels	32
Les puces à expression	32
Limitation	32
1.4.3 Le séquençage NGS	33
La capture des parties à séquencer, avantages et inconvénients	34
L'amplification	35

La réaction de séquence	37
1.5 L'analyse bioinformatique des données de NGS	41
1.5.1 Les données fournies par le NGS	41
Un <i>read</i> c'est quoi ?	41
Le format FASTQ	42
1.5.2 L'alignement	42
1.5.3 L'appel des variants	43
1.5.4 L'annotation des variants, filtrage et priorisation	45
1.5.5 Conclusion NGS	48
Chapitre 2 : Investigation génétique et physiologique de la globozoospermie	51
2.1 Introduction sur la globozoospermie	51
2.2 Résultats	54
2.2.1 Les mécanismes mutationnels	54
Confirmation de l'excès de délétion	54
Détermination du taux de recombinaison <i>de novo</i> de <i>DPY19L2</i>	56
Autres résultats	57
2.2.2 La transcriptomique	58
2.3 Conclusion	62
Chapitre 3 : Mise en place d'une stratégie pour l'analyse des données exomiques – application en recherche clinique	63
3.1 Intro	63
3.2 Résultats	64
3.2.1 Description de la pipeline	64
3.2.2 Utilisation du pipeline dans des cas familiaux	67
Description des familles	67
Résultats des exomes	68
Discussion	82
3.2.3 Etude d'une large cohorte de patients MMAF	84
Description de la cohorte	84
Application de la pipeline - Résultats	85
Analyse des listes de gènes	88
Chapitre 4 : MutaScript	89
4.1 Introduction	89
4.2 Matériel & Méthodes	91
4.2.1 Récupération et filtrage des données	91
4.2.2 Validation du score	92
4.2.3 Les figures	92
4.3 Résultats	92
4.3.1 Résultat de l'annotation	92
4.3.2 Détermination de la formule du score	92
Le SLAC et le WSLAC	92

4.3.3	Analyse du score	93
4.4	Comparaison avec RVIS et pLI	94
4.5	Conclusion	94
Conclusion	95
Annexe A : The First Appendix	97
Annexe B : The Second Appendix, for Fun	99
References	101

Liste des tableaux

1.1	Durée de vie moyenne des cellules germinales humaines	7
3.1	Liste simplifiée des conséquences prédites par VEP avec leur description et impact associée	66
3.2	Tableau récapitulatif des familles séquencées et de leur phénotype . .	67
3.3	Liste des gènes ayant passé l'ensemble des filtres pour chaque famille	81
3.4	Liste des différents projets de séquençages effectués	84
3.5	liste des variants passant les filtres et chevauchant le gène *DNAH1*	87

Table des figures

1.1	Schéma anatomique du testicule humain	8
1.2	Les différentes phases de la spermatogénèse d'après [medizin-kompakt](http://www.medizin-kompakt.de/spermatogenese) : description à écrire!!!	10
1.3	Les différentes étapes de la méiose gamétique masculine	11
1.4	Les différentes étapes de la première division méiotique masculine adapté	13
1.5	Les différentes étapes de la deuxième division méiotique masculine adapté	14
1.6	Schéma simplifié d'un enjambement chromosomique (crossing-over) .	15
1.7	Principales étapes et modifications structurales lors de la spermiogénèse	17
1.8	Anatomie simplifiée du spermatozoïde	18
1.9	Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes	20
1.10	Structure simplifiée de l'axonème	21
1.11	Structure du flagelle d'un spermatozoïde	22
1.12	Différentes anomalies morphologiques du spermatozoïde selon la classification de David modifiée adapté... TABLEAU à adapter et à insérer!!!!	25
1.13	Représentation schématique du chromosome Y adapté	27
1.14	Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée	34
1.15	Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS	37
1.16	Exemple de séquençage CRT tel qu'il est effectué par Illumina	38
1.17	Exemple de séquençage SNA tel qu'il est effectué par Ion Torrent .	39
1.18	Exemple de séquençage SBL tel qu'il est effectué par SOLiD	40
1.19	présentation d'un fichier FASTQ (FIGURE A CHANGER)	42
1.20	Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel	44
1.21	Diagramme de Venn des prédictions de pathogénicités de six logiciels	47
2.1	Observation au microscope à balayage d'un spermatozoïde normal (**A**) et d'un spermatozoïde globozoocéphale (**B**) (changer les photos avec celles sur lesquelles on voit l'acrosome colorés)	52
2.2	Comparaison entre les spermatozoïdes des souris *Dpy19l2* ^{-/-} (à gauche) et les souris sauvages *Dpy19l2* ^{+/+} (à droite)	53

2.3	Représentation schématique du mécanisme de NAHR causé par les séquences LCR flanquant le gène *DPY19L2*	55
2.4	Ratio des délétions / duplications *de novo* observées au locus *DPY19L2* déterminé par PCR digital à partir d'ADN spermatique de trois donneurs	57
2.5	Fonctions moléculaires affectées par au moins un gène dérégulé	61
2.6	Détails des sites ciblés par les gènes dérégulés ayant la fonction moléculaire "Binding"	61
3.1	Listes des différentes conséquences prédites par VEP et leur positionnement sur le transcrit	65
3.2	Processus simplifié du contrôle qualité des *reads*	69
3.3	Contrôle qualité des variants appelés	72
3.4	Annotation des variants par VEP	74
3.5	Filtrage des transcrits jugés "non pertinents" et des variants les chevauchant	76
3.6	Nombre d'individus composant la cohorte contrôle de chaque famille	77
3.7	Comparaison de l'efficacité de chacun des six filtres utilisés	79
3.8	Nombre de gènes passant l'ensemble des filtres par famille	83
3.9	Résultats de l'appel des variants par individus et par projet de séquençage	86
3.10	TODoooooooooooooooo	88

Remerciements

Je remercie ...

->

Résumé

Résumé de ma thèse

Second paragraph of abstract starts here.

Abstract

Même chose en anglais

Chapitre 1

Introduction

1.1 La spermatogénèse

La spermatogenèse des mammifères est un processus long et complexe contrôlé par plusieurs mécanismes étroitement liés (Gnessi, Fabbri, & Spera, 1997, KIERSZENBAUM (1994)), **Sharpe1994 à trouver !!!**. C'est au cours de celle-ci qu'à partir de cellules germinales, seront produits les spermatozoïdes matures. Ce processus est divisé en trois phases principales : La phase de multiplication, la phase de division (appelée la méiose) et la phase de maturation. Chez les hommes, ces étapes se déroulent en continu dans la paroi des tubes séminifères du testicule depuis la puberté jusqu'à la mort et implique trois types de cellules germinales : les spermatogonies, les spermatocytes et les spermatides. Le temps nécessaire pour obtenir un spermatozoïde mature à partir de cellules germinales est de 74 jours et la production quotidienne de spermatozoïde s'élève environ à 45 million par testicules (JOHNSON, PETTY, & NEAVES, 1980). Le cycle spermatogénétique est défini comme la succession chronologique des différents stades de différenciation d'une génération de cellules germinales (depuis la spermatogonie jusqu'au spermatozoïde). Chacune des étapes du cycle spermatogénétique a une durée fixe et constante selon les espèces (**Table : 1.1**).

Table 1.1 – Durée de vie moyenne des cellules germinales humaines

Cellules germinales	Durée de vie moyenne (jours)
Spermatogonies Ap	16-18
Spermatogonie B	7.5-9
Spermatocytes primaires	23
Spermatocytes secondaires	1
Spermatides	1

1.1.1 Rappels sur le testicule

Les testicules sont les organes sexuels masculins. Ils possèdent deux fonctions principales (plus ou moins exprimées selon les périodes de la vie de l'individu) : une fonction endocrine caractérisée par la synthèse des hormones stéroïdes sexuelles masculines (la stéroïdogenèse) et une fonction exocrine au cours de laquelle seront produits les gamètes masculins. Chez un individu adulte en bonne santé, le testicule présente une forme ovoïde ayant un volume moyen de 18 cm³. Chez l'homme, comme chez la plupart des mammifères terrestres, ils sont localisés sous le pénis dans une poche de peau appelée scrotum et reliés à l'abdomen par le cordon spermatique (**Figure : 1.1**). Cette externalisation des testicules permet leur maintien à une température plus basse que celle du reste du corps nécessaire à la spermatogenèse.

L'intérieur du testicule contient des tubes séminifères enroulés ainsi que du tissu entre les tubules appelé espace interstitiel. Les tubes séminifères sont de longs tubes compactés sous forme de boucles et dont les deux extrémités débouchent sur le *rete testis* (**Figure : 1.1**). C'est le long des parois du tube séminifère que se déroulera l'ensemble des étapes de la spermatogenèse.

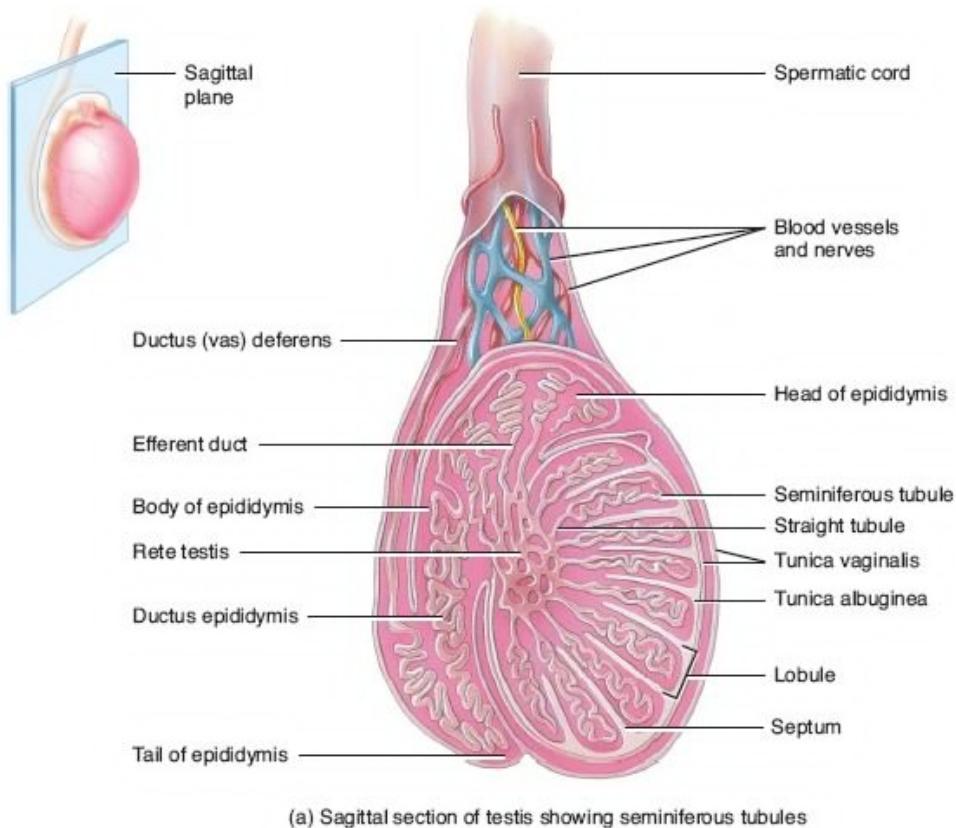


Figure 1.1 – Schéma anatomique du testicule humain

1.1.2 La phase de multiplication

La phase de multiplication est la phase au cours de laquelle les spermatogonies se divisent par mitoses pour aboutir au stade de spermatocytes primaires. Les spermatogonies sont des cellules diploïdes à l'origine de l'ensemble des autres cellules germinales humaines. Pour cela, elles vont s'auto-renouveler par mitoses successives afin de maintenir une production continue de spermatozoïdes tout au long de la vie de l'individu. Ces cellules sont localisées dans le compartiment basal des tubes séminifères. Les analyses histologiques ont permis de distinguer trois types de spermatogonies en fonction de leur contenu en hétérochromatine (Clermont, 1963, Clermont (1966), Goossens & Tournaye (2013)) :

1. Les spermatogonies de type A dark (ou Ad)
2. Les spermatogonies de type A pale (ou Ap)
3. Les spermatogonies de type B

Chez l'Homme, les spermatogonies Ad ont une activité mitotique au cours de la spermatogénèse et servent de réserve. Elles vont au cours d'une première mitose former une spermatogonie Ad et un spermatogonie Ap (**Figure : 1.2**). Cette propriété permet à la fois de se différencier en spermatocytes tout en constituant un compartiment de réserve de spermatogonies Ad pour la régénération de la population de cellules germinales au sein de l'épithélium séminifère. L'entrée en division des spermatogonies Ap se fait par groupes cellulaires tous les 16 jours. Les cellules d'une même génération maintiennent entre elles des ponts cytoplasmiques jusqu'à la spermiogénèse ce qui permet la synchronisation parfaite du développement gamétique de toutes les cellules filles issues d'un groupe de spermatogonies Ap. Ce phénomène est appelé onde spermatogénétique. Chaque spermatogonie Ap va, lorsqu'elle se divise par mitose, former deux spermatogonies B qui elles-mêmes se diviseront en deux spermatocytes primaires diploïdes (**Figure : 1.2**).

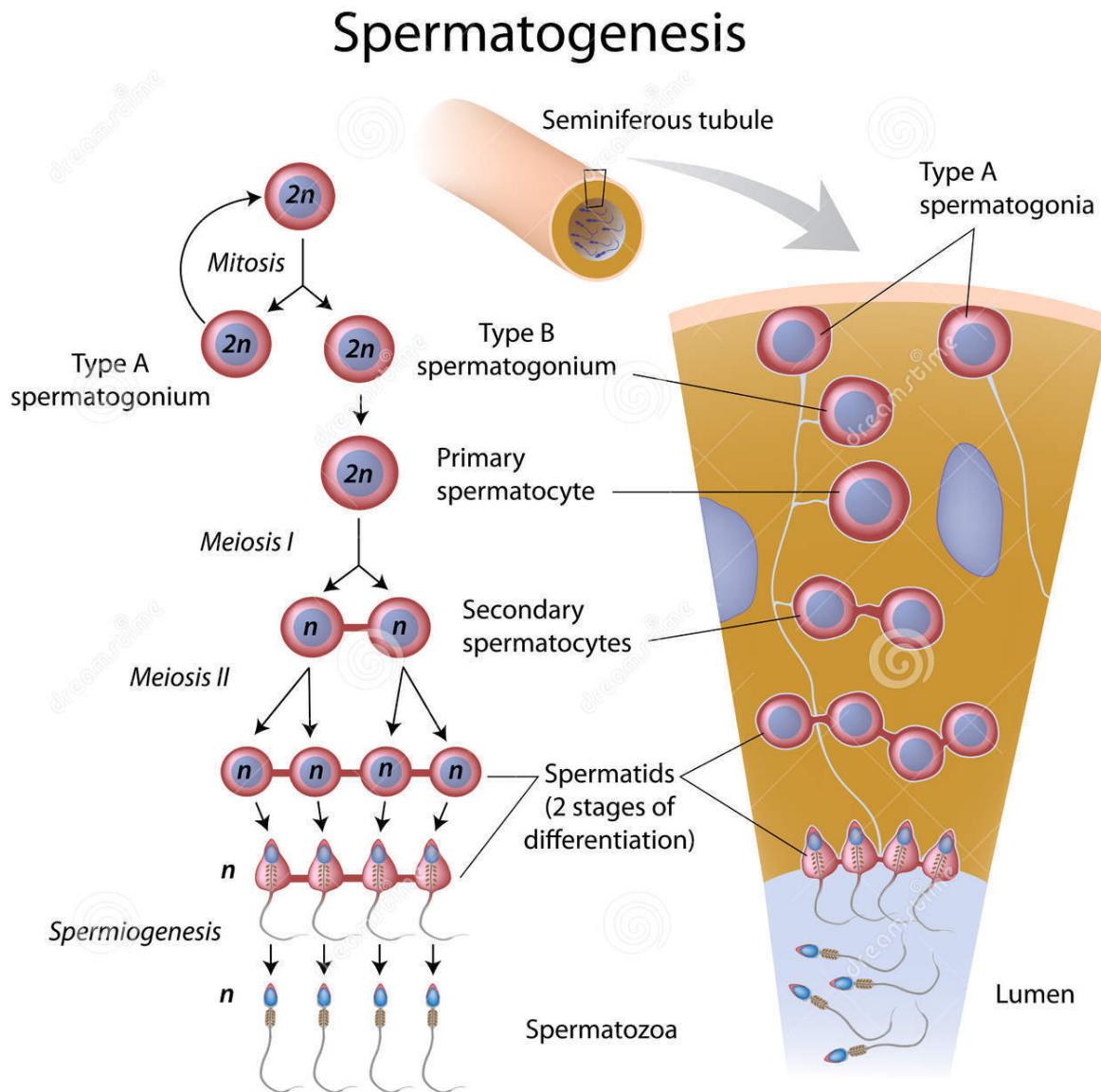


Figure 1.2 – Les différentes phases de la spermatogénèse d'après [medizin-kompakt](<http://www.medizin-kompakt.de/spermatogenese>) : description à écrire!!!

1.1.3 La méiose

La méiose, ou phase de maturation, est l'étape au cours de laquelle, à partir de cellules diploïdes (les spermatogonies B) vont se former des cellules haploïdes, les spermatocytes secondaires (spermatocytes II). Ce résultat est le fruit de deux divisions successives (Figure : 1.3) appelée respectivement méiose réductionnelle ou méiose I (MI) et méiose équationnelle ou méiose II (MII). La MI va séparer les chromosomes homologues, produisant deux cellules et réduisant la ploïdie de diploïde à haploïde

(d'où son non *réductionnelle*). En plus de son rôle de division vu précédemment, la méiose joue un rôle clef dans le brassage génétique (mélange des gènes) et ce, grâce à deux mécanismes de brassage : le brassage inter-chromosomique, lorsque les chromosomes sont séparés et le brassage intra-chromosomique impliquant notamment des enjambements chromosomiques (crossing-over) (**Figure : 1.6**).

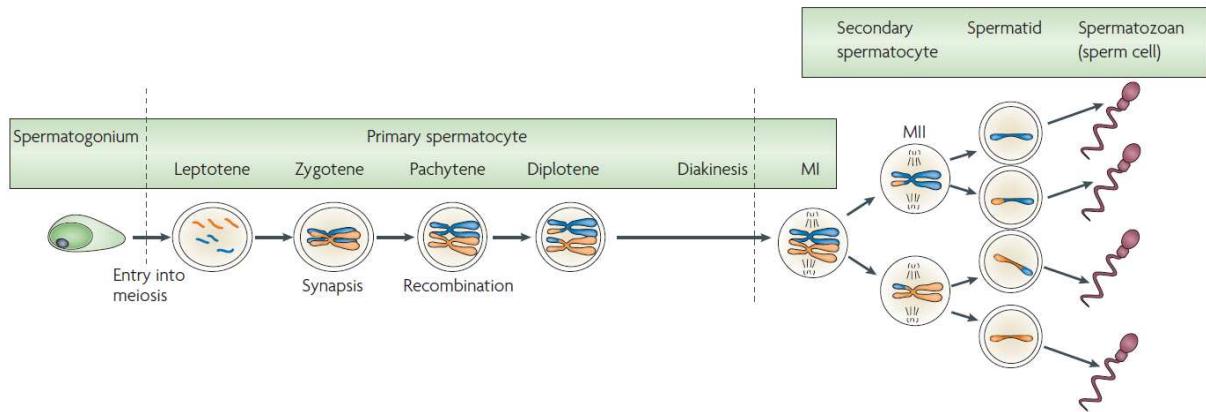


Figure 1.3 – Les différentes étapes de la méiose gamétique masculine d'après Sasaki et Matsui, 2008

La méiose est initiée dès la fin de la phase de multiplication à partir des spermatocytes primaires issus de la division des spermatogonies de type B. Ces cellules nouvellement formées se situent dans le compartiment basal du tube séminifère. C'est là qu'ils vont tout d'abord subir une interphase (stade préleptotène) durant entre 2 et 4 jours. Au cours de cette phase a lieu la réPLICATION de l'ADN. Cette réPLICATION se fait lorsque l'ADN est à l'état de chromatine, pendant la phase S (pour synthèse) de l'interphase. À l'issue de cette phase, chaque chromosome sera composé de deux chromatides reliées entre elles par le centromère, le matériel génétique de chaque cellule ayant donc été multiplié par 2. Par la suite, ces cellules vont subir deux divisions méiotiques, chacune composées de 4 étapes distinctes (**Figure : 1.3**) :

1. Méiose réductionnelle : (**Figure : 1.4**)

- a. **La prophase I :** Cette longue étape dure 23 jours chez l'homme et peut être subdivisée en 5 phases successives : leptotène, zygotène, pachytène, diplotène et diacénèse.
 - i. **Leptotène :** condensation de la chromatine et formation des chromosomes.
 - ii. **Zygotène :** Appariement des chromosomes homologues par paires appelées bivalents grâce l'intermédiaire d'une structure multi-protéique : le complexe synaptonémal.

- iii. **Pachytène** : Ce stade dure 16 jours et est le plus long de la prophase I. C'est au cours de celui-ci qu'à lieu l'échange de matériel génétique par le biais des crossing-over (**Figure** : 1.6) entre les chromatides non-sœurs appelés nODULES de recombinaison.
 - iv. **Diplotène** : La dissociation du complexe synaptonémal va permettre aux chromosomes homologues d'initier leur séparation. Certains sites d'appariement étroits nommés chiasmas demeurent néanmoins liés permettant une séparation plus progressive des chromosomes et réduisant ainsi le risque d'aneuploïdies (nombre anormal de chromosomes) (Handyside, 2012).
 - v. **Diacinèse** : Cette étape marque la fin de la méiose I et fait office de transition avec la méiose II. Elle est caractérisée par une condensation maximale des chromosomes et la disparition de la membrane nucléaire et du nucléole. Le fuseau méiotique commence à s'assembler, les centromères des chromosomes homologues s'éloignent et les chiasmas glissent progressivement vers les télomères.
-
- b. **La métaphase I** : phase au cours de laquelle les chromosomes vont s'aligner à l'équateur de la cellule pour former la plaque équatoriale.
 - c. **L'anaphase I** : les chromatides sœurs (ou les chromosomes homologues en fonction de la phase méiotique) vont se séparer et migrer aux pôles opposés de la cellule.
 - d. **La télophase I** : qui est l'étape finale, les chromosomes se décondensent et l'enveloppe nucléaire se reforme autour des chromosomes. La cellule mère se sépare alors en deux cellules filles appelées spermatocytes secondaires.

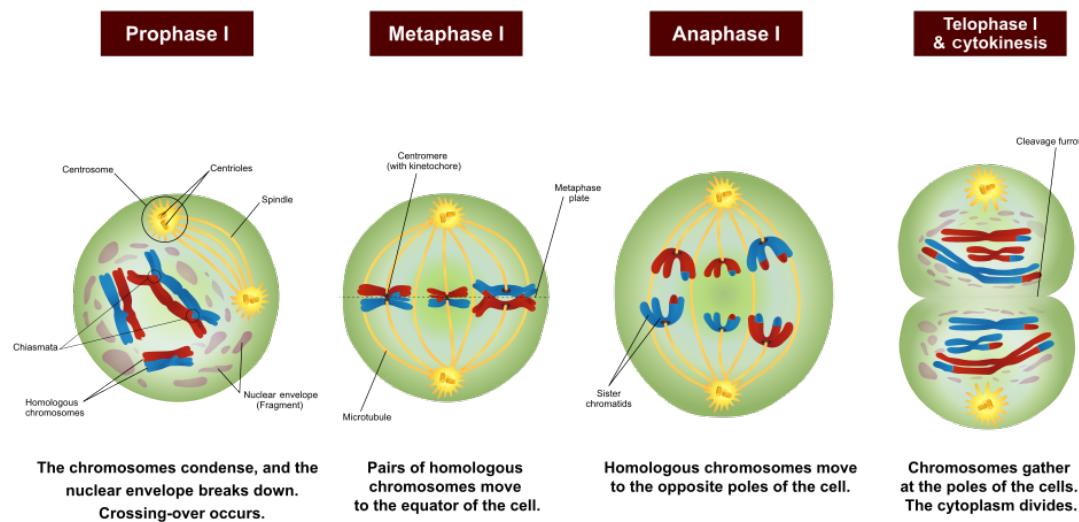


Figure 1.4 – Les différentes étapes de la première division méiotique masculine adapté d'après [Wikipédia](<https://en.wikipedia.org/wiki/Meiosis>)

2. Méiose équationnelle : (Figure : 1.5) La MII est similaire à une division mitotique,

- La prophase II :** Contrairement à la prophase I, la prophase II est très courte. Les chromosomes alors formés de deux chromatides sœurs se dirigent vers la plaque équatoriale.
- La métaphase II :** À ce stade, les chromosomes sont alignés le long de la plaque équatoriale au niveau de leur centromère.
- L'anaphase II :** Les centromères de chaque chromosome se séparent permettant aux chromatides sœurs de se diriger vers les pôles opposés des spermatoctyes II.
- La télophase II :** Comme en télophase I, les cellules mères se séparent en deux cellules filles haploïdes appelées spermatides, contenant chacune n chromosomes.

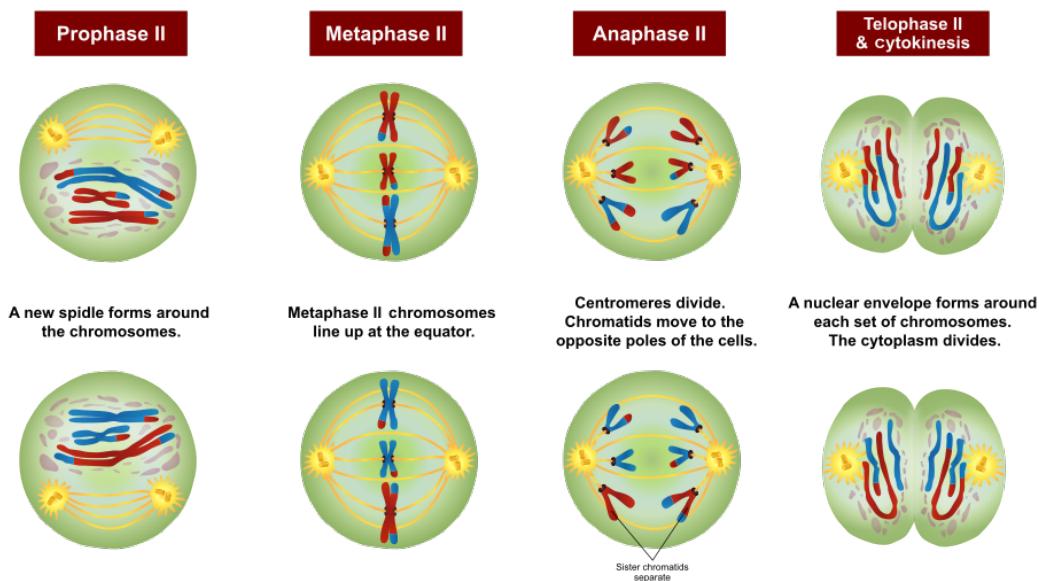


Figure 1.5 – Les différentes étapes de la deuxième division méiotique masculine adapté d'après [Wikipédia](<https://en.wikipedia.org/wiki/Meiosis>)

La première division méiotique aboutit à la formation des spermatocytes secondaires (spermatocytes II). À ce stade, les cellules sont haploïdes et chaque chromosome est composé de deux chromatides sœurs. Après, cette brève étape (environ 1 jour) ainsi qu'une très courte interphase sans réPLICATION de l'ADN, les spermatocytes II vont entrer en deuxième division méiotique. Cette deuxième division est très semblable à une division mitotique. La prophase II, à la différence de la prophase I, est très courte. Lors de cette étape, les chromosomes constitués de chromatides sœurs se dirigent vers la plaque équatoriale. En métaphase II, les chromosomes s'alignent au niveau de leurs centromères. En anaphase II, les chromatides sœurs se séparent l'une de l'autre et migrent vers les pôles opposés des spermatocytes II. Lors de la télophase II, on observe la formation de cellules filles haploïdes appelées spermatides, contenant chacune n chromosomes.

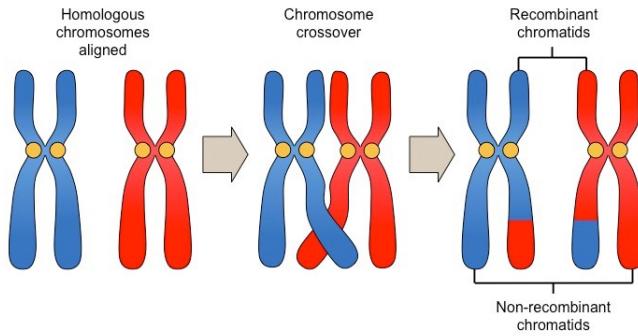


Figure 1.6 – Schéma simplifié d'un enjambement chromosomique (crossing-over)

1.1.4 La spermiogénèse

La spermiogénèse est la phase finale de la spermatogénèse. Elle dure environ 23 jours chez l'humain et peut être subdivisée en sept étapes (Figure : 1.7). La spermiogénèse définit la cytodifférentiation des spermatides en spermatozoïdes. C'est au cours de cette phase que les caractéristiques morphologiques et fonctionnelles du spermatozoïde seront déterminées (Clermont & Oko 1993 à trouver!!!). Elle est caractérisée par 3 événements majeurs : la formation de l'acrosome, la compaction de l'ADN nucléaire et la formation du flagelle. Le développement de l'acrosome et la formation du flagelle commencent au niveau des spermatides rondes (D. Escalier et al., 1991). Pendant l'elongation de la spermatide, le noyau se condense et devient hautement polarisé (Hamilton, D. W., Waites, 1990).

Les spermatides sont situées dans le compartiment adluminal, à proximité de la lumière du tube séminifère. Ce sont de petites cellules (8 à 10 µm) que l'on peut schématiquement diviser en trois classes :

1. **Les spermatides rondes** (Figure : 1.7 1-2) : L'identification de ces cellules représente une difficulté technique. Elles ont cependant pu être décrites en détail par différentes techniques de coloration sous microscope optique (Clermont, 1963, Papic, Katona, & Skrabalo (1988), Schenck & Schill (n.d.), Adelman & Cahill (1989), World Health Organization (1992)). Plusieurs études animales ont pu démontré le potentiel des spermatides rondes à donner la vie à des individus sains et fertiles, (a Ogura, Matsuda, & Yanagimachi, 1994), A. Ogura, Matsuda, Asano, Suzuki, & Yanagimachi (1996), Sasagawa & Yanagimachi (1997)], la même chose ayant été également observée plus récemment chez l'homme (A. Tanaka et al., 2015) bien que le taux de fécondation et d'implantation soit extrêmement faible (Asimakopoulos, 2003). Ils possèdent un noyau rond avec une chromatine pâle et homogène. C'est à partir de ces étapes que démarre la biogenèse de l'acrosome avec la production par l'appareil de Golgi des vésicules pro-acrosomales (phase de Golgi). Les

deux centrioles contenus dans le cytoplasme vont se déplacer au futur pôle caudal. Le centriole proximal est inactif alors que le centriole distal donne naissance à un ensemble de microtubules à l'origine de l'axonème du futur flagelle.

2. **Les spermatides en élongation (Figure : 1.7 3-4) :** Ils peuvent aussi permettre la fécondation et d'initier des grossesses avec un meilleur taux que les spermatides rondes et engendreraient théoriquement moins de risques d'anomalies génétiques (Asimakopoulos, 2003). **A compléter**
3. **Les spermatides en condensation (Figure : 1.7 5-7) :** C'est le stade final de la différentiation de la spermatide en spermatozoïde. À ce stade le noyau est très allongé, avec une partie caudale globulaire et une partie antérieure saillante. La chromatine est sombre et condensée. L'axonème va continuer à s'allonger pour former le flagelle mature. Les différentes organelles inutiles pour la physiologie spermatique et l'excès de cytoplasme vont former la gouttelette cytoplasmique qui va se détacher et donner le corps résiduel qui va ensuite être phagocyté par les cellules de Sertoli (Hermo, Pelletier, Cyr, & Smith, 2010).

Une fois ces étapes de différentiation finies, les spermatides sont relâchées en tant que spermatozoïdes dans la lumière du tube séminifère. Ce procédé est appelé spermatiation.

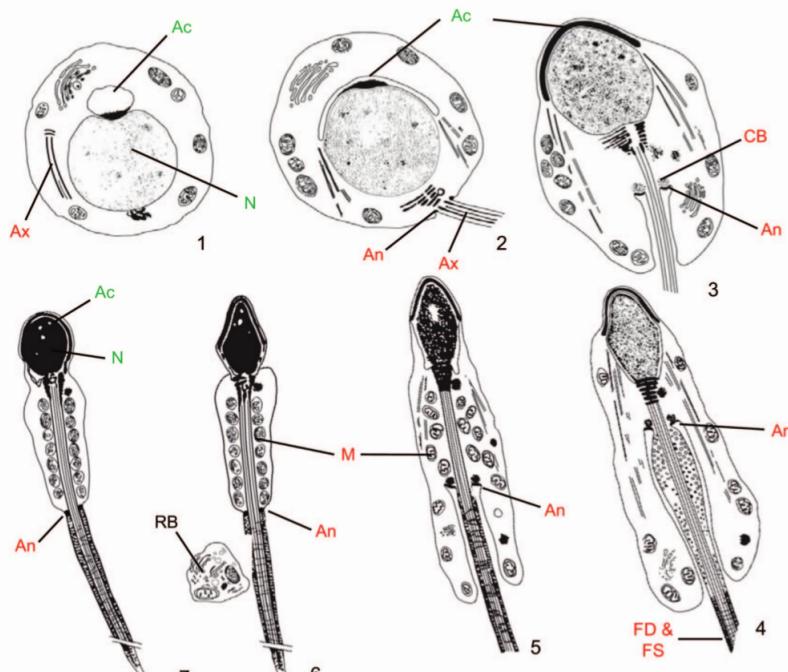


Figure 1.7 – Principales étapes et modifications structurales lors de la spermiogénèse : 1. La spermatide immature avec un gros noyau arrondi. La vésicule acrosomale est attachée au noyau, l'ébauche du flagelle n'atteint pas le noyau. 2. La vésicule acrosomale a augmenté de taille et apparaît aplatie au niveau du noyau. Le flagelle entre en contact avec le noyau. 3-7. Formation de l'acrosome, condensation du noyau et développement des structures flagellaires. Ac, acrosome ; Ax, axonème ; CC, corps chromatoïdes ; CR, corps résiduel ; FD, fibres denses ; GF, gaine fibreuse ; M, mitochondrie ; Ma, manchette. d'après [Toure2011]

1.2 Structure et fonction du spermatozoïde

1.2.1 Anatomie du spermatozoïde

Le spermatozoïde est une cellule hautement différenciée dont la taille, l'orientation et la symétrie sont déterminées. La morphologie générale du spermatozoïde éjaculé est similaire à celle du spermatozoïde testiculaire. Le spermatozoïde humain normal mature mesure environ 60 µm de long et est essentiellement constitué de deux parties : la tête et le flagelle (Figure : 1.8).

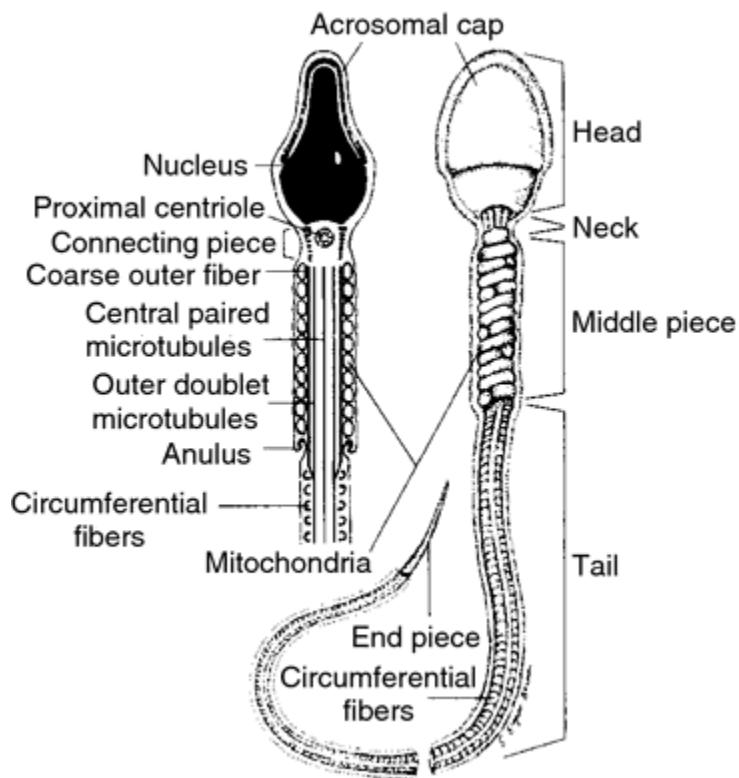


Figure 1.8 – Anatomie simplifiée du spermatozoïde

La tête

1. **L'acrosome** : C'est une vésicule de sécrétion géante située dans la moitié supérieure de la tête du spermatozoïde. Elle se développe à partir de l'appareil de Golgi lors de la spermiogénèse. Au cours de sa formation, l'acrosome forme tout d'abord un granule sphérique qui se colle sur la partie apicale du noyau. En s'aplatissant contre celui-ci, l'acrosome va prendre une forme hémisphérique recouvrant la membrane nucléaire formant la coiffe céphalique. Le rôle de l'acrosome est fondamental dans le processus de fécondation puisqu'il permet

d'excéter notamment l'acrosine, une enzyme de digestion permettant au spermatozoïde de traverser la zone pellucide qui entoure les ovocytes. Ce processus de relargage est appelé réaction acrosomale.

2. **L'acoplaxome** : L'acoplaxome est une structure cytosquelette composée de microfilaments d'actine (F- actine) et de kératine 5. Cette structure est positionnée en face de l'appareil de golgi et contre le noyau et sert de point d'attachement ainsi que de guide aux vésicules pro-acrosomales (Abraham L Kierszenbaum & Tres, 2004). C'est une structure transitoire qui disparaît pour être remplacée par la thèque périnucléaire dans le spermatozoïde mature.
3. **Le noyau** : C'est une structure cellulaire présente dans la majorité des cellules eucaryotes. Il contient l'essentiel du matériel génétique. Le noyau du spermatozoïde est caractérisé par une compaction extrêmement importante de l'ADN. Dans les cellules somatiques l'ADN est enroulé par unité de 146 paires de bases autour d'un octamère d'histones dit de cœur (H2A, H2B, H3 et H4) afin d'organiser les 3 milliards de paires de bases du génome humain dans un noyau de quelques microns (**Figure** : 1.9). L'ADN des spermatides va subir une réorganisation chromatinienne plus importante au cours de la spermatogénèse afin d'augmenter sa compaction. Ainsi, les octamères d'histones présents dans les cellules somatiques sont remplacés par les protéines de transition (TPN1, TPN2) puis par les protamines (PRM1, PRM2) deux protéines riches en arginine et en cystéine (**Figure** : 1.9). L'intégrité des deux protéines composant ce dimère est nécessaire pour la procréation (C. Cho et al., 2001). Cette compaction extrême permet de réduire la taille du noyau, mais aussi de protéger l'ADN d'agents de dégradation comme l'oxydation des bases. Parallèlement à cette condensation chromatinienne se produit un arrêt des processus de transcription cellulaire (A L Kierszenbaum & Tres, 1978). Le noyau du spermatozoïde est donc un noyau au repos, transcriptionnellement inactif (Ward, 1994)

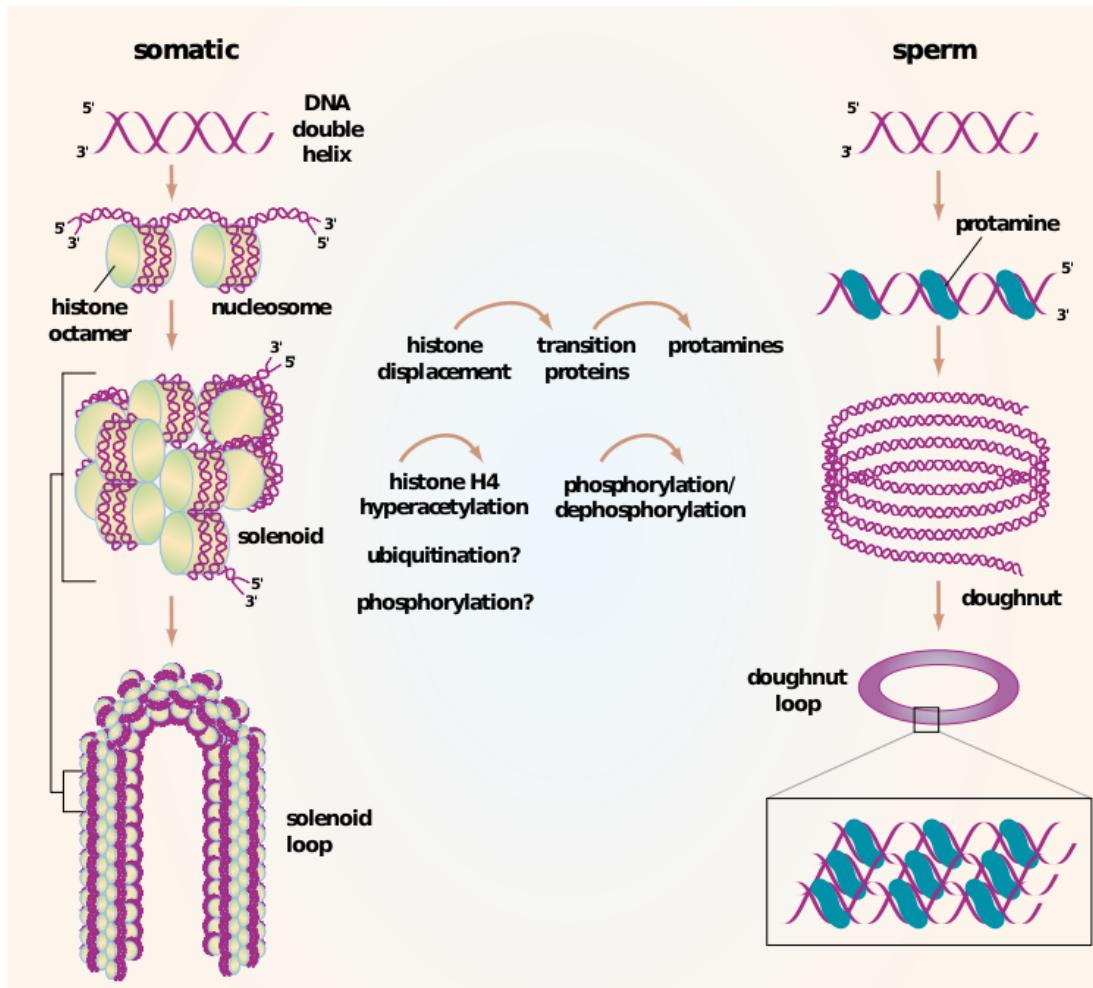


Figure 1.9 – Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes : D'après Braun (2001)

Le flagelle

Le flagelle représente la queue du spermatozoïde. Celui-ci permet, par mouvement d'oscillation à haute vitesse, le déplacement du spermatozoïde. Cette mobilité est générée par un cytosquelette interne extrêmement conservé durant l'évolution appelé l'axonème. Celui-ci est composé de neuf doublets de microtubules périphériques et de deux doublets internes (Inaba, 2003) (Figure : 1.10), on parle alors de structure “9 + 2”. Les doublets externes sont reliés entre eux par des ponts de nexine et au doublet central par des ponts radiaires. Les doublets externes sont également reliés entre eux par les complexes protéiques qui forment les dynéines externes et internes. Ce sont ces protéines qui en exerçant une contraction alternée permettent le mouvement du spermatozoïde.

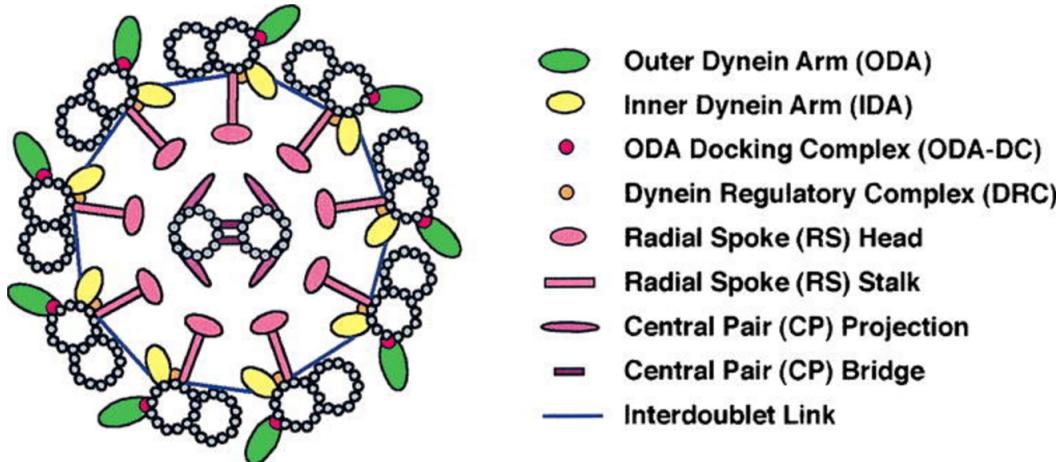


Figure 1.10 – Structure simplifiée de l’axonème d’après [@Inaba2003] : L’axonème est constitué de neuf doublets de microtubules périphériques reliés entre eux par des liens de nexine et d’un doublet central relié aux doublets périphériques par des ponts radiaires

Le flagelle du spermatozoïde peut être divisé en trois parties distinctes (**Figure : 1.11**) :

1. **La pièce intermédiaire** : Elle fait jonction avec la tête du spermatozoïde et est composée de la gaine de mitochondrie qui fournira une partie de l’énergie nécessaire au battement flagellaire (grâce à la phosphorylation oxydative qui produit de l’ATP). L’axonème qui se prolonge dans la pièce principale et un ensemble de neuf faisceaux de fibres denses.
2. **La pièce principale** : Ici, la gaine de mitochondrie a disparue ainsi que deux des faisceaux de fibres denses présents dans la pièce intermédiaire. On note cependant la présence d’une structure supplémentaire, la gaine fibreuse. Cette gaine entoure l’axonème et comporte deux épaississements diamétralement opposés, appelés colonnes longitudinales sur lesquelles s’insèrent les fibres denses 3 et 8. C’est le long de la gaine fibreuse qu’est produite la majorité de l’énergie nécessaire au glissement des microtubules (Eddy, 2007).
3. **La pièce terminale** : Elle est située au niveau de l’extrémité distale du flagelle et ne contient que l’axonème (Inaba, 2003).

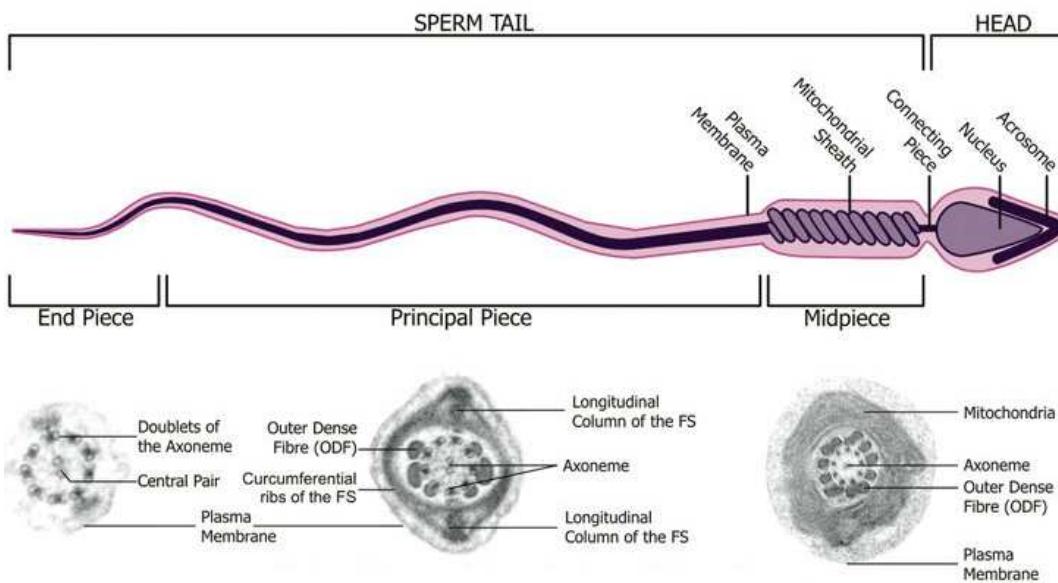


Figure 1.11 – Structure du flagelle d'un spermatozoïde d'après Borg et al. (2010) : Coupes transversales en microscopie électronique. Le flagelle se compose de trois parties : la pièce intermédiaire, contenant les mitochondries, la pièce principale et la pièce terminale. L'axonème, en position centrale, parcourt tout le flagelle. Des structures périaxonémiales sont observables : les fibres denses dans la pièce intermédiaire et principale, et la gaine fibreuse dans la pièce principale seulement.

1.2.2 Fonction du spermatozoïde

En plus d'être unique dans sa morphologie, le spermatozoïde l'est aussi dans sa fonction puisque c'est la seule cellule produite de manière endogène et dont l'action est exercée de manière exogène.

1.3 L'infertilité masculine

L'organisation mondiale de la santé définit l'infertilité comme étant : “une pathologie du système reproductif définie par l'échec d'une grossesse clinique après 12 mois ou plus de rapports sexuels réguliers non protégés” (Who.int. 2013-03-19. Retrieved 2013-06-17). L'étude de l'infertilité représente un des enjeux scientifique et médical majeur de ces dernières années. On estime qu'environ 10 à 15% des couples humains font face à des problèmes d'infertilité soit plus de 70 millions de personnes dans le monde (Boivin, Bunting, Collins, & Nygren, 2007). Dans la moitié des cas, la cause sous-jacente serait masculine. On estime que les facteurs causaux sous-jacents de l'infertilité masculine peuvent être attribués à des toxines environnementales, des troubles systémiques tels que la maladie hypothalamo-hypophysaire, les cancers testiculaires et l'aplasie des cellules germinales. Les facteurs génétiques, y compris les aneuploïdies et les mutations de gènes uniques, contribuent également à l'infertilité masculine. Cependant, aucune cause n'est identifiée dans près de la moitié des cas. Comme nous avons pu le voir, la spermatogénèse est une succession de processus complexes qui s'effectue de manière coordonnée, de fait la moindre altération génétique affectant une seule de ces étapes est susceptible d'entrainer un phénotype d'infertilité (Barratt, 1995 **A TROUVER**).

1.3.1 Les différents phénotypes d'infertilité masculine

Chez l'homme, l'infertilité est associée à une altération quantitative et / ou qualitative des spermatozoïdes présents dans l'éjaculat. L'ensemble de ces altérations peuvent être détectées et quantifiées dans des laboratoires spécialisés par réalisation d'un spermogramme. Au cours de celui-ci, plusieurs critères tels que le volume de sperme sécrété, son pH, la quantité et la vitalité des spermatozoïdes qu'il contient seront évalués. La proportion de cellules immatures sera elle aussi analysée. Ces cellules rondes, se retrouvent à la fois dans l'éjaculat des individus ayant une quantité de spermatozoïdes “normale” (Michael & Joel, 1937, M. Tomlinson et al. (1993)), chez les individus présentant une quantité basse de spermatozoïdes (MacLeod, 1970, M. J. Tomlinson, Barratt, & Cooke (1993)) ou en étant dépourvu (Kurilo, Liubashevskaya, Dubinskaia, & Gaeva, 1993). Cependant, leur nombre augmente tandis que la quantité de spermatozoïde diminue (SPERLING & KADEN, 1971).

Anomalies liées à la quantité spermatique

Chez l'humain, l'arrêt de la spermatogénèse est défini comme l'incapacité des cellules spermatogénétiques à devenir des spermatozoïdes matures. Elle peut survenir à n'importe quelle étape de la formation des cellules germinales. Les blocages méiotiques,

au stade de spermatocyte I sont les plus fréquents, suivis par l'arrêt au niveau des spermatides et moins fréquemment au niveau des spermatogonies (Girgis, Etriby, Ibrahim, & Kahil, 1969).

1. **L'oligozoospermie** : L'oligozoospermie est définie comme un phénotype d'infertilité masculine caractérisé par une production inférieure à 15 millions de spermatozoïdes par ml de sperme (T. G. Cooper et al., 2010). Un arrêt de la spermatogénèse a été observé dans 4 à 30% des biopsies testiculaires des hommes présentant une oligospermie sévère (Colgan, Bedard, Strawbridge, Buckspan, & Klotz, 1980, Levin (1979), Soderström & Suominen (1980), WONG, STRAUS, & WARNER (1973)). Cet arrêt a longtemps été considéré comme sans espoir pour les couples désirant concevoir, jusqu'à l'émergence de l'injection mécanique d'un spermatozoïde dans l'ovocyte appelé *intracytoplasmic sperm injection* (ICSI) (Palermo, Joris, Devroey, & Van Steirteghem, 1992)
2. **L'azoospermie** : Comme l'oligozoospermie, l'azoospermie est un phénotype d'infertilité masculine cette fois-ci caractérisé par l'absence totale de spermatozoïdes dans l'éjaculat. On distingue des causes excrétoires empêchant l'excrétion des spermatozoïdes, on parle alors d'azoospermie obstructive et des causes sécrétoires, les plus fréquentes, accompagnées d'un défaut de la spermatogenèse, on parle alors d'azoospermie non-obstructive.

Anomalies liées à la morphologie

Ces anomalies sont observables en effectuant un spermocytogramme. Plusieurs classifications ont été établies, cependant, c'est la classification de David modifiée (**Table : 1.12**) qui est la plus retenue en France. Pour ce faire, on procède généralement à une observation de 100 spermatozoïdes au cours de laquelle l'ensemble des anomalies observées sont relevées et quantifiées permettant ainsi de définir un index d'anomalies multiple (nombre total d'anomalies/nombre de spermatozoïdes anormaux) révélant le nombre moyen d'anomalies par spermatozoïde.

Table 1. Morphological abnormalities^a and sperm pathology.

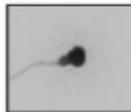
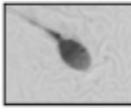
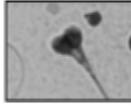
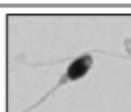
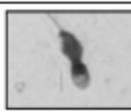
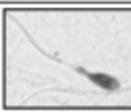
Sperm defect ^a (Light microscopy, final magnification [$\times 1\,000$])	Possible related TEM observations
Elongated head Major axis ↑ Minor axis =	 Abnormally shaped head and abnormally condensed chromatin
Thin head Major axis = Minor axis ↓	 Abnormally shaped head and abnormally condensed chromatin
Microcephalous head Major axis ↓ Minor axis ↓	 Excessive shrinking of the nucleus and abnormally condensed chromatin
Macrocephalous head Major axis ↑ Minor axis ↑	 Insufficient shrinking of the nucleus and abnormally condensed chromatin
Multiple heads More than one head	 Two or more closed or dissociated heads with or without a common acrosome or midpiece
Abnormal postacrosome region All outline and texture anomalies of the region	 Abnormally shaped post-acrosomal component and disorganization of the cape structures; abnormal DNA condensation
Abnormal acrosome region All outline, size and texture anomalies of the region	 Absent or abnormally shaped or sized acrosome, incomplete acrosome and/or abnormal appearance of the underlying nucleus
Abnormal residual cytoplasm Residual cytoplasm > 30% of head size	 Abnormally wide cytoplasmic remnant containing subcellular components
Thin midpiece Diameter of midpiece < diameter of the proximal principal piece	 Partial or absent mitochondrial sheath

Figure 1.12 – Différentes anomalies morphologiques du spermatozoïde selon la classification de David modifiée adapté... TABLEAU à adapter et à insérer !!!!! d'après [@Auger2010]

Anomalies liées à la mobilité

Le succès du passage du spermatozoïde le long du tractus génital féminin dépend en grande partie de la mobilité et de la vitesse du spermatozoïde (Lindholmer, 1974, Björndahl (2010)). La vitesse moyenne d'un spermatozoïde étant de 25 µm/s. Une mauvaise mobilité observée dans plus de 50% des spermatozoïdes éjaculés se révèle être un prédicteur de l'échec de la fécondation (Aitken, Sutton, Warner, & Richardson, 1985).

1.3.2 La génétique de l'infertilité

Comme il a déjà été dit, il est estimé que 10 à 15% des couples humain font face à des problèmes d'infertilité. Par ailleurs, 30% des infertilités restent inexplicées et près de 40% ont des causes incertaines. Ainsi, l'infertilité masculine d'origine génétique pourrait concerner près de 1 homme sur 40 (Tüttelmann et al., 2011).

Les causes fréquentes

1. **Les microdélétions du chromosome Y :** Le chromosome Y est un petit chromosome atteignant une taille d'environ 53 Mb et est porteur de 78 gènes principalement impliqués dans la différentiation sexuelle masculine et la spermatogénèse (Skaletsky et al., 2003). De fait, le chromosome Y représente une région d'intérêt évidente dans l'étude de facteur génétique liés à l'infertilité masculine. L'évolution des technologies a permis de mettre en évidence des délétions invisibles au caryotype dans la région du facteur AZF (*Azoospermia Factor*). Cette région peut être subdivisée en trois sous-parties, AZFa, AZFb et AZFc (**Figure : 1.13**). Depuis plusieurs années, de nombreuses séries de patients azoospermiques ou oligozoospermiques ont été étudiées et publiées et tendent à montrer que les microdélétions du chromosome Y seraient responsables de 10% des cas d'azoospermie non-obstructive et chez 5% des cas d'oligozoospermie sévère (<5 millions de spermatozoïdes/ml) (Hotaling & Carrell, 2014).

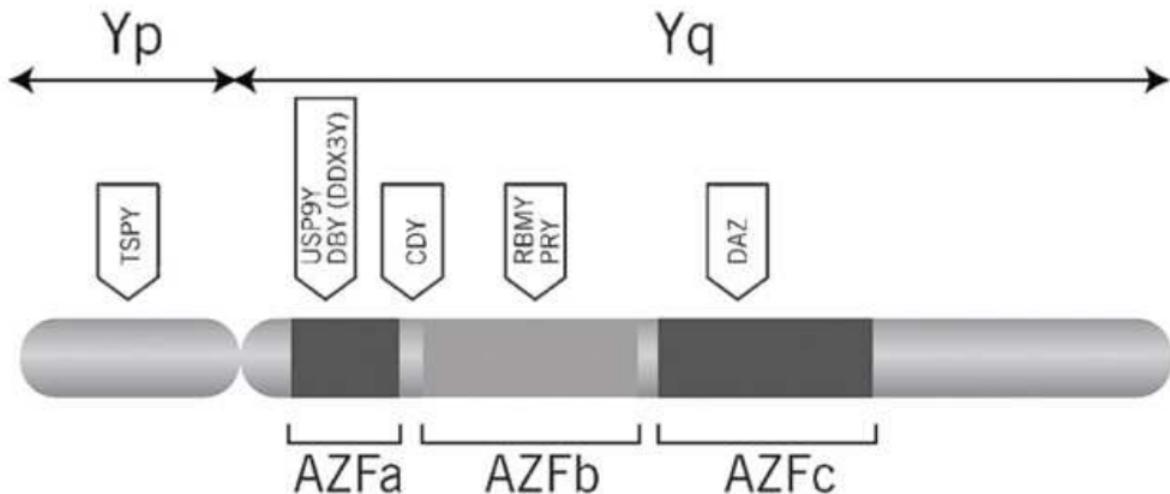


Figure 1.13 – Représentation schématique du chromosome Y adapté d'après [OflynnOBrien2010] : Visualisation la région AZF ainsi que des trois sous-régions AZF a, b, c et des principaux gènes compris dans chacune des sous-régions

2. **Anomalies chromosomiques** : Des anomalies chromosomiques de nombre ou de structure impliquant les autosomes ou, le plus souvent, les gono-somes, peuvent être impliquées dans des cas d'infertilité masculine. Le pourcentage d'individus concernés varie entre 2 et 8% et peut atteindre 15% pour les patients azoospermiques soit 10 à 20 fois la fréquence retrouvée dans la population générale (Ravel, Berthaut, Bresson, Siffroi, & Genetics Commission of the French Federation of CECOS, 2006).
 - a. **Syndrome de Klinefelter** : Le syndrome de Klinefelter (ou 46, XXY) fut décrit pour la première fois en 1942 par Harry F. Klinefelter et décrit une affection due à la présence d'un chromosome X supplémentaire suite à une erreur de ségrégation des chromosomes au moment de la méiose. Sa prévalence dans la population générale est estimée à environ 1 sur 1200 (1 homme sur 600) (Bojesen & Gravholt, 2011) mais elle est environ 50 fois supérieure chez les patients infertiles azoospermiques (Gekas et al., 2001).
 - b. **Les anomalies de structure** : Les translocations et les inversions sont les anomalies de structures retrouvées le plus fréquemment chez les patients infertiles.
 - i. La translocation est définie comme l'échange de matériel génétique entre deux chromosomes non homologues. On en distingue deux types, les translocations réciproques et les translocations robertsonniennes [TODO : définir ces deux types de translocation]. Les premières sont retrouvées 4 à 10 fois plus fréquemment chez les patients infertiles que dans la population générale (Elliott & Cooke, 1997), les secondes sont retrouvées chez 1.6% des patients oligozoospermiques et 0.09%

- des patients azoospermiques (O'Flynn O'Brien, Varghese, & Agarwal, 2010).
- ii. Les inversions chromosomiques caractérisent le mécanisme de cassure d'un fragment de chromosome suivi de son retournement à 180° et sa réintégration à la même position. Ces inversions vont gêner l'appariement des chromosomes homologues (formation d'une boucle d'inversion) pendant la méiose et sont, comme les translocations, retrouvées plus fréquemment chez les patients infertiles que dans la population générale (Krausz & Forti, 2000).
 - c. **Autres anomalies chromosomiques** : Parmi les anomalies chromosomiques responsables d'infertilité masculine, on peut par exemple citer les hommes de formule 46,XX. Ces patients sont généralement totalement infertiles et présentent une azoospermie par absence des sous- régions AZF a, b et c (Vorona, Zitzmann, Gromoll, Schüring, & Nieschlag, 2007) bien qu'ils aient un phénotype masculin normal. Ces anomalies sont souvent le fait de la translocation du gène SRY sur un des chromosomes X du patient.
3. **Mutations du gène *CFTR*** : L'identification du gène *CFTR* (*Cystic Fibrosis Transmembrane conductance Regulator*) chez les patients atteints de mucoviscidose et présentant une agénésie bilatérale des canaux déférents (ABCD) a permis d'associer ce gène au phénotype d'azoospermie obstructive. Cette malformation serait responsable de 2% des cas d'infertilité masculine et de 25% des cas d'azoospermie obstructive (J. Yu, Chen, Ni, & Li, 2012).

Bien que la prévalence de ces anomalies génétiques varie en fonction du phénotype concerné, il est estimé que ces défauts soient seulement retrouvés chez 5% des cas d'infertilité masculine tous phénotypes confondus. Cette observation suggère fortement l'implication de nombreux autres gènes encore inconnus dans les différents phénotypes d'infertilité masculine (Nieschlag et al., 2010 A trouver).

Les nouveaux gènes

1. Les anomalies morphologiques liées à la tête du spermatozoïde :

- a. **La macrozoospermie** : Ce phénotype d'infertilité masculine rare est caractérisé par la présence de 100% des spermatozoïdes de l'éjaculat présentant une tête anormalement grosse ainsi que plusieurs flagelles. Il fut observé pour la première fois en 1978 (Nistal, Paniagua, & Herruzo, 1978), mais ce n'est qu'en 2007 qu'une explication génétique fut enfin trouvée. Une étude portant sur 14 patients nord Africains a permis d'identifier la délétion c144delC du gène *AURKC* (*Aurora kinase C*) comme responsable du phénotype de l'ensemble des individus de l'étude (Dieterich et al., 2007).

Depuis, d'autres études ont permis d'associer d'autre variants sur ce même gène à ce phénotype [INSERT REF]. Des anomalies du gène *AURKC* seraient ainsi responsables d'environ 83.7% des cas macrozoospermie chez des patients non apparentés [INSERT REF]. Le gène *AURKC*, étant impliqué dans la méiose, conduit lorsqu'il est muté à un blocage de la première division méiotique entraînant la production de spermatozoïdes tétraploïdes, c'est à dire, portant une quantité de matériel génétique quatre fois supérieure à la normale [INSERT Dieterich 2009].

- b. **La globozoospermie :** La globozoospermie est aussi un phénotype rare d'infertilité dont la prévalence est estimée à de 0,1%. Il fut identifié pour la première fois en 1971 et est caractérisé par la présence dans l'éjaculat d'une majorité de spermatozoïde dépourvu d'acrosome empêchant ainsi le spermatozoïde de franchir la zone pellucide de l'ovocyte et compromettant ainsi la fécondation (A. Dam et al., 2006, C. G. S. Sen, Holstein, & Schirren (1971), A. F. Holstein, Schirren, & Schirren (1973)). En 2007, une étude familiale a permis de lier ce phénotype à la mutation c.848G>A dans le gène *SPATA16* (*spermatogenesis-associated protein 16*) (A. H. Dam et al., 2007) dont la protéine va, au cours de la spermatogénèse fusionner avec les vésicules proacrosomales pour former l'acrosome (A. H. Dam et al., 2007, L. Lu, Lin, Xu, Zhou, & Sha (2006)). Plus tard, en 2011, une étude portant sur 20 patients tunisiens permit d'identifier une délétion homozygote de 200 kb emportant la totalité du gène *DPY19L2* (*Dpy-19 Like 2*) chez 15 des 20 patients (Harbuz et al., 2011). cf globo
- c. **Spermatozoïdes acéphaliques :** Ce phénotype rapporté plusieurs fois (Hector E. Chemes & Rawe, 2010, Panidis et al. (2001), H E Chemes et al. (1987)) caractérise les patients présentant des spermatozoïdes dépourvus de tête dans leur éjaculat. Une étude récente a pu lier ce phénotype à une mutation c.824C>T homozygote ainsi qu'à deux variants hétérozygotes composites c.1006C>T et c.485T>A dans le gène *SUN5* (F. Zhu et al., 2016) qui avait précédemment été décrit comme localisant à la jonction noyau / flagelle du spermatozoïde (S. Yassine et al., 2015).

2. Les anomalies liées au flagelle et à la motilité :

- a. **Phénotype MMAF :** Le phénotype MMAF (*Multiple morphological abnormalities of the sperm flagella*) décrit les patients atteints d'asthenozoospermie dont les spermatozoïdes présentent de multiples anomalies morphologiques touchant en particulier les flagelles. Plus précisément, ce phénotype décrit les asthenozoospermie résultant d'une mosaïque d'anomalies morphologiques au niveau du flagelle tel que l'absence totale de flagelle, des flagelles enroulés, courts, anguleux... (C. Coutton, Escoffier, Martinez, Arnoult, & Ray, 2015, Ben Khelifa et al. (2014)). Récemment, le

gène *DNAH1* (*Dynein Axonemal Heavy Chain 1*) codant pour une dynéine de la chaîne lourde de l'axonème a été retrouvé muté chez près d'un patient sur trois dans sa cohorte comportant 18 patients (Ben Khelifa et al., 2014). Deux autres études ont retrouvé des mutations dans le gène *DNAH1* chez des patients venant de Chine, d'Iran et d'Italie, laissant suggérer que ce gène est l'un des acteurs majeurs dans le syndrome MMAF (X. Wang et al., 2017, Amiri-Yekta et al. (2016)).

3. **Les échecs de fécondation du spermatozoïde :** Au moment de la fécondation, l'activation ovocytaire repose sur le relargage par le spermatozoïde de "facteurs spermatiques" qui déclenchent un signal de calcium, constitué d'oscillations Ca^{2+} . Ce processus est médié par une protéine spécifique du spermatozoïde, la *phospholipase C Zeta 1* (PLC ζ 1) codée par le gène *PLCZ1* (Nomikos, Kashir, Swann, & Lai, 2013, Amdani, Jones, & Coward (2013)). Plusieurs cas d'échec d'activation ovocytaire ont été liés à l'absence ou la mauvaise localisation de la protéine PLC ζ 1. Malgré cela, aucune preuve génétique directe n'avait été reportée jusque récemment où deux mutations au sein du gène *PLCZ1* furent retrouvées chez un patient (Heytens et al., 2009) et un peu plus tard une mutation homozygote chez deux frères consanguins (Escoffier et al., 2016).

1.4 Les techniques d'analyses génétiques

L'acide désoxyribonucléique (ADN) a été identifié comme étant le porteur de l'information génétique par Oswald Theodore Avery en 1944. Sa structure en double hélice composée par quatre bases, la thymine (T), l'adénine (A), la guanine (G) et la cytosine (C) fut caractérisée en 1953 par James D. Watson et Francis Crick. Cependant, l'existence “d'entités d'information génétiques discrètes” que sont les gènes fut suggéré dès la deuxième moitié du XIX^{ème} siècle grâce aux travaux de Gregor Mendel portant sur l'hérédité de certains traits chez le pois. Depuis, de nombreuses méthodes permettant de lier le phénotype d'un individu à son génotype ont vu le jour au gré des améliorations technologiques.

1.4.1 Approche “gènes candidats”

L'approche gène candidat consiste à rechercher des mutations chez un patient dans un gène cible. Le choix du gène cible se fera en fonction de plusieurs critères. Le premier d'entre eux est l'étude de gènes reliés à des phénotypes proche du phénotype étudié dans différents modèles animaux et notamment murins. Dans ce cas, les mutations seront recherchées sur le gène orthologue humain (Boer, Vries, & Ramos, 2015). Une autre possibilité consiste à rechercher des variants dans des gènes paralogues à un gène précédemment identifié avec l'idée sous-jacente que leur structure proche implique une fonction similaire. Enfin la dernière méthode consiste à étudier des gènes connus comme étant des partenaires de gènes déjà identifiés dans cette pathologie en supposant que si un variant dans un gène donné entraîne une pathologie, un variant dans un partenaire de ce gène pourrait entraîner le même phénotype. Cette approche est bien souvent infructueuse dû en grande partie à l'hétérogénéité génétique des phénotypes étudiés, au nombre limité de patients testés (Elinati et al., 2012) et aux connaissances souvent incomplètes sur le phénotype. De fait, cette approche a quasiment disparu au profit des méthodes à haut débit que sont les puces et le séquençage nouvelle génération (NGS), néanmoins, cette méthode compte à son actif plusieurs succès retentissants [INSERT PETITE LISTE].

1.4.2 Les puces

Les puces à ADN / ARN ont

1. Bref historique de la technologie
2. A quoi ça sert
3. Comment ça marche

Les puces à SNP, le génotypage... (titre à revoir)

Depuis leur emmergence, les puces à ADN ont largement été utilisé comme plateforme de génotypage pour les SNP connus.

Les puces à indels

1. MLPA
2. CGHarray

Les puces à expression

L'utilisation principale des puces à ADN a été de mesuré l'expression des gènes dans un tissus donné. Dans cette application, l'ARN est extrait des cellules d'intérêt puis est généralement convertit en ADNc. L'ADNc est ensuite hybride à la puce qui subira ensuite une étape de lavage. L'intensité de fluorescence est ensuite mesurée à chaque spot de la puce. L'intensité du signal sera ensuite le reflet du niveau d'expression d'un gène.

Limitation

Bien qu'elles aient été extrêmement utiles dans un champ d'application varié, cette technologie présente plusieurs limitations. La première est que la mesure que la puce renvoie

pour les génomes complexes tel que le génome humain, il est extrêmement difficile (voire impossible) de ne pas

En effet, les puces à ADN peuvent détecter uniquement ce pourquoi elles ont été. Ainsi, si la solution que l'on hybride sur la puce contient des séquences d'ADN ou d'ARN pour lesquelles il n'y a aucune sonde complémentaire sur la puce, celles-ci ne seront pas détectées. Cela peut avoir de grandes répercussions puisque par exemple dans le cas des gènes à expression, les gènes qui n'ont pas encore été annotés risquent de ne pas être représentés sur la puce.

1.4.3 Le séquençage NGS

Le terme séquençage de l'ADN fait référence à l'ensemble des techniques permettant de déterminer l'ordre des nucléotides A, T, C et G de l'intégralité ou d'une partie d'une molécule d'ADN. Avant de parler des nouvelles technologies de séquençage (NGS) faisons un bref historique du séquençage de l'ADN. En 1977 Frederick Sanger développe une technologie de séquençage d'ADN basée sur la méthode *chain-termination*. Ce procédé est désormais connu sous le nom de séquençage Sanger. D'autres méthodes furent développées à la même période, notamment celle de Walter Gilbert basée sur la modification chimique de l'ADN, cependant sa grande efficience et sa faible utilisation de la radioactivité permirent au séquençage Sanger de s'imposer comme référence dans la "première génération" de séquenceur à application commerciale et de recherche. Apparu en 1998, les instruments de séquençage automatique ainsi que les logiciels associés utilisant le séquençage par capillarité et la technologie Sanger furent les outils principaux qui permirent la complétion du *human genome project* en 2001 (F. S. Collins, Morgan, & Patrinos, 2003).

Contrairement à la méthode Sanger, le NGS "lit" des fragments d'ADN, provenant d'un génome entier, de manière aléatoire. On parle alors de séquençage de génomes entiers ou *whole genome sequencing* (WGS). Pour cela, la molécule d'ADN est "coupée" en plusieurs fragments d'une taille donnée. Ce sont ensuite ces fragments qui seront, après une étape d'amplification spécifique aux différentes plateformes, séquencés simultanément. C'est pourquoi on parle souvent de séquençage parallèle massif pour décrire le NGS. Le produit de ce séquençage est appelé *read*. Cette technologie est avantageuse de par la masse de *reads* qu'elle produit et par son faible coût par bases séquencées (Metzker, 2010). Ces caractéristiques ont permis au séquençage Haut-débit d'être couramment utilisé dans le domaine de la recherche clinique.

La taille des *reads* obtenus par séquençage NGS est nettement inférieure à celle atteinte par le séquençage Sanger. À l'heure actuelle, les *reads* obtenus par séquençage NGS ont une taille comprise entre 50 et 500 pb pour la plupart des plateformes contre une taille d'environ 800 nucléotides obtenus par Sanger (**Figure : 1.14**), c'est pour cela que les résultats du séquençage NGS sont appelés des *reads courts* ou *short reads*. Étant donné que le NGS produit à l'heure actuelle des *reads* courts la notion de couverture est importante et représente l'un des critères majeur à considérer dans l'analyse des données (D. Sims, Sudbery, Ilott, Heger, & Ponting, 2014). La couverture est définie comme le nombre de *reads* qui, après l'étape d'alignement, se chevauchent les uns les autres au sein d'une région génomique spécifique. Par exemple, une couverture de 30x pour le gène XXXX signifie que chaque nucléotide de ce gène est chevauché par au moins 30 *reads* distincts.

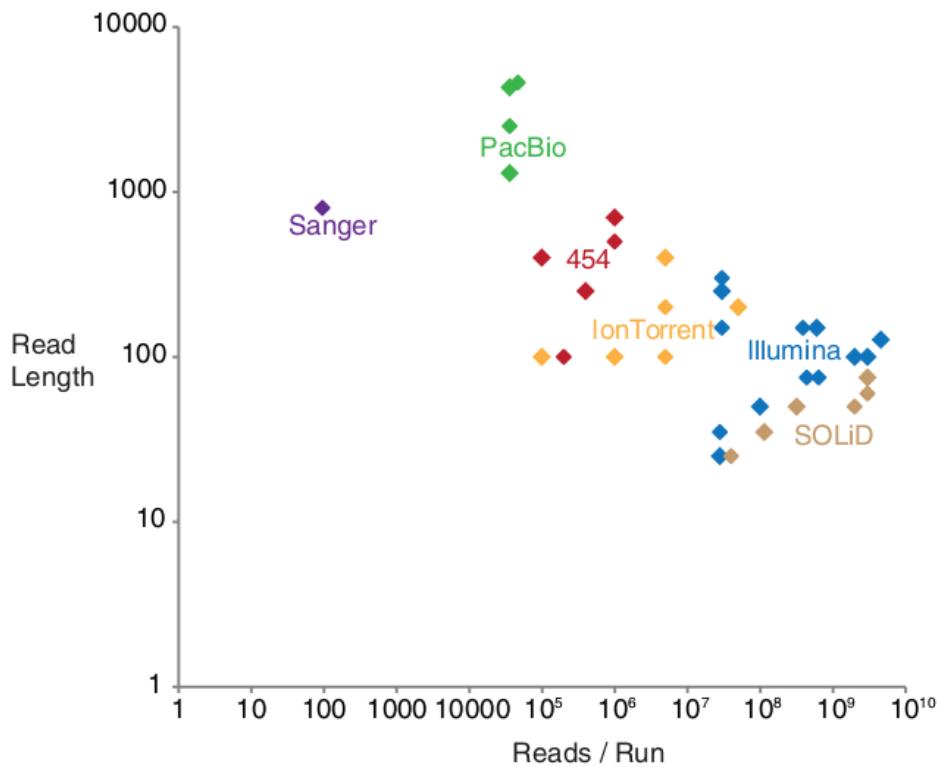


Figure 1.14 – Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée d'après [@Hodkinson2015] : Chaque point représente une plateforme de séquençage, la couleur détermine la marque du séquenceur

La capture des parties à séquencer, avantages et inconvénients

Pour de nombreuses applications, il peut être intéressant de ne séquencer qu'une partie du génome et non pas son intégralité. Dans cette sous partie de génome ciblé on peut trouver par exemple : une région génomique spécifique à laquelle une pathologie a déjà été associée, l'ensemble des exons de certains gènes candidats, ou encore l'intégralité des exons de l'ensemble des gènes codant pour une protéine. Dans ce dernier cas on parle alors de séquençage exomique ou *whole exome sequencing* (WES). Les principaux avantages du WES par rapport au WGS sont son coût réduit ainsi qu'une masse de données moins importantes à stocker et à analyser. En effet, l'ensemble de l'exome ne représente qu'environ 1% du génome entier. On considère cependant que ces parties codantes contiennent plus de 90% des anomalies responsables de pathologies génétiques chez l'homme. Pour ces raisons, le WES considéré comme le standard dans le cadre de recherche sur des pathologies génétiques et révèle être un outil puissant pour l'identification de variants associés à des pathologies (S. B. Ng et al., 2010). Le

procédé de séquençage est identique au WGS, il est simplement précédé d'une étape d'enrichissement au cours de laquelle les exons sont capturés par hybridation à des sondes. De fait les exons capturés sont donc dépendants du kit de capture utilisé, cette technique permet donc de séquencer uniquement les exons connus et ciblés par les sondes. Il faut également noter que depuis quelques années, plusieurs études ont remis en cause l'intérêt du WES au profit du WGS, notamment car le WGS fournit une meilleure couverture sur l'exome que le WES (Lelieveld, Spielmann, Mundlos, Veltman, & Gilissen, 2015, Meienberg, Bruggmann, Oexle, & Matyas (2016)). De plus le WES montre une plus grande sensibilité au pourcentage de GC contenu dans la région à séquencer et à la sélection des kits de capture utilisés (Meienberg et al., 2016). Ainsi, bien que le WES soit encore à l'heure actuelle le choix privilégié dans la majorité des études (citation...), la réduction des coûts de séquençage et du stockage des données, pourraient permettre prochainement au WGS de remplacer totalement le WES ainsi que l'ensemble des techniques impliquant la capture de séquences ciblées (Meienberg et al., 2016).

L'amplification

Dans la plupart des technologies, la phase de séquençage est précédée par une étape d'amplification de l'ADN. Cette amplification se fait dans la grande majorité des cas sur une surface solide exceptée pour la PCR en émulsion qui s'effectue en phase aqueuse. Elle permet d'obtenir dans une région définie plusieurs milliers de copies du même fragment d'ADN, appelés des clones. Cette étape assure que le signal émis lors du séquençage pourra être distingué du bruit. Chacun de ces *spots* d'amplification appelés aussi centre de réaction, se retrouve donc être le représentant d'un unique fragment d'ADN. Ceux-ci seront ensuite séquencés parallèlement aux autres *spots*. Une plateforme de séquençage peut gérer plusieurs millions de ces centres de réactions simultanément, séquençant ainsi plusieurs millions de molécules d'ADN en parallèle, donnant ainsi le nom de séquençage massif en parallèle à ces techniques. Cette étape d'amplification est généralement précédée d'une phase de fragmentation de l'ADN. Cette fragmentation peut être physique, enzymatique ou bien chimique. Ce sont les résidus d'ADN résultant de cette fragmentation qui seront ensuite amplifiés. Il existe quatre stratégies utilisées pour le clonage de l'ADN dans le cadre du NGS :

1. **La PCR en émulsion ou emPCR (Figure : 1.15 - a)** : Le patron d'ADN fragmenté simple brin est lié à une séquence adaptatrice complémentaire et est capturé par une gouttelette aqueuse appelée micelle contenant une bille recouverte d'adaptateur complémentaire à celui fixé sur le fragment d'ADN ainsi que tous les composants nécessaires à la réaction de PCR. En respectant un ratio nombre de molécules d'ADN / nombre de billes, on va fixer un seul fragment d'ADN sur chaque bille. Chacune de ces billes seront donc, en fin de réaction, recouverte par plusieurs milliers de copies de la même séquence d'ADN.

2. **L'amplification par pont sur face solide (Figure : 1.15 - b)** : Les fragments d'ADN sont liés à des séquences adaptatrices et lié par une de leurs extrémités à une amorce fixée sur un support solide. Du fait de la dilution, les molécules d'ADN se trouvent éloignées les unes des autres. L'extrémité libre du fragment interagit avec les amorces situées à proximité formant une structure en pont, d'où le nom de PCR en pont ou *bridge-PCR*. La PCR va alors synthétiser un deuxième brin complémentaire aux fragments immobilisés sur le support. En procédant à des cycles de température comme pour une réaction PCR classique, on obtient à l'emplacement de chaque molécule d'ADN un massif de molécules fixé sur la plaque, toutes identiques à la molécule initiale.
3. **Amplification par modèle mobile ou *walking-template* (Figure : 1.15 - c)** : L'ADN fragmenté est lié à un adaptateur et lié à une amorce complémentaire fixée sur un support solide. Le brin complémentaire du fragment sera synthétisé par PCR à partir de l'amorce fixée. La molécule double brin nouvellement formée sera ensuite partiellement dénaturée permettant à l'extrémité libre de se fixer à une séquence amorce voisine. Des amorces *reverse* sont ensuite utilisées pour resynthétiser un fragment d'ADN libre à partir des fragments fixés sur le support.
4. **(Figure : 1.15 - d) : PAS DU TOUT COMPRIS LE MECHANISME !!!**

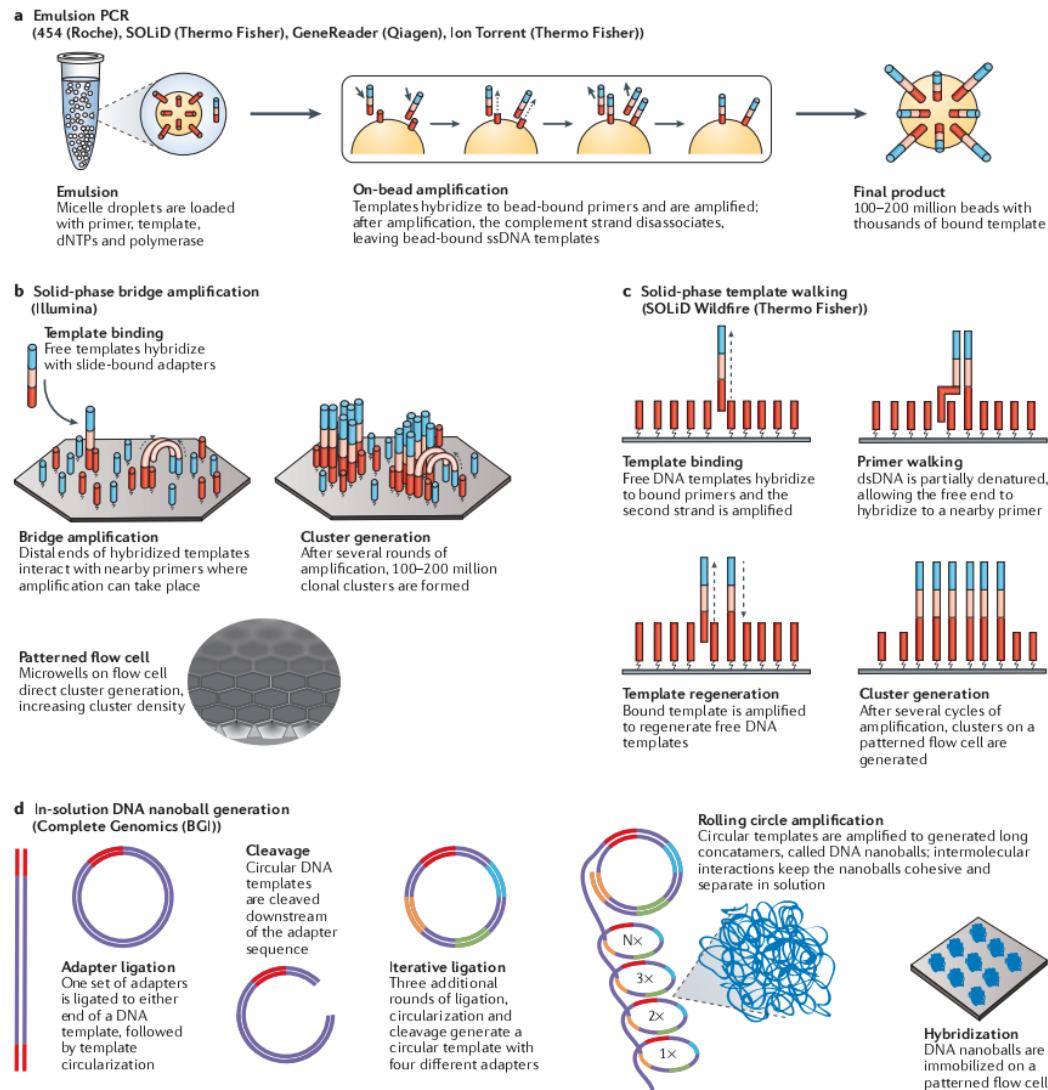


Figure 1.15 – Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS d'après [@Goodwin2016] : **a** : PCR en émulsion. **b** : amplification par pont. **c** : Amplification par modèle mobile. **d** :

La réaction de séquence

La réaction de séquence est l'étape suivant l'amplification et consiste à déterminer l'ordre dans lequel se succèdent les nucléotides de l'ensemble des clones générés dans la phase d'amplification. Il existe deux technologies principales permettant le séquençage de *reads* courts :

1. Séquençage par synthèse (SBS) : Ce type de séquençage regroupe l'ensemble

des méthodes utilisant l'ADN polymérase pour synthétiser de l'ADN. En 2016, Sahra Goodwin et ses collègues ont différenciées deux catégories de séquençage par synthèse (Goodwin, McPherson, & McCombie, 2016) :

- Terminaison par cycle réversible, cyclic reversible termination (CRT)** (**Figure : 1.16**) : Cette méthode est caractérisée par l'utilisation de molécules terminatrices auxquelles le groupement 3' – OH est modifié de sorte à éviter l'elongation (J. Guo et al., 2008), on parlera de groupement 3' – bloqué. Une amorce liée au fragment d'ADN permettra l'initialisation du processus de polymérisation. À chaque cycle, un mix comprenant l'ensemble des quatre désoxynucléotides (dNTPs), préalablement labélisés par un fluorophore 3' – bloqué, est mis en contact du fragment. Après l'incorporation d'un unique dNTP au fragment, les dNTPs non liés sont éliminés et la nature du dNTP ajouté est identifiée grâce à son fluorophore. Le fluorophore et le groupement 3' – bloqué sont retirés permettant ainsi à un nouveau cycle de commencer.

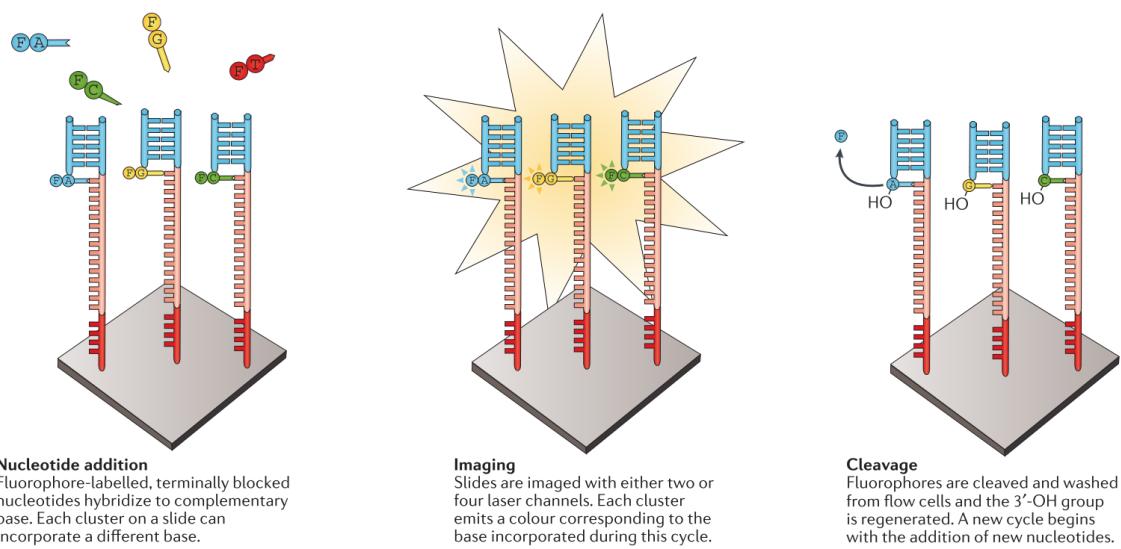


Figure 1.16 – Exemple de séquençage CRT tel qu'il est effectué par Illumina d'après [@Goodwin2016] : **a** : ajout d'un dNTP labellisé par un fluorophore et 3'-bloqué. **b** : identification du dNTP ajouté grâce au fluorophore. **c** : le fluorophore est clivé du dNTP et le groupement 3'-OH est reformé à partir du groupement 3'-bloqué permettant ainsi l'elongation

- Addition de nucléotide unique, single nucleotide addition (SNA)** (**Figure : 1.17**) : L'initialisation de la méthode SNA est identique à celle de la méthode CRT. La différence se fait donc au moment de la phase d'elongation. Contrairement à la méthode CRT, le mix contenant les dNTPs ne contient qu'un seul type de dNTP. Quatre mixs différents sont donc présentés successivement au fragment d'ADN à séquencer, ceux-ci se fixeront uniquement s'ils sont complémentaires à

la séquence. Ces dNTPs n'ont donc pas besoin d'être 3' – bloqué puisqu'un seul dNTP est ajouté à chaque itération. Après avoir présenté un mix, on vérifie si un dNTP s'est lié au fragment. Lors des séquences homopolymériques (plusieurs nucléotides identiques successifs dans la séquence), plusieurs dNTPs sont donc liés simultanément, cela sera détecté car le signal émis est proportionnel au nombre de nucléotides ajoutés.

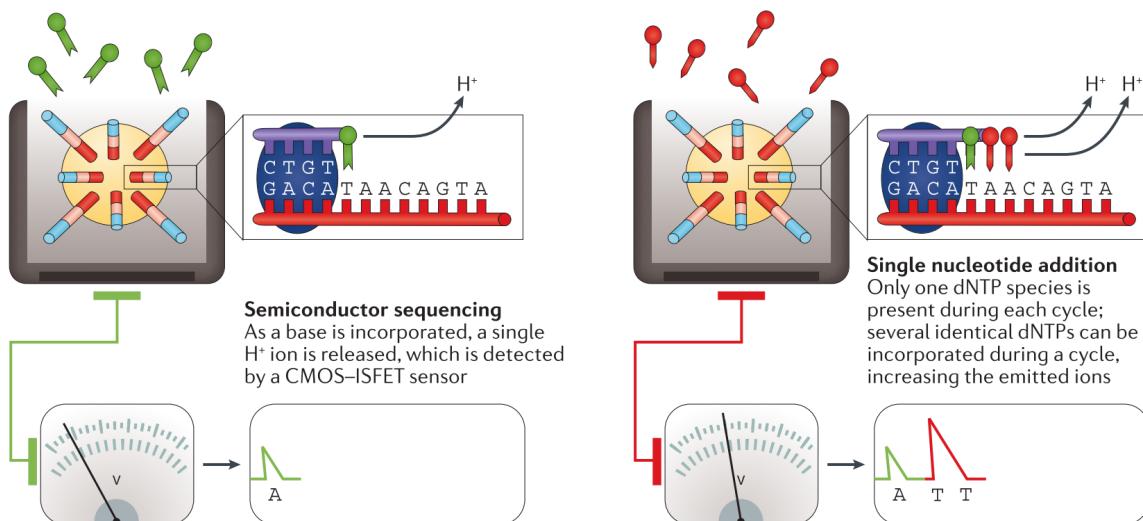


Figure 1.17 – Exemple de séquençage SNA tel qu'il est effectué par Ion Torrent d'après [Goodwin2016] : **a** : Mise en présence du patron d'ADN à séquencer avec un mix contenant un seul type de dNTP, si le dNTP est complémentaire au patron, il se fixe et libère un proton permettant d'identifier la liaison. **b** : Dans le cas d'homopolymère, autant de proton sont relâchés de bases constituant l'homopolymère, le signal émit est donc plus fort permettant d'identifier le nombre des dNTPs liés

2. **Séquençage par ligation (SBL)** : Par définition, cette méthode est basée sur l'hybridation et la ligation de l'ADN à une sonde liée à un fluorophore (Tomkinson, Vijayakumar, Pascal, & Ellenberger, 2006). Ce processus utilise les caractéristiques de la ligase, une enzyme qui a pour fonction de catalyser la liaison de deux brins d'ADNs par des liaisons phosphodiester. La sonde est constituée d'une ou deux bases connues, on parle alors de *one-base-encoded probes* ou de *two-bases-encoded probes* suivis d'une succession de bases "dégénérées" ou universelle, c'est à dire, des bases capables de s'apparier avec n'importe laquelle des quatre bases de l'ADN.

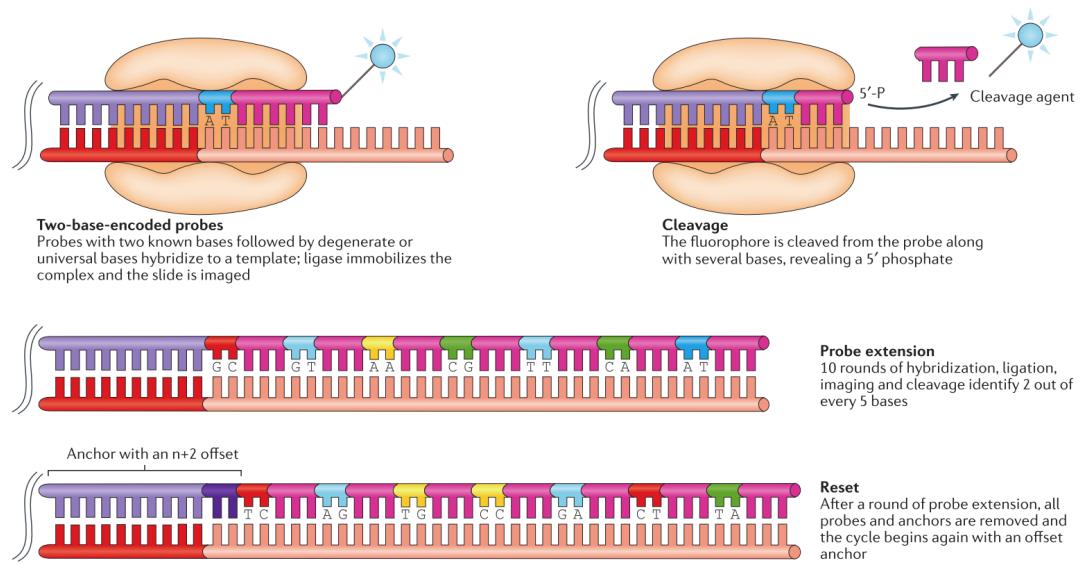


Figure 1.18 – Exemple de séquençage SBL tel qu'il est effectué par SOLiD d'après [Goodwin2016] :

1.5 L'analyse bioinformatique des données de NGS

La stratégie consistant à séquencer en parallèle plusieurs milliers de *reads* courts a engendré plusieurs nouveaux défis bioinformatique dans l'analyse et l'interprétation des données de séquençage et la recherche de variants dans le génome humain (Wold & Myers, 2007, M. Q. Yang et al. (2009)). Ces techniques ont été appliquées dans différents contextes, notamment la métagénomique (J. Qin et al., 2010), la détection de SNPs (Van Tassell et al., 2008) et de variants structuraux (Alkan et al., 2010, Medvedev, Stanciu, & Brudno (2009)) mais également dans des études portant sur la méthylation de l'ADN (K. H. Taylor et al., 2007), l'analyse de l'expression des ARNs messagers (Sultan et al., 2008), dans la génétique du cancer (Guffanti et al., 2009) et la médecine personnalisée (Auffray, Chen, & Hood, 2009). Cependant, pour l'ensemble de ces applications, la grande quantité de données générées par chaque analyse pose plusieurs défis informatiques (Horner et al., 2009). En effet, les progrès techniques des dernières décennies ont rendu possible le séquençage de plusieurs millions de *reads* d'ADN en un temps relativement court et à coûts raisonnable. Ainsi, l'émergence du séquençage haut débit et notamment du WGS et du WES a permis de réunir une quantité jusqu'à présent inégalée d'informations sur les variations génétiques, et sur les gènes et leurs fonctions (E. R. Mardis, 2008, Bentley (2006)). Cependant, de par leur nature et leur quantité, l'acquisition de ces nouvelles données a engendrée de nouvelles problématiques qui freinent les biologistes dans leurs recherches.

1.5.1 Les données fournies par le NGS

Un *read* c'est quoi ?

Après la phase d'amplification, chaque clone est analysé puis, la séquence composant chacun de ce clone est déterminée. La taille de cette séquence varie en fonction des plateformes de séquençage mais est généralement comprise entre 40 et 150 pb pour le NGS (**Figure : 1.14**). Depuis quelques années, un nouveau type de *read* est apparu, le *read paired-end*. Contrairement au *reads* classiques (single-end), les deux extrémités (les *ends*) du fragment d'ADN sont désormais séquencées. La distance séparant les deux extrémités du *read* étant connue, cela permet aux aligneurs d'utiliser cette information afin d'améliorer leur précision, notamment dans les zones répétées (H. Li et al., 2008). En plus de SNP, ce format permet de mettre en évidence des variants structuraux (Korbel et al., 2009).

Le format FASTQ

Le format FASTQ (**Figure : 1.19**) est actuellement le format de donnée le plus couramment utilisé dans le cadre du séquençage haut-débit. Sa création est cependant antérieure à l'émergence du NGS puisqu'il fut inventé à la fin du XX^{ième} par Jim Mullikin au Wellcome Trust Sanger Institute alors que le séquençage commençait à prendre de l'ampleur grâce à des projets tels que le Projet Génome Humain. La quantité de données générées par ces programmes a nécessité une analyse automatisée, c'est ainsi que chaque base séquencée s'est vue associer un score de qualité appelé *Phred-score*. Chaque séquence générait ainsi deux fichiers, un fichier FASTA contenant les séquences et un fichier QUAL contenant les scores *Phred* associés à chaque base du fichier FASTA Cock2009. Plus tard, afin de n'avoir à manipuler qu'un seul fichier, les fichiers FASTA et QUAL furent fusionnés en ce que l'on appelle désormais le fichier FASTQ. Ce format est aujourd'hui le plus utilisé par les différents séquenceurs on peut cependant noter certaines différences dans les formats FASTQ provenant des différentes plateformes puisqu'à l'époque, aucune spécification officielle n'avait été donnée (Cock, Fields, Goto, Heuer, & Rice, 2009).

```

@HC9D00P01AN1VB rank=0000246 x=156.0 y=3301.0 length=309
ACACATACGCACTGGCGTAAAGGGCGCGAGCGCGTCAGAGCGTCGGCTCAAAGTCCACCGCTTAACGGTGGAGGCCTG
+HC9D00P01AN1VB
FFFFFFFFFFFGD554A6911144442AAABDFFIIIIIIIIIIIIIIHHHFFFFFFFA@CFDFDFDFC???CFFFFFFFI
@HC9D00P01AWYAE rank=0000402 x=258.0 y=772.0 length=373
ACACATACGCACTGGGCATAAAAGGGCACGTAGGCCGATTGTAAGTCAGGGGTGAATCCGGGCGTCAACCTCGGAACTGCCT
+HC9D00P01AWYAE
IIIIIIIIIIIIIIHHHII;666MHIIIIIIIIIIICCIIIIFDC2//.<-//93.....---9?CCCCFEECCCCIIIIDI
@HC9D00P01A3C8R rank=0000675 x=331.0 y=1081.0 length=373
ACACATACGCACTGGGTAAAGGGTCGTAGGCCGCTTAAAGTCAGGGGTGAATCCTGGAGCTCAACTCCAGAACTGCCT
+HC9D00P01A3C8R
IIIIIIIIIIII3//...---4AIIIECCE466GH974EEIAC@.0004.000>9@CEEEIIIIIIIIIIIIIIIIHHI
@HC9D00P01AW8TJ rank=0000926 x=261.5 y=2133.0 length=373
ACACATACGCACTGGGTAAAGCGCACGTAGGCCGATTGCTAAGTCAGGGGTGAATCCTGGAGCTCAACTCCAGAACTGCCT
+HC9D00P01AW8TJ
IIIIIIHHIIIIHHHII;;IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@HC9D00P01AU18Y rank=0000952 x=230.0 y=2656.0 length=372
ACACATACGCACTGGCATAAAAGAGCGCGTAGGCCGCTTGTAGTCAGGTGAAGGCCCTGGCTTAACCCGGGAAGCGCGC
+HC9D00P01AU18Y
IIIIIIIIHHHIIII;;IIIIIIIIIIIIIIIIIIIIIIIIIIIIII?666DHHHIHFEIIIIIC;555994?FIGI
@HC9D00P01AU1W rank=0000977 x=228.0 y=226.0 length=372

```

Figure 1.19 – présentation d'un fichier FASTQ (FIGURE A CHANGER) : **a** : identifiant du *read*. **b** : séquence du *read*. **c** : score de qualité associé

1.5.2 L'alignement

L'alignement constitue la première étape de l'analyse des données de NGS lorsqu'un génome de référence est disponible. L'objectif de l'alignement est de déterminer la position correcte de chacun des *reads* séquencés le long du génome de référence. Cette référence est souvent construite à partir des données de séquençage de plusieurs donneurs et ne représente donc pas la séquence d'un individu en particulier mais est

censé représenter la séquence consensus d'une espèce donnée. Par exemple, la séquence de référence humaine GRCh37 (*Genome Reference Consortium human build 37*) a été créé à partir de 13 volontaires anonymes New-Yorkais. Dès lors, cette référence servira de patron aux aligneurs afin qu'ils replacent correctement les différents *reads* des individus séquencés. Cette étape peut être comparée à la reconstruction d'un puzzle dans laquelle les *reads* seraient les pièces et le génome de référence le modèle. Elle constitue probablement l'étape la plus importante de l'analyse des données issues du séquençage haut débit (Flicek & Birney, 2009) car elle est la base sur laquelle repose l'ensemble des étapes effectuées en aval, notamment l'appel des variants (R. Nielsen, Paul, Albrechtsen, & Song, 2011). Cependant, l'étape d'alignement est sujette à de nombreuses erreurs dont certaines proviennent directement des erreurs survenues lors de l'étape de séquençage, d'autres, sont dues aux caractéristiques des régions séquencées comme par exemple les séquences répétées (Ben Langmead & Salzberg, 2012) qui pourront entraîner l'alignement d'un même *read* à plusieurs régions du génome (Treangen & Salzberg, 2013). De nombreux aligneurs ont émergés afin de répondre au mieux à cette problématique tel que Bowtie (B Langmead, Trapnell, Pop, & Salzberg, 2009), Bowtie2 (Ben Langmead & Salzberg, 2012), BWA, NovoAlign, MAGIC (Su et al., 2014). De nombreuses études ont cependant montré de grandes différences entre ces aligneurs, au niveau du temps de calcul, de leur coût en mémoire et de leur taux d'erreur (Ruffalo, Laframboise, & Koyutürk, 2011, Thankaswamy-Kosalai, Sen, & Nookaew (2017), S. Bao et al. (2011)).

1.5.3 L'appel des variants

L'appel des variants, ou *variant calling*, fait référence à l'ensemble des méthodes permettant d'identifier des SNVs ou des indels à partir des résultats de l'alignement. Cette étape est souvent différenciée de l'alignement, cependant, les résultats de l'appel étant extrêmement dépendants de l'alignement, il est conseillé d'effectuer son appel en tenant compte de l'aligneur choisi (R. Nielsen et al., 2011, M. A. DePristo et al. (2011), Lunter & Goodson (2011)). On appellera variant toute différence de séquence observée entre un individu et la séquence de référence utilisée. Pour reprendre la comparaison avec la construction d'un puzzle, cette étape consiste à détecter quelles sont les pièces qui présentent des différences avec le modèle. De nombreux logiciels d'appels des variants, ou *caller*, basés sur des algorithmes différents ont émergés ces dernières années pour répondre à cette problématique. Parmi les plus connus on note SAMtools (H. Li et al., 2009), Genome Analysis Tool Kit - HaplotypeCaller (GATK-HC) (McKenna et al., 2010), Freebayes, SOAPindel et TVC. Les quatre premiers cités, peuvent être utilisés pour analyser des données provenant de tout type de plateforme de séquençage contrairement à TVC qui a été développé spécifiquement pour les données provenant de Ion Proton. Les données issues de NGS peuvent présenter un taux d'erreur important. Ce taux d'erreur est multifactoriel et inclus notamment les erreurs de l'alignement. L'un des éléments clef à prendre en compte pour pouvoir effectuer un appel de qualité est la couverture de la position appelée (D. Sims et al.,

2014). Cependant, malgré la prise en compte de cet élément, l'appel de variants reste un processus difficile souvent lié à plusieurs erreurs. Plusieurs de ces erreurs sont même directement liées à la plateforme de séquençage utilisée en amont, et les différents logiciels ne présentent pas les mêmes performances en fonction de ces différentes plateformes (Hwang, Kim, Lee, & Marcotte, 2015), c'est pourquoi il convient d'adapter le logiciel d'appel en fonction de la plateforme de séquençage utilisée préalablement. Les erreurs d'appel sont généralement classées en deux catégories principales et certains aligneurs auront tendance à être plus sujets à l'un de ces types d'erreur qu'à l'autre (**Figure : 1.20**) :

1. Oubli de l'allèle de référence (**IR**, *ignore the reference allele*) : représente un variant appelé homozygote correspondant en réalité à un variant hétérozygote composé de l'allèle de référence et d'un allèle variant.
2. Ajout de l'allèle de référence (**AR**, *adding the reference allele*) : représente un variant appelé hétérozygote composé de l'allèle de référence et d'un allèle variant correspondant en réalité à un variant homozygote composé de deux allèles variants.

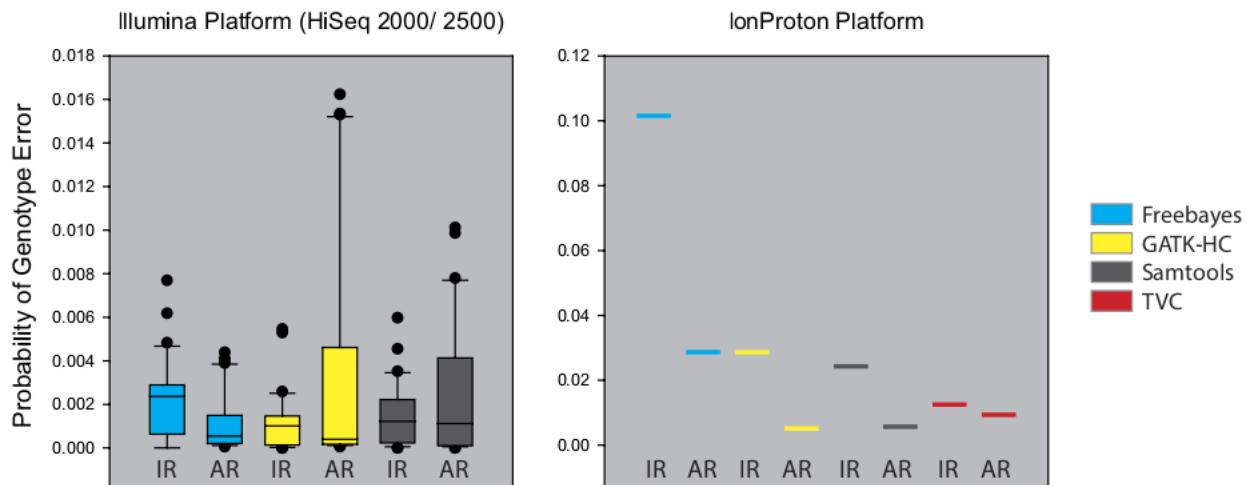


Figure 1.20 – Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel d'après [Hwang2015] : Pour la plateforme Illumina, on peut voir que Freebayes favorise les appels variant-homozygote tandis que GATK-HC et Samtools favorisent les appels hétérozygotes. Pour la plateforme Ion Proton, les 4 logiciels entraînent des erreurs de type IR

De même que pour l'aligneur, le choix du logiciel d'appel est crucial car il existe de nombreuses différences dans les variants appelés par différents logiciels se basant sur les mêmes données brutes (Baes et al., 2014, O'Rawe et al. (2013), Rosenfeld,

Mason, Smith, Wallin, & Diekhans (2012)). En effet, en 2013, une étude comparant les résultats de 5 *callers* montraient que seulement 57,4% des variants étaient appelés par les 5 *callers* et que 80,7% des variants étaient appelés par au moins 3 d'entre eux. Ce taux chutait drastiquement pour les indels puisque la concordance était cette fois seulement de 26,8% pour les indels non retrouvés par les 3 *callers* (O’Rawe et al., 2013). Ces résultats sont cependant à pondérer avec une étude de 2015 comparant 4 *callers* et montrant que 91,7% des SNVs séquencés sur une plateforme Illumina étaient appelés par 3 *callers*, cependant, pour les variants séquencés sur Ion Proton, seulement 27,3% des variants étaient appelés par au moins 3 *callers* et 57,4% des variants n'étaient appelés que par un seul des *callers* (Hwang et al., 2015).

1.5.4 L'annotation des variants, filtrage et priorisation

Traditionnellement, les scientifiques et les laboratoires dans lesquels ils travaillaient développaient leur expertise dans un nombre de pathologies et de gènes associés limité. L'émergence du NGS est en train de remettre en cause cette pratique car la totalité de l'exome ou du génome peut permettre de couvrir tous les gènes quelque en une seule et même analyse. De nombreux praticiens maintiennent cependant une spécialisation pour certains groupes de pathologies qui est précieuse pour l'analyse des données et l'obtention d'un diagnostic. En effet il est courant de retrouver entre 20.000 et 25.000 variants différents par exome (Gonzaga-Jauregui, Lupski, & Gibbs, 2012). Afin de pouvoir lier un variant à une pathologie, il est indispensable d'annoter cet ensemble de variants, c'est à dire d'associer à ces variants l'ensemble des informations qui les caractérisent afin de pouvoir les replacer dans leur contexte biologique. Ces informations serviront ensuite d'indicateur afin filtrer ou prioriser un variant. Cette dernière étape de l'analyse est elle aussi cruciale puisqu'elle permet de réduire le nombre de variants à considérer. On peut généralement distinguer deux niveaux d'annotations d'un variant :

1. **Au niveau du variant** : Ce niveau d'annotation regroupe l'ensemble des informations spécifiques à un variant
 - a. **Informations issues des résultats du séquençage** : la couverture du variant ainsi que la qualité qui lui est associée peuvent permettre de considérer un variant comme étant fiable ou non. Le génotype associé à ce variant est également une information importante.
 - b. **La fréquence du variant dans la population générale** : l'émergence du séquençage haut-débit a permis de de gros consortium tel que ESP6500 [CITATION], 1KG [CITATION]. Ces consortiums ont pu mettre à disposition du public de données de séquençage exomique de 6503 individus pour ESP et de 2504 pour la phase 3 du 1000Genomes. On peut également noter l'*Exome Aggregate Consortium* (ExAC) (Lek et al., 2016) qui n'a

effectué aucun séquençage mais qui a regroupé les données de plusieurs gros jeux (notamment 1000Genome et ESP) afin de leur appliquer la même analyse bioinformatique harmonisant ainsi les données provenant de 60.706 individus non apparentés. Cette masse d'information permet de se faire une idée de la fréquence d'un variant dans la population générale et même au sein de sous population humaine. On considère qu'un variant fréquent ne peut pas être délétère, sinon il aurait été contre-sélectionné au cours de l'évolution.

- c. **Son impact sur le transcrit** : Dans la plupart des analyses phénotype-génotype, les chercheurs se limitent aux variants chevauchant des transcrits codants pour une protéine. Il est donc important de savoir l'impact d'un variant sur ce transcrit, c'est à dire si le variant va causer une mutation synonyme, un faux-sens ou une mutation tronquante. Des logiciels tel que *Variant Effect Predictor* (VEP) (W. McLaren et al., 2016), SnpEff (Cingolani et al., 2012) ou encore ANNOVAR [@] vont prédire l'impact qu'aura un variant sur les différents transcrits qu'il chevauche. D'autres logiciels tel que SIFT (P. Kumar, Henikoff, & Ng, 2009), PROVEAN (Y. Choi, Sims, Murphy, Miller, & Chan, 2012), Polyphen2, ou encore CADD vont eux chercher à prédire la pathogénicité de ce variant, c'est à dire la probabilité que ce variant soit délétère pour la fonction de la protéine. Bien que cette information soit importante, elle est à pondérer étant donné le peu de concordance qu'il existe entre les prédictions de ces différents logiciels (**Figure : 1.21**).

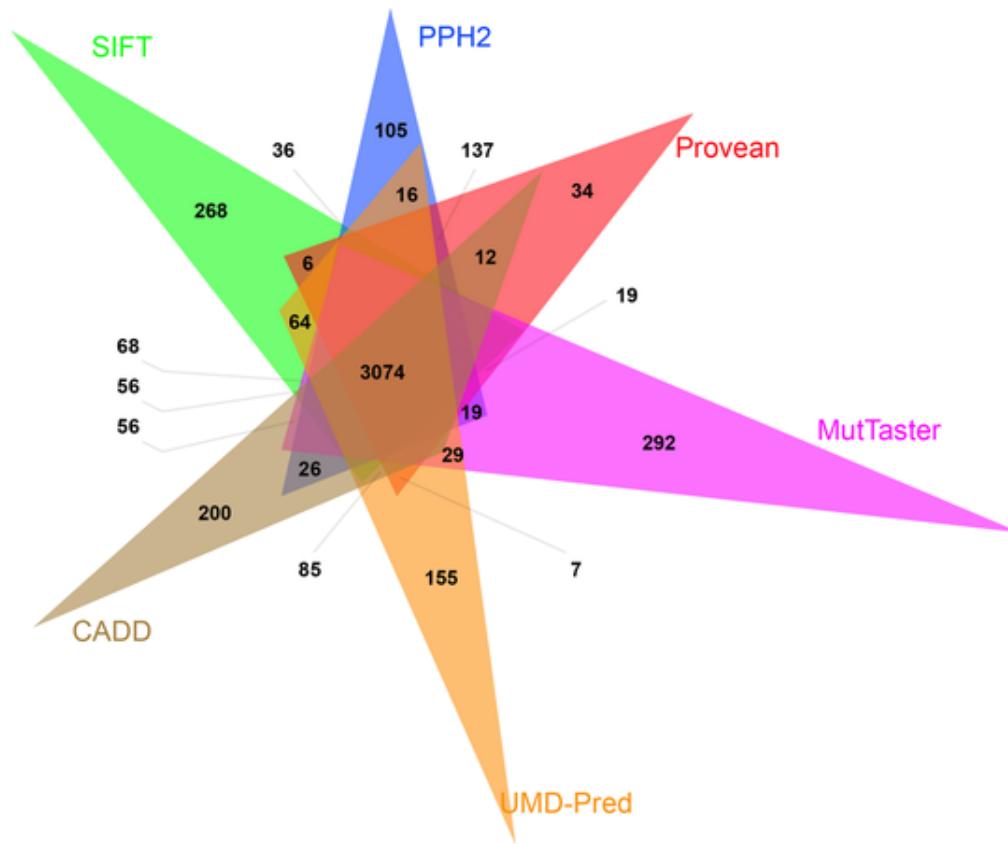


Figure 1.21 – Diagramme de Venn des prédictions de pathogénicités de six logiciels d'après [Salgado2016] :

2. **Au niveau de l'unité génétique** : DÉCRIRE UNITÉ GÉNÉTIQUE (gène, transcript). L'annotation au niveau de l'unité génétique consiste à récupérer l'ensemble des informations disponibles non plus sur le variant uniquement mais sur la ou les unités génétiques qu'il impacte. Ce “dézoom” permet d'ajouter des informations complémentaires particulièrement utiles notamment lorsque peu d'informations sont disponibles sur le variant lui-même. En pratique, la plupart des variants connus pour impliquer une pathologie sont des variants privés, c'est à dire spécifiques à une famille ou un individu limitant ainsi la quantité d'information disponible sur ce variant. Élargir l'annotation au niveau des unités génétiques impactées par des variants permet d'augmenter considérablement la quantité d'information disponible et permet donc d'améliorer la capacité des algorithmes à filtrer et / ou prioriser les variants rendant donc les analyses plus efficaces. On peut relever certains logiciels tel que le *Protein A Nalysis THrough Evolutionary Relationships* (PANTHER) (Mi et al., 2017) qui permet par exemple de classer une liste de gènes en fonction de leurs fonctions moléculaires, des processus biologiques et des voies de signalisation dans lesquels ils sont impliqués. On peut également noter *the Human Phenotype Ontology project* (HPO) (Köhler et al., 2014) qui fournit une classification (À compléter). Plus récemment, on a pu voir émerger des “scores mutationnel” tel que RVIS (Petrovski et al., 2013)

ou encore le pLI (Lek et al., 2016). En se basant sur les bases de données telle que ESP ou encore ExAC, ces scores permettent de classer les gènes en fonction de leur tolérance (ou intolérance) aux variations avec l'idée sous-jacente que "les gènes impliqués dans des pathologies à transmission Mendéliennes" devraient être moins tolérants aux variations que les autres.

Comme nous l'avons vu, le développement d'outils permettant l'analyse des données NGS est extrêmement important puisqu'il permet aux biologistes de faire face à la masse de données générées par le séquençage haute débit l'aidant ainsi dans ses prises de décisions. Il est à noter que la plupart de ces données filtrées sont extrêmement dépendantes du jeu de gènes utilisés, les prédictions seront donc différentes si l'on se base sur les gènes RefSeq, Ensembl ou UCSC (D. J. McCarthy et al., 2014, S. Zhao & Zhang (2015)) bien que les gènes du *Consensus Coding Sequence project* (CCDS) soient bien représentés par ces trois listes (K. D. Pruitt et al., 2009). De même, pour une même liste de gène, de nombreuses différences seront observées en fonction du ou des logiciels de prédition utilisés (D. J. McCarthy et al., 2014, Salgado, Bellgard, Desvignes, & Bourd (2016)).

1.5.5 Conclusion NGS

En moins de 10 ans, les technologies NGS sont passées du séquençage de panels de gènes (environ 100 Mb pour le Roche GS FLX system) au séquençage de génomes entiers (environ 1500 GB pour l'Illumina Hiseq 4000) et d'une utilisation exclusive à la recherche à l'analyse en routine dans un cadre de diagnostics cliniques. Le nombre croissant d'études utilisant le WGS ou le WES démontre le pouvoir de ces approches dans des analyses phénotypes-génotypes impliquant des pathologies à transmission Mendélienne. De plus, la diminution constante des coûts par génomes / exomes séquencés laisse supposer que ces technologies deviendront d'ici peut le fer de lance de la génétique clinique moderne. Cependant, cette quantité de données produites crées de nouvelles problématiques pour les généticiens qui se retrouvent désormais face au "déluge de données génétiques" (Schatz & Langmead, 2013). Le succès d'une étude n'étant plus lié aux capacités de séquençage mais aux compétences dans l'analyse et l'interprétation des données produites. Bien que de nombreux efforts soient faits pour palier la contrainte instaurée par les *reads* courts dans le cadre d'analyse génomique, les solutions informatiques et bioinformatiques proposées jusqu'à présent restent en dessous des besoins créés par NGS (J. D. McPherson, 2009). Cette masse de données produite, à l'origine du succès du séquençage haut-débit dans le domaine de la génomique et de la post-génomique, se trouve désormais être un frein à la compréhension et l'interprétation des réseaux de gènes et leurs implications dans des pathologies. La limitation de cette technologie n'est donc plus le séquençage d'un, de plusieurs, ou de l'ensemble des gènes, mais plutôt l'analyse et l'interprétation des données générées. Le processus allant de l'extraction de l'ADN à l'identification d'un variant responsable d'une pathologie comprend de nombreuses étapes apportant avec elles leur lot d'erreurs. Bien que dans chacune de ces phases, de nombreux acteurs

soient en concurrence et cherchent à atteindre une solution idéale, celle-ci n'a toujours pas été trouvée et la prolifération des logiciels et algorithmes d'analyses, bien que nécessaire, peut également parfois augmenter la confusion.

Malgré les dizaines de milliers d'exomes et de génomes ayant été jusqu'à présent étudiés, notre compréhension des mécanismes moléculaires qui sous-tendent la variété génomique humaine reste limitée, et ce particulièrement dans le contexte de l'analyse de pathologies génétiques. En effet, à l'heure actuelle, plus de 3700 pathologie à transmission Mendélienne ont été caractérisées mais un nombre similaire ont toujours une cause inconnue (Amberger, Bocchini, & Hamosh, 2011). L'élucidation de ces mystères passera probablement par une harmonisation des méthodes de production des données ainsi que par l'amélioration des techniques d'analyses.

Chapitre 2

Investigation génétique et physiologique de la globozoospermie

2.1 Introduction sur la globozoospermie

Comme expliqué précédemment, La globozoospermie est phénotype rare (< 0.1% des patients infertiles) mais néanmoins sévère (C. G. S. Sen et al., 1971) de teratozoospermie menant à l'infertilité masculine. Cette anomalie est caractérisée par la présence de spermatozoïdes présentant une tête ronde dépourvue d'acrosome et d'une pièce intermédiaire désorganisée dans l'éjaculat (Singh, n.d., Pedersen & Rebbe (1974)) (**Figure : 2.1**). En plus des anomalies morphologiques, les spermatozoïdes globozoocéphales présentent également des désorganisations au niveau moléculaire. Par exemple, le facteur spermatique PLC ζ requit pour l'activation ovocitaire, est absent ou en quantité infime dans les spermatozoïdes globozoocéphales (Heytens et al., 2009, S. Taylor et al. (2010), S.-Y. Yoon et al. (2008)) compromettant ainsi l'activation ovocytaire et expliquant le faible taux de fécondation observés en IVF (*in vitro* fertilization) et en ICSI (intra cytoplasmic sperm injection) (A. Dam et al., 2006). On distingue la globozoospermie totale avec 100% des spermatozoïdes présentant le phénotype ou partielle en fonction du taux de spermatozoïdes atteints. Bien que l'infertilité masculine soit souvent la résultante de plusieurs facteurs, les premières études présentant des patients atteints par un phénotype complet (C. G. S. Sen et al., 1971) suggéraient que la globozoospermie était une exception. De plus les caractéristiques morphologiques très typiques des spermatozoïdes laissaient penser à une cause monogénique. En 2007, une étude portant sur une famille juive ashkénaze comprenant six frères dont trois atteints a pu lier ce phénotype à une mutation homozygotes sur le gène *SPATA16* présente chez les trois frères atteint (A. H. D. M. Dam et al., 2007). Cependant, dans la même étude, 29 autres patients présentant le même phénotype ont été analysé, et pour ceux-ci, aucun variant du gène *SPATA16* n'a pu être lié au

phénotype (A. H. D. M. Dam et al., 2007) indiquant clairement que les mutations de ce gène n'étaient pas les seules responsables. En 2011, une autre étude portant sur une cohorte de 20 patients Tunisiens a pu mettre en évidence une délétion homozygote de 200kb emportant la totalité du gène *DPY19L2* chez 15 des 20 patients analysés (Harbuz et al., 2011). Les études effectuées ultérieurement sur ce phénotype ont ensuite pu montrer que les altérations du gène *DPY19L2*, et notamment cette délétion, étaient responsables de la majorité des cas de globozoospermie (Ray & Arnoult, 2011, ElInati et al. (2012)).

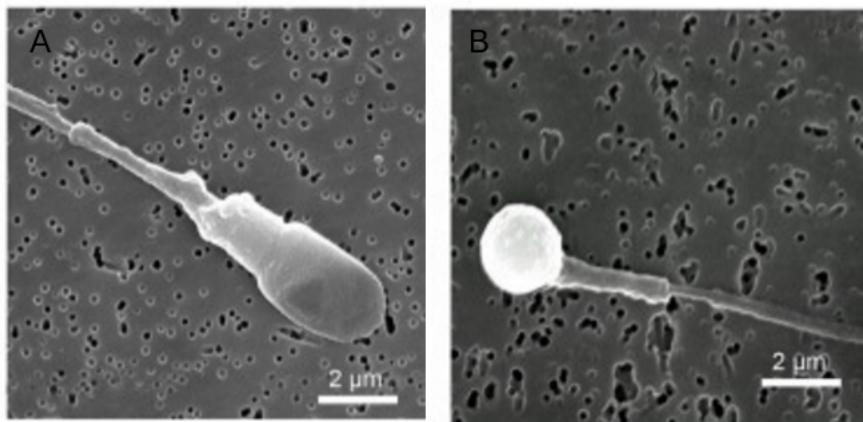


Figure 2.1 – Observation au microscope à balayage d'un spermatozoïde normal (**A**) et d'un spermatozoïde globozoocéphale (**B**) (changer les photos avec celles sur lesquelles on voit l'acrosome colorés)
adapté d'après [@Harbuz2011]

En 2012, le développement d'un modèle murin KO *Dpy19l2*^{-/-} a permis de mieux comprendre les mécanismes moléculaires impliqués dans la globozoospermie causée par la délétion du gène *DPY19L2* chez l'humain (V. Pierre et al., 2012). Tout d'abord car ce modèle de souris KO présentait les mêmes caractéristiques que les patients humains. Tout d'abord, ces souris étaient infertiles et présentaient des spermatozoïdes globozoocéphales (**Figure : 2.2**) mais aussi et surtout, l'ensemble des autres dysfonctionnements étaient retrouvés, c'est à dire : l'absence de l'acrosome, les défauts morphologiques du noyau, de l'enveloppe nucléaire et de l'acoplaxome ainsi que le mauvais positionnement de la manchette (V. Pierre et al., 2012). Ainsi il a pu être démontré que la protéine Dpy19l2 étaient principalement exprimé dans le spermatides et plus spécifiquement dans la membrane nucléaire interne faisant face à la vésicule acrosomale et que l'absence de cette protéine entraînait la déstabilisation à la fois de la lamina nucléaire, de la jonction entre l'acoplaxome et l'enveloppe nucléaire (V. Pierre et al., 2012).

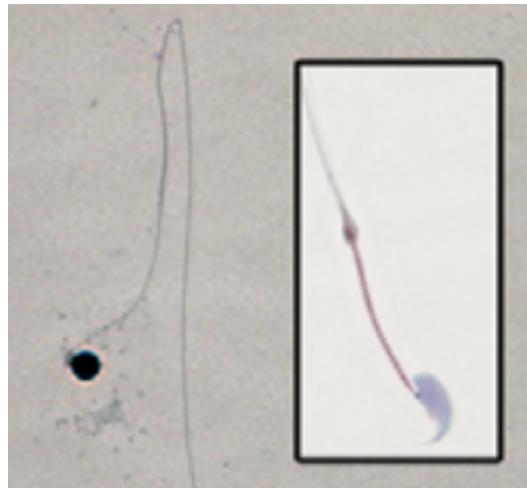


Figure 2.2 – Comparaison entre les spermatozoïdes des souris $*\text{Dpy19l2}^{*--}$ (à gauche) et les souris sauvages $*\text{Dpy19l2}^{*++}$ (à droite) d'après [@Pierre2012]

Dans ce chapitre je détaillerai les résultats de trois articles dont je suis co-auteur :

1. **Fine Characterisation of a Recombination Hotspot at the DPY19L2 Locus and Resolution of the Paradoxical Excess of Duplications over Deletions in the General Population** : au cours de cette étude j'ai participé à divers manipulation de biologie moléculaire tel que l'extraction d'ADN spermatique, quantification des délétions / duplications *de novo*. De même, j'ai pu contribuer au divers analyses statistiques.
2. **Comparative testicular transcriptome of wild type and globozoospermic Dpy19l2 knock out mice** : Dans cette étude j'ai pu effectuer l'intégralité des manipulation de biomoléculaire (extraction de l'ARN testiculaire de souris et analyse sur puce) de même que l'intégralité de l'analyse bioinformatique des résultats.

2.2 Résultats

2.2.1 Les mécanismes mutationnels

Confirmation de l'excès de délétion

Chez les mammifères il existe trois paralogues de *DPY19L2* de fonction encore inconnue et un pseudogène présentant une très forte homologie de séquence (> 95%) (Carson, Cheung, & Scherer, 2006). Chez l'Homme, ce gène est flanqué de deux séquences présentant une forte homologie (>95%) d'une taille de 28kb (**Figure : 2.3 - 1**). Ces séquences appelées LCRs (*low copy repeats*) représentent une large portion du génome humain (Cheung et al., 2003, Bailey et al. (2002)) et vont, de par leur homologie favoriser les duplications de gènes jouant ainsi un rôle important dans l'évolution des génomes des vertébrés (Walsh, 2003, Ohno (1970)). Dans le cas de *DPY19L2*, ces LCRs vont, au cours de la méiose entraîner la venue de recombinaison homologues non-allélique (NAHR) donnant lieu soit à une délétion du gène *DPY19L2* et la formation d'un ADN circulaire comprenant le gène (**Figure : 2.3 - 2**) soit à un allèle possédant deux copies du gène tandis que l'autre n'en possède aucune (**Figure : 2.3 - 3**).

Ainsi, le mécanisme de NAHR devrait, en théorie, engendrer la formation de plus d'allèles délétés que d'allèles dupliqués puisque les cas présentés en figures 2.3 - 2 et 2.3 - 3 induisent la formation d'un allèle délété tandis que seul le cas 2.3 - 3 forme un allèle dupliqué. Cependant, les données mises à disposition par la base de donnée *Database of Genomic Variants* (DGV) (MacDonald, Ziman, Yuen, Feuk, & Scherer, 2014) indique un excès de duplication puisque sur un total de 6575 individus analysés, 83 duplications et de 26 délétions hétérozygotes ont été observées pour le locus de *DPY19L2*. Afin de confirmer ces résultats et ainsi écarter l'hypothèse d'un biais causé par la présence du pseudogène *DPY19L2P1* très homologue avec *DPY19L2* (Carson et al., 2006) notre équipe a procédé à la ré-analyse du locus de *DPY19L2* de 1699 individus par CGHarray. Un total de 15 duplication et de 3 délétions hétérozygotes furent observés corroborant les données fournies par DGV et confirmant ainsi l'excès de l'allèle dupliqué dans la population générale.

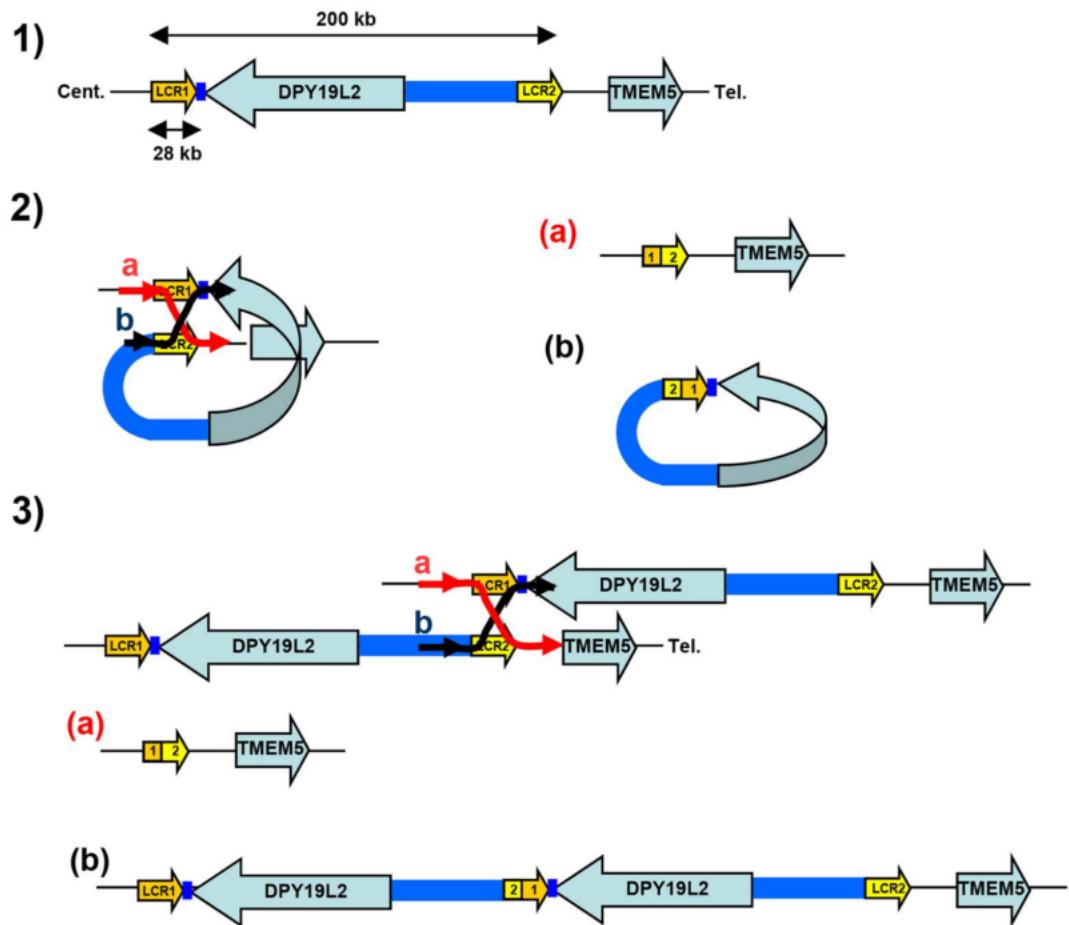


Figure 2.3 – Représentation schématique du mécanisme de NAHR causé par les séquences LCR flanquant le gène *DPY19L2* : **1** : Le gène *DPY19L2* est entouré par les séquences LCR1 et LCR2 qui correspondent respectivement au LCR centromérique et télomérique. Ces LCRs sont séparés par environ 200kb et chacun d'eux mesurent approximativement 28kb. **2** : NAHR résultant du mauvais alignement des LCRs 1 et 2 du même chromatide entraînant la formation d'un allèle délété (a) et d'un ADN circulaire comprenant le gène *DPY19L2*. Cet ADN circulaire sera éliminé au cours du cycle cellulaire. **3** : NAHR résultant du mauvais alignement des LCRs 1 et 2 provenant de deux chromatides différents conduisant à la formation d'un allèle délété (a) et d'un allèle dupliqué (b)

Détermination du taux de recombinaison *de novo* de DPY19L2

Alors que les résultats précédents confirment un excès de l'allèle dupliqué de *DPY19L2* dans la population générale, nous avons par la suite cherché à déterminer les fréquences de duplications / délétions *de novo* de ce même locus. Ceci ayant pour but de déterminer si cet excès est dû à une sélection de l'allèle dupliqué ou au fait que celui-ci était effectivement produit plus fréquemment que l'allèle délété. Pour ce faire nous avons quantifié le taux d'apparition de ces événements génétiques à partir d'ADN spermatique. Les spermatozoïdes étant le produits direct de la méiose, ils sont donc les reflets d'haplotypes produits *de novo*. Pour cela, nous avons analysé par PCR digitale l'ADN spermatique de trois donneurs ainsi que l'ADN spermatique constitué d'un mix provenant de ces trois donneurs. Leur ADN a tout d'abord été dilué en série de sorte à ce qu'environ 25% des 96 puits de la PCR contiennent un événement (délétion ou duplication). Ainsi, en acceptant qu'un génome haploïde humain représente 3pg, 50ng d'ADN spermatique furent déposés dans chaque puit pour la PCR spécifique à la délétion, et 100ng dans chaque puit spécifiques à la duplication. Chaque puit contient donc une partie de cette charge d'ADN initiale. La distribution de cette charge d'ADN au sein des 96 puits peut donc s'apparier à un tirage sans remise, la probabilité qu'un puit soit positif pour un événement chromosomique (duplication ou délétion) peut donc être modélisé par une loi hypergéométrique (**Équation** : (2.1)). Nous permettant ainsi d'estimer la fréquence duplication / délétion λ pour chaque donneur (**Équation** : (2.2)) (**Figure** : 2.4).

$$\frac{\frac{(N-R)!}{W!(N-R-W)!}}{\frac{N!}{W!(N-W)!}} = \frac{(N-R)!(N-W)!}{N!(N-W-R)!} = \prod_{i=0}^{R-1} \frac{N-W-i}{N-i} \quad (2.1)$$

$$\lambda = \frac{R}{N} \quad (2.2)$$

Où :

- . N : représente le nombre de copie de chromosome 12 dans la charge d'ADN initiale (1.6×10^6 pour la PCR spécifique à la délétion, 3.2×10^6 pour la PCR spécifique à la duplication)
- . $W = \frac{N}{96}$ correspond au nombre de copie de chromosome 12 par puit
- . R représente le nombre total de recombinaison observées

L'intervalle de confiance (IC) à 95% est ensuite calculé grâce à une loi binomiale de sorte à modéliser la dilution initiale pour obtenir l'ADN d'*entrée*. Le puit contenant le *pool* des trois ADN spermatique est donc celui ayant les résultats les plus robustes, l'IC étant le plus resserré (**Figure** : 2.4), et permet donc d'établir le taux de délétion *de novo* à 1.8×10^{-5} (IC 95% : 1.4×10^{-6} ; 2.2×10^{-6}) tandis que le taux de duplication *de novo* est estimé à 7.7×10^{-6} (IC 95% : 6.1×10^{-6} ; 9.7×10^{-6}) montrant un

enrichissement environ deux fois supérieur des délétion par rapport aux duplications sur le site de *DPY19L2* (**Figure : 2.4**).

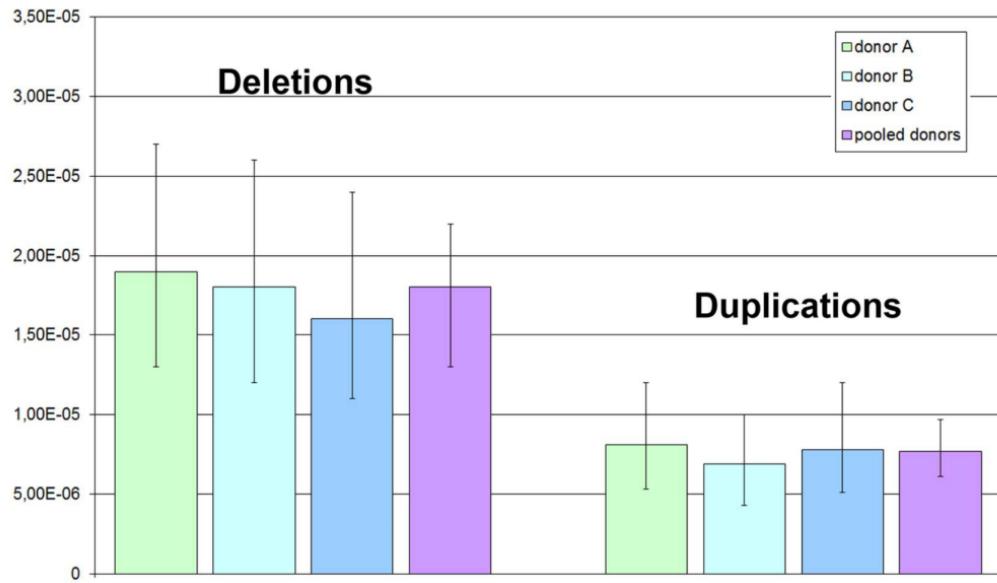


Figure 2.4 – Ratio des délétions / duplications *de novo* observées au locus *DPY19L2* déterminé par PCR digital à partir d'ADN spermatique de trois donneurs : [TODO]

[TODO : expliquer brièvement notre hypothèse expliquant ce phénomène]

Autres résultats

Cette étude a également été pour notre équipe l'occasion d'effectuer une étude plus approfondie des LCRs flanquant le locus de *DPY19L2*. Ainsi, nous avons pu mettre en évidence que les LCRs 1 et 2 contenaient 5 répétitions du site de reconnaissance consensus de PRDM9 (CCNCCNTNNCCNC), une protéine connue pour [TODO : décrire le rôle de PRDM9]. De même, nous avons pu mettre en évidence que les recombinaisons s'effectuaient le long de 5 points de cassures distincts répartis le long des LCRs 1 et 2 et que comme attendu, l'incidence des points de cassures dans les recombinaisons du locus *DPY19L2* était proportionnelle à la taille des régions [TODO : expliquer mieux].

2.2.2 La transcriptomique

Karaouzène et al. *Basic and Clinical Andrology* 2013, **23**:7
<http://www.bacandrology.com/content/23/1/7>



BASIC AND CLINICAL
ANDROLOGY

RESEARCH ARTICLE

Open Access

Comparative testicular transcriptome of wild type and globozoospermic *Dpy19l2* knock out mice

Thomas Karaouzène^{1,2}, Michèle El Atifi^{3,4,5}, Jean-Paul Issartel^{3,4,5}, Marianne Grepillat^{1,2,6}, Charles Coutton^{1,2,7}, Delphine Martinez^{1,6}, Christophe Arnoult^{1,2} and Pierre F Ray^{1,2,6*}

Abstract

Background: Globozoospermia is a male infertility phenotype characterized by the presence in the ejaculate of near 100% acosomeless round-headed spermatozoa with normal chromosomal content. Following intracytoplasmic sperm injection (ICSI) these spermatozoa give a poor fertilization rate and embryonic development. We showed previously that most patients have a 200 kb homozygous deletion, which includes *DPY19L2* whole coding sequence. Furthermore we showed that the *DPY19L2* protein is located in the inner nuclear membrane of spermatids during spermiogenesis and that it is necessary to anchor the acosome to the nucleus thus performing a function similar to that realized by Sun proteins within the *LINC-complex* (Linker of Nucleoskeleton and Cytoskeleton). SUN1 was described to be necessary for gametogenesis and was shown to interact with the telomeres. It is therefore possible that *Dpy19l2* could also interact, directly or indirectly, with the DNA and modulate gene expression during spermatogenesis.

In this study, we compared the transcriptome of testes from *Dpy19l2* knock out and wild type mice in order to identify a potential deregulation of transcripts that could explain the poor fertilization potential of *Dpy19l2* mutated spermatozoa.

Methods: RNA was extracted from testes from *DPY19L2* knock out and wild type mice. The transcriptome was carried out using GeneChip® Mouse Exon 1.0 ST Arrays. The biological processes and molecular functions of the differentially regulated genes were analyzed with the PANTHER software.

Results: A total of 76 genes were deregulated, 70 were up-regulated and 6 (including *Dpy19l2*) were down-regulated. These genes were found to be involved in DNA/RNA binding, structural organization, transport and catalytic activity.

Conclusions: We describe that an important number of genes are differentially expressed in *Dpy19l2* mice. This work could help improving our understanding of *Dpy19l2* functions and lead to a better comprehension of the molecular mechanism involved in spermatogenesis.

Keywords: Male infertility, Globozoospermia, Spermatogenesis, *Dpy19l2*, Transcriptome

* Correspondence: pray@chu-grenoble.fr

¹Université Joseph Fourier, Grenoble F-38000, France

²Laboratoire AGIM, CNRS FRE3405, Equipe "Génétique, Infertilité et Thérapeutiques", La Tronche F-38700, France

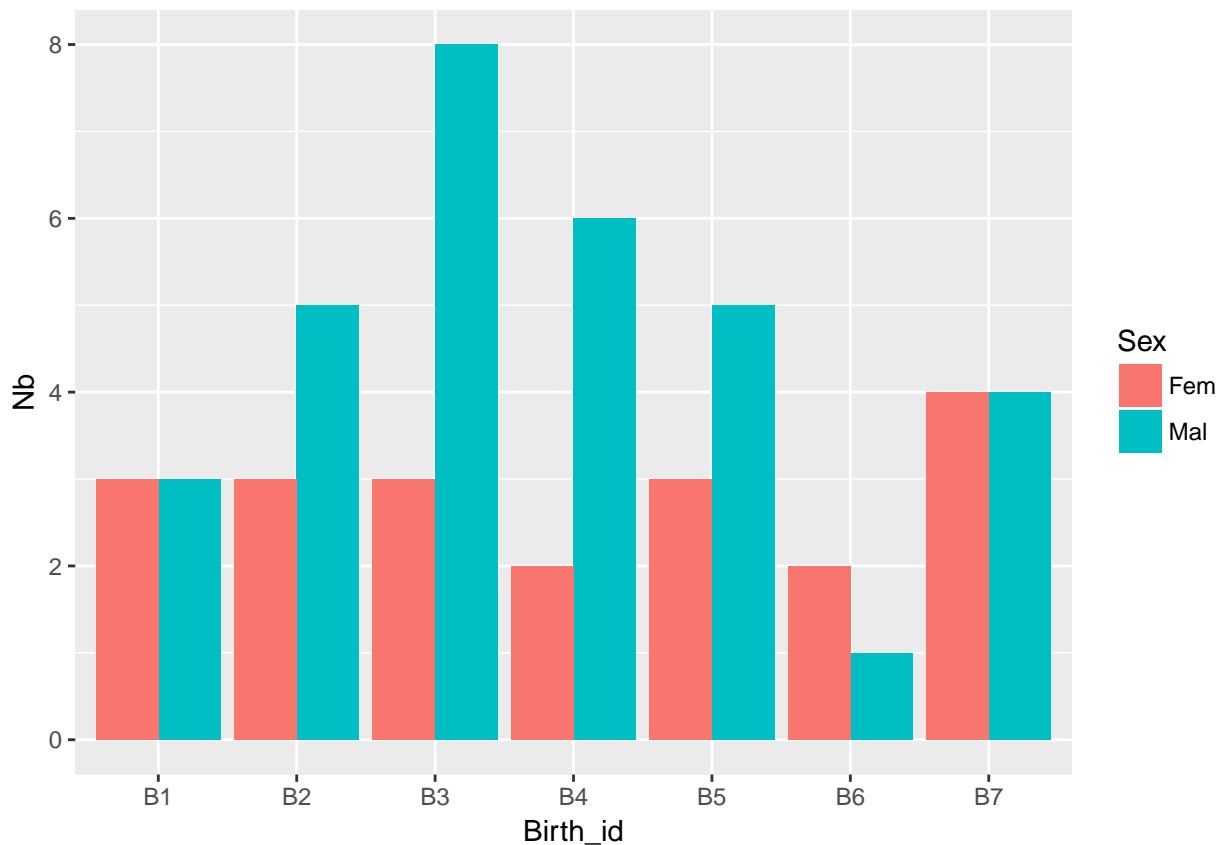
Full list of author information is available at the end of the article

Dans des études précédentes, notre équipe a réussi à démontrer que la protéine *DPY19L2* était localisée dans la membrane interne des noyaux des spermatides pendant la spermatogénèse et qu'elle est nécessaire pour fixer l'acrosome au noyau [TODO : insert ref]. De même, nous avons pu mettre en évidence que dans des cellules HEK cette protéine colocalisait avec la protéine SUN5 et que *Dpy19l2* pourrait être un partenaire de SUN5(V. Pierre et al., 2012). nous avons cherché à observer si, comme la protéine SUN5. Chez la souris, la protéine Sun1 est elle aussi nécessaire à la gamétogénèse et est connue pour permettre l'interaction entre le noyau et les télomères (Ding et al., 2007). Dans cette étude nous avons donc chercher à savoir si l'absence de la protéine Dpy19l2 pouvait entraîner des dérèglement transcriptionnel qui pourrait, entre autre, expliquer l'absence de la protéine PLC ζ dans les spermatozoïdes globozoocéphales. De plus, au cours de l'elevage des souris *Dpy19l2* KO au sein de notre laboratoire nous avons pu observé un excès de naissance de souris mâle lorsque l'on croisait deux souris *Dpy19l2^{+/−}* [TODO : insérer tableau et figure montrant le biais de sexe].

Afin de tenter d'expliquer ces observations, nous avons effectué une analyse comparative du transcriptome testiculaire de deux souris *Dpy19l2^{+/+}* (S1+ et S2+) et deux souris *Dpy19l2^{-/-}* (S1- et S2-) ayant pour but de mettre en évidence d'éventuels dérèglement transcriptionnels chez la souris KO. Pour effectuer ces analyses, nous avons donc extrait l'ARN testiculaire des 4 souris que nous avons ensuite hybridé sur des puces à ADN Affymetrix GeneChip® Mouse Exon 1.0 contenant des sondes pour 35.557 gènes murins. Cette étape nous a alors permis d'obtenir pour chacune des 4 souris les valeurs d'expression testiculaire de l'ensemble de leurs gènes. Pour chacun de ces gènes, nous avons donc chercher à savoir s'ils étaient différemment exprimés chez les souris S1- et S2- lorsqu'on comparait leur expression avec celle des souris S1+ et S2+. Pour cela, nous avons calculé quatre ratios (R1, R2, R3 et R4) (**Équation** : (2.3)). Les gènes pour lesquels au moins 3 de leurs ratio étaient $\geq 1,7$ furent considérés comme sur-exprimés tandis que ceux pour lesquels 3 de leurs ratio étaient $\leq 0,58$ ($\frac{1}{1,7}$) furent considérés comme sous-exprimés.

$\forall gene \in \{genes\}$:

$$\begin{aligned} R1_{gene} &= \frac{\exp_{gene}(S1-)}{\exp_{gene}(S1+)} & R2_{gene} &= \frac{\exp_{gene}(S2-)}{\exp_{gene}(S1+)} \\ R3_{gene} &= \frac{\exp_{gene}(S1-)}{\exp_{gene}(S2+)} & R4_{gene} &= \frac{\exp_{gene}(S2-)}{\exp_{gene}(S2+)} \end{aligned} \quad (2.3)$$



Ainsi cette étude a pu mettre en évidence la sous-expression de 6 gènes (incluant *Dpy19l2*) et la sur-expression de 70 gènes chez les souris *Dpy19l2^{-/-}*. Parmi ces gènes, nous ne figurait pas *Plcζ* indiquant que l'absence de cette protéine chez les spermatozoïdes globozoocéphales n'étaient pas directement dûe à un dysfonctionnement transcriptionnel. Afin de prédire les fonction moléculaire dans lesquels étaient impliqués ces gènes, nous nous sommes servi du logiciel PANTHER (Mi et al., 2017). Ainsi, nous avons pu constater que 23 gènes codant pour des protéines de liaison étaient dérégulés (**Figure : 2.5**), dont 12 sont des protéines de liaison aux acides nucléiques (**Figure : 2.6**) suggérant que *Dpy19l2* pourrait effectivement intéragir avec l'ADN. D'autre fonctions moléculaires telles que l'activité catalytique, la régulation de la transcription et des protéines ayant des fonctions structurelles étaient également dérégulées chez les souris KO. Ces derniers sont particulièrement intéressant lorsque l'on sait que les spermatozoïdes globozoocéphales sont caractérisés par plusieurs défauts structurels.

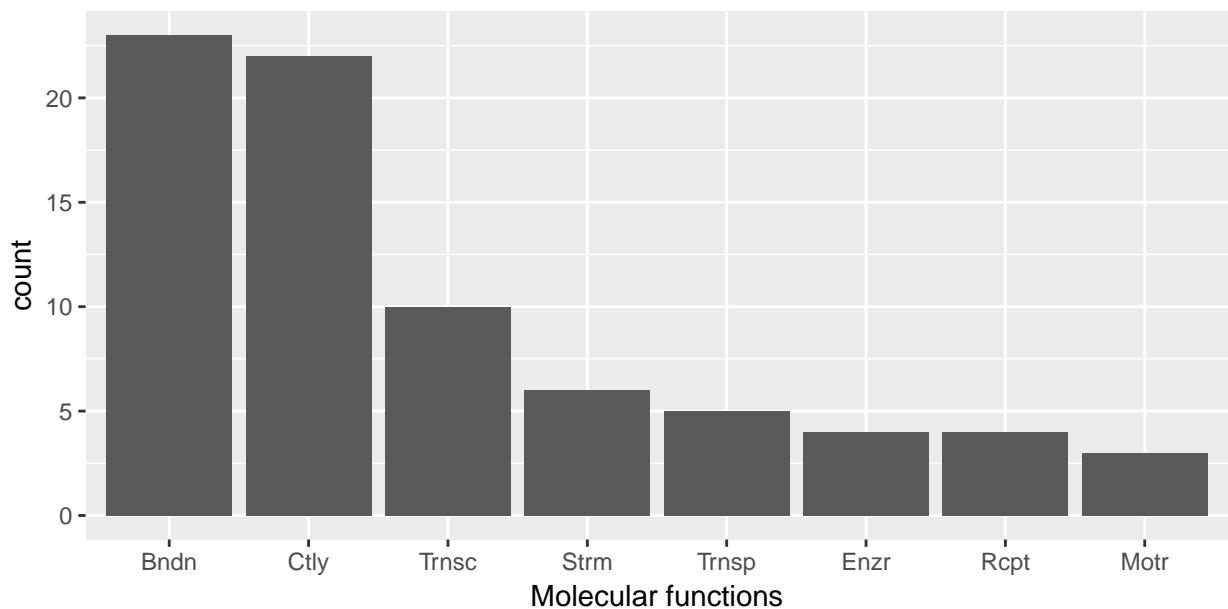


Figure 2.5 – Fonctions moléculaires affectées par au moins un gène dérégulé : Bndn = Binding, Ctly = Catalytic, Trnsc = Transcription, Strm = Structural molecule, Enzr = Enzyme regulator, Rcpt = Receptor, Motr = Motor

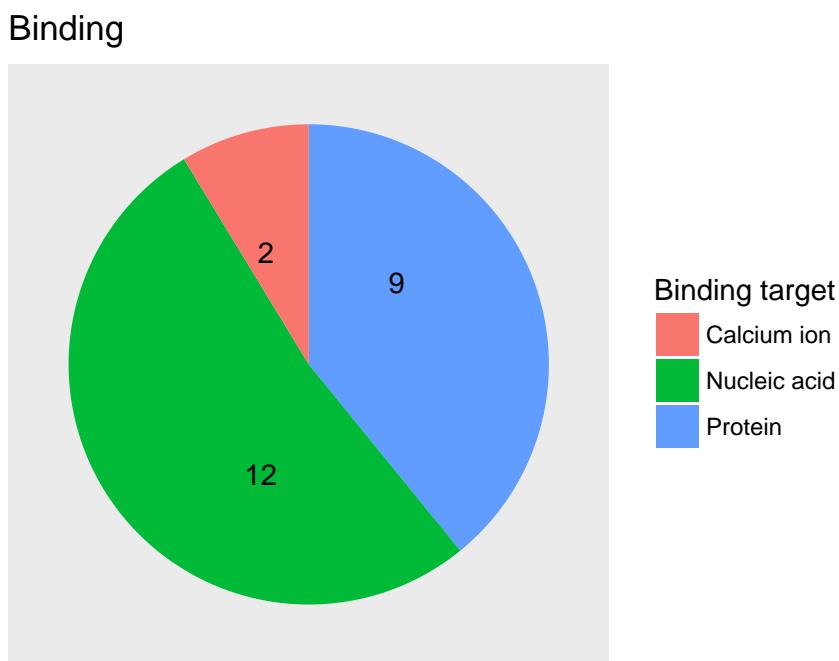


Figure 2.6 – Détails des sites ciblés par les gènes dérégulés ayant la fonction moléculaire "Binding" : Les nombres présent au milieu des parts indiquent la quantité de gène impliqués dans cette fonction moléculaire

Au cours de cette étude nous n'avons pas pu mettre en évidence la cause expliquant le biais de sexe observé dans les naissance des souris. De même aucune dérégulation dans l'expression du gène *Plc ζ* n'a pu être observé. Cependant, et que ceux-ci soient modestes, nous avons montrés que les souris *Dpy19l2*^{-/-} présentaient des dérèglements transcriptionnels affectant plusieurs fonctions moléculaires pouvant ainsi expliquer, du moins en partie, les nombreux défauts morphologiques caractérisant les spermatozoïdes globozoocéphales. De même, nous avons pu observer un dérèglement de nombreux gènes impliqués dans la liaison d'acide nucléique et de protéine pouvant ainsi expliquer les défauts d'ancre de l'acrosome au noyau chez les spermatozoïdes globozoocéphales.

2.3 Conclusion

Chapitre 3

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

3.1 Intro

3.2 Résultats

3.2.1 Description de la pipeline

Notre pipeline d'analyse effectue l'ensemble des étapes allant de l'alignement des données jusqu'au filtrage des variants

1. **L'alignement** : L'alignement des *reads* le long du génome de référence est effectué par le logiciel MAGIC (Su et al., 2014). Celui-ci l'intégralité pour l'ensemble des analyses en aval l'ensemble des *reads* dupliqués et / ou s'alignant à plusieurs zones du génome. Au cours de cette étape, MAGIC va produire également 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 (R. Nielsen et al., 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** : 3.1) (**Table** : 3.1). 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.

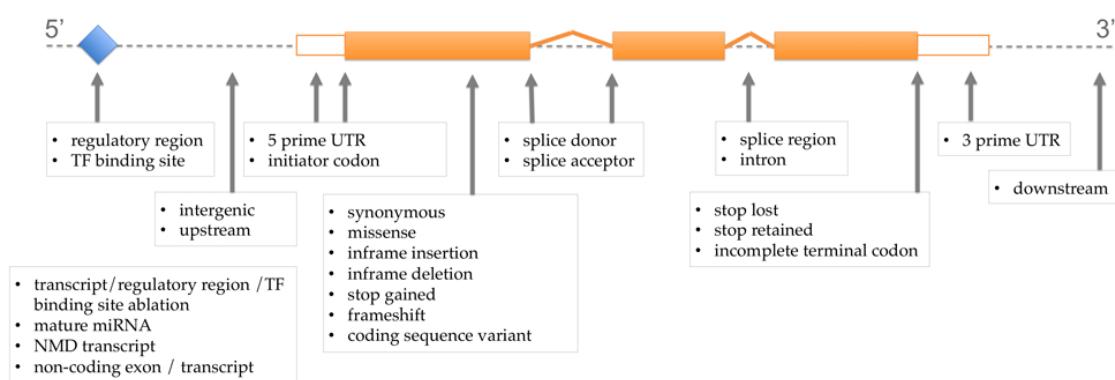


Figure 3.1 – Listes des différentes conséquences prédictes par VEP et leur positionnement sur le transcrit d'après [VEP site](<http://www.ensembl.org/info/genome/variation/consequences.jpg>)

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 modifiée 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 3.1 – Liste simplifiée des conséquences prédictes par VEP avec leur description et impact associé

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 codon
Inframe insertion / deletion	MODERATE	An inframe non synonymous variant that inserts / deletes bases into in the coding sequence
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

3.2.2 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 atteints par une mosaïque d'anomalie morphologique du flagelle.

Un récapitulatif des familles et de leur phénotype est disponible dans la table 3.2.

Table 3.2 – Tableau récapitulatif des familles séquencées et de leur phénotype

Family	Individuals	Phenotype	Year	Place
Az	2	Azoospermia	2012	Mount Sinai Institut
FF	2	Fertilization failure	2014	Genoscope (Evry)
MMAF1	2	MMAF	2014	Genoscope (Evry)
MMAF2	2	MMAF	2014	Genoscope (Evry)
MMAF3	2	MMAF	2014	Genoscope (Evry)
MMAF4	3	MMAF	2014	Genoscope (Evry)

Resultats 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 exceptés, c'est à dire une médiane de 91438630 *reads*. Seul les deux frères AZ1 et AZ2 se distinguent près de 3 fois plus de *reads* que pour les autres patients. Cette différence peut être expliquée car ces deux patients sont les deux seuls à avoir 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 (**Table : 3.2, Figure : 3.2 - A**).

L'ensemble de nos exomes ayant été réalisés en *paired-end*, les deux extrémités de chaque fragment sont séquencés 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 représentent entre 1.2 et 5.5 % des *reads* de nos individus (**Figure : 3.2 - B**).

Une fois cela fait, nous vérifions la “compatibilité” des deux *ends* composant chacun des *reads* s'étant correctement alignés. Un *read* 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 environ ... % (médiane) des *reads* s'étant correctement alignés. (**Figure : 3.2 - 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 : 3.2 - C**).

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

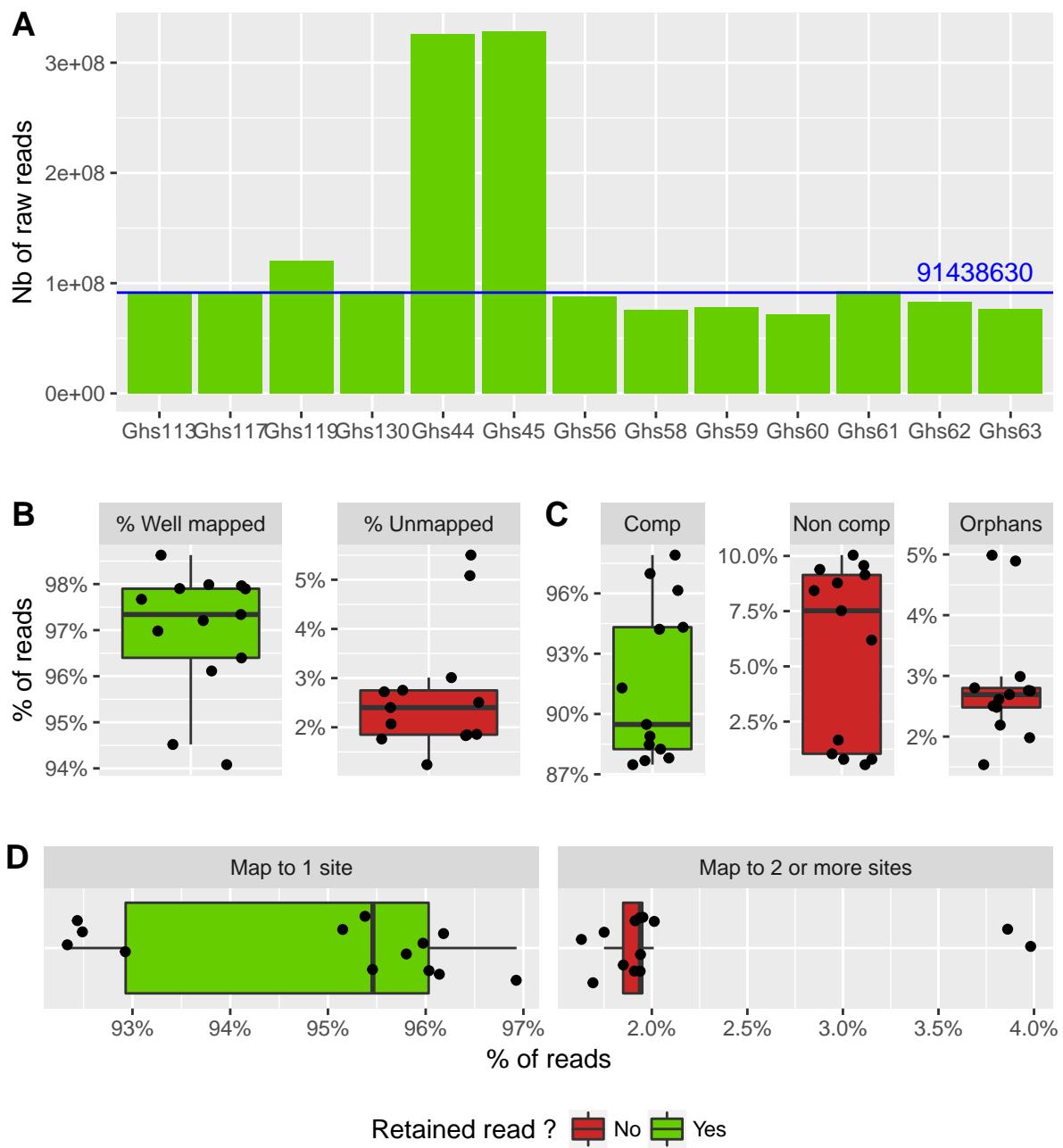


Figure 3.2 – Processus simplifié du contrôle qualité des *reads* : Pour chacun des graphiques, les *reads* représentés en vert sont conservés tandis que ceux en rouge sont filtrés. ****A**** : Quantité de *reads* bruts générés pour chaque patient au cours de l'étape de séquençage. La médiane des *reads* 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 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 vu 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 vu 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 : 3.3 - 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 filtrer, ainsi environ 86.3 % des variants DS passent ce filtre. 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 : 3.3 - 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 portent 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 : 3.3 - C**).

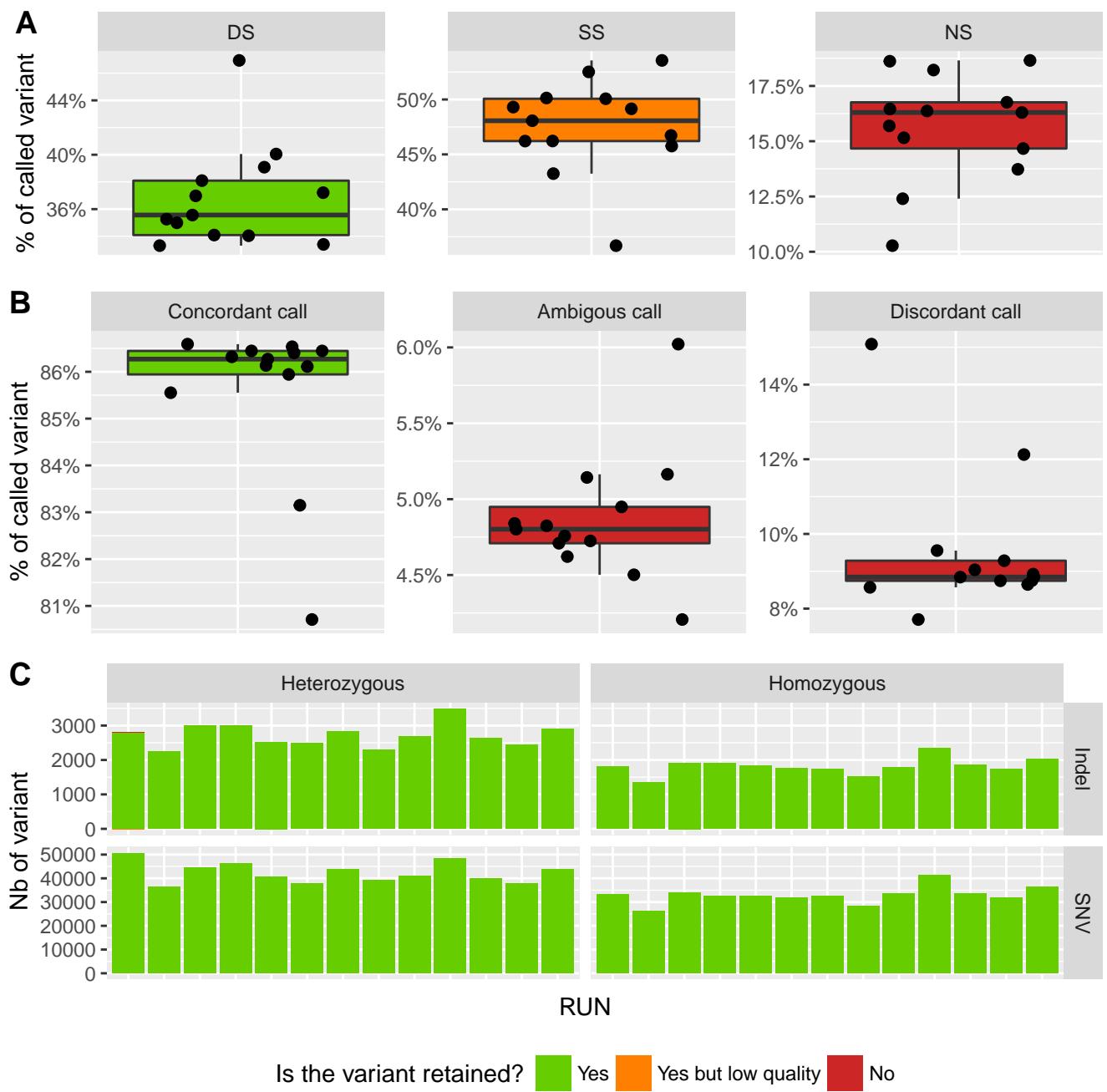


Figure 3.3 – Contrôle qualité des variants appelés : Pour chacun des graphiques, les variants représentés en vert et en orange sont conservés tandis que ceux en rouge sont filtrés. **A** : Distribution du *stranding* des appels pour chaque patient. **B** : Comparaison des appels entre les deux *ends* 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 connaître les prédictions de pathogénicités fournies par SIFT et PolyPhen pour les variants faux-sens.

Après avoir annoter 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 environs 5 transcrits par gènes) (**Figure : 3.4 - 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 :3.1**). 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 :3.1**), 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 : 3.4 - 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 : 3.4 - C**) (À discuter!!!!).

(À discuter!!!!) (**Figure : 3.4 - D**)

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

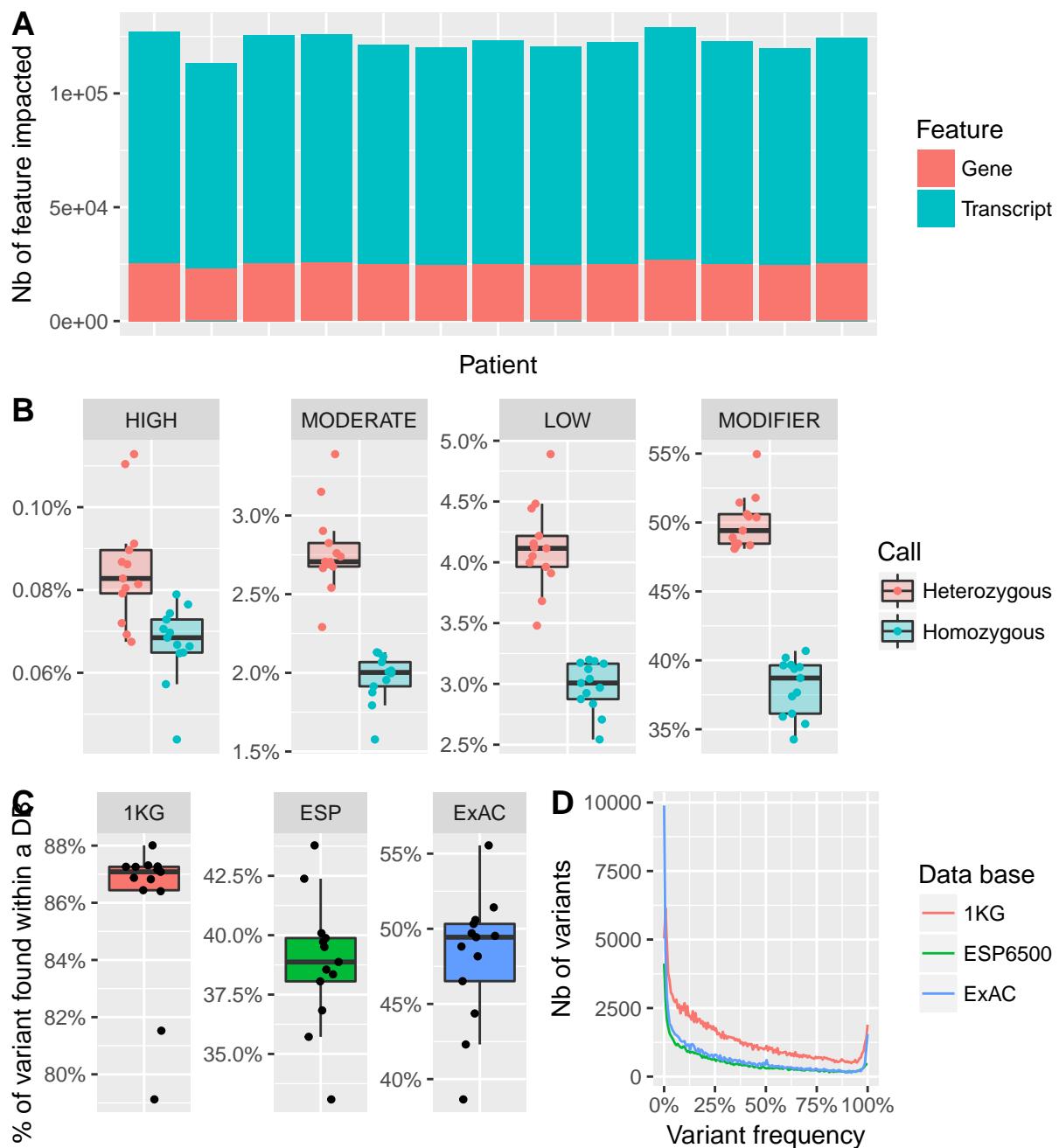


Figure 3.4 – Annotation des variants par VEP : **A** : Quantification du nombre de gènes (en bleu) / transcrits (en rose) impactés par au moins un variant pour chaque patient chacun représentés par une barre. **B** : Distribution des impacts HIGH MODERATE LOW et MODIFIER en fonction des patients et du génotype du variant. **C** : Pourcentage de nos 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 entre autres 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 duos ou des trios d'individus provenant de même fratries et étant caractérisés par le même 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 par l'ensemble des membres de la fratrie.
2. **Filtre 2 : Génotype des variants** : Dans ces études, nous avons émis l'hypothèse d'une transmission récessive du phénotype. Ainsi, seul les variants homozygotes ont été conservés. (**Figure** : ??, ??).
3. **Filtre 3 : Impact du variant** : Afin de ne conserver que les variants ayant un effet potentiellement tronquant 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 séquences protéique. Pour les variants faux sens (changement d'un seul aa de la séquence protéique) il est plus difficile de se décider [TODO insert citation] nous avons donc utilisé les logiciels SIFT (P. Kumar et al., 2009) et Polyphen (I. A. Adzhubei et al., 2010) et filtré l'ensemble des faux-sens prédit 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 transcrits annotés NMD soient responsables du phénotype. Dès lors, ces transcrits furent eux aussi 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 systématiquement filtrer entre 36576 et 44581 transcrits différents par patients, soit une moyenne de NaN variants par

individus (**Figure : 3.5**).

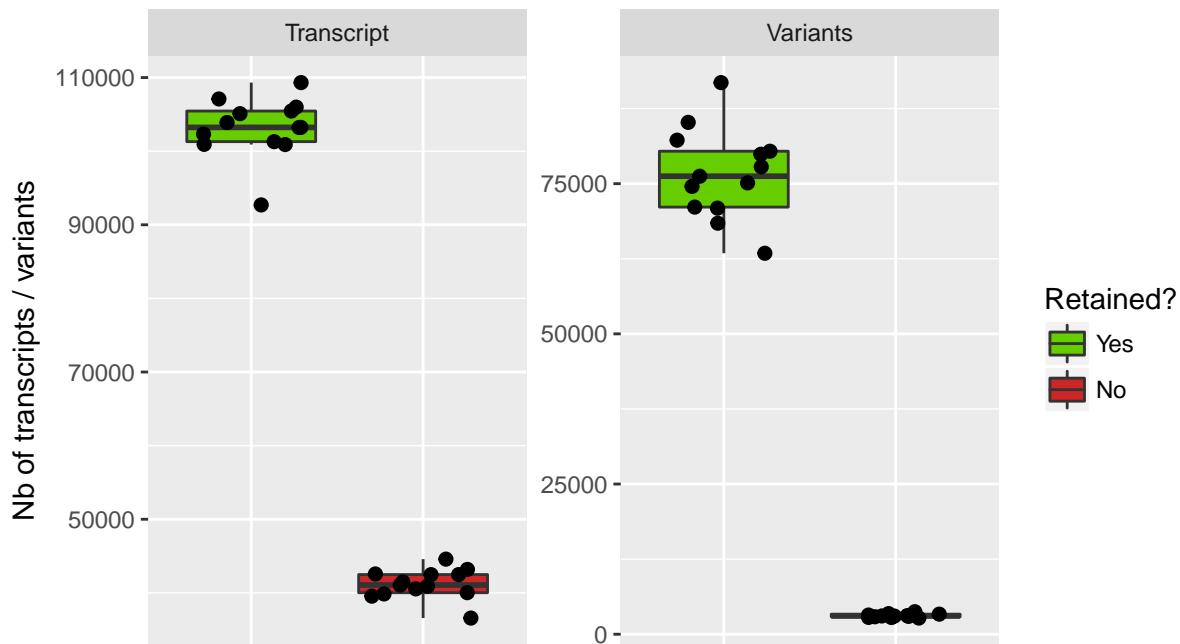


Figure 3.5 – Filtrage des transcrits jugés "non pertinents" et des variants les chevauchant : Pour chaque patients nous avons filtrer les transcrits jugés "non pertinents" pour l'analyse, c'est à dire ceux ne codant pas pour une protéine et ceux annoté NMD. Dès lors, l'intégralité des variants chevauchant 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 un avis sur l'effet délétère de celui-ci. En effet, il est peu probable qu'un 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 134. L'ensemble de ces individus peuvent être soit sains soit présenter l'un des 6 phénotypes étudié au cours de nos différentes études (**Table : ??**). Ces phénotypes étant très différent, il n'est pas aberrant d'émettre l'hypothèse qu'ils que leurs causes génétiques le soient également. 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 : 3.6**). 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.

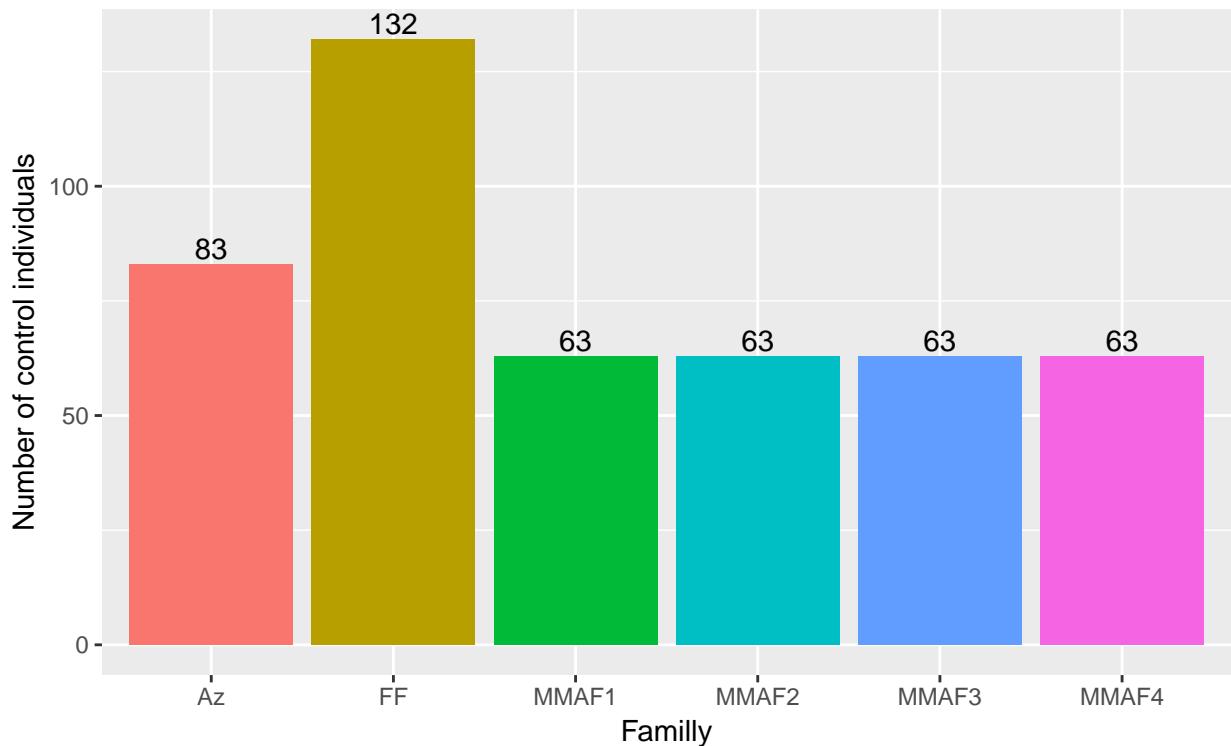


Figure 3.6 – Nombre d’individus composant la cohorte contrôle de chaque famille : Ici, chaque barre représente une famille et sa hauteur est déterminée par le nombre d’individus composant la cohorte contrôle à laquelle elle a été confronté. Chaque individu de la cohorte contrôle a été séquencés en WES 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 individu 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 : 3.7**). En effet, tandis que le filtre “Transcript relevance” (filtre n°4) élimine 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 : 3.7 - 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 : 3.4 - 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 : 3.4 - 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 filtrer 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 stringeance reste quasi identique (**Figure : 3.7 - B**) il est tout de même intéressant de noter que désormais le filtre “Variant impact” apparaît 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 : 3.7 - B**).

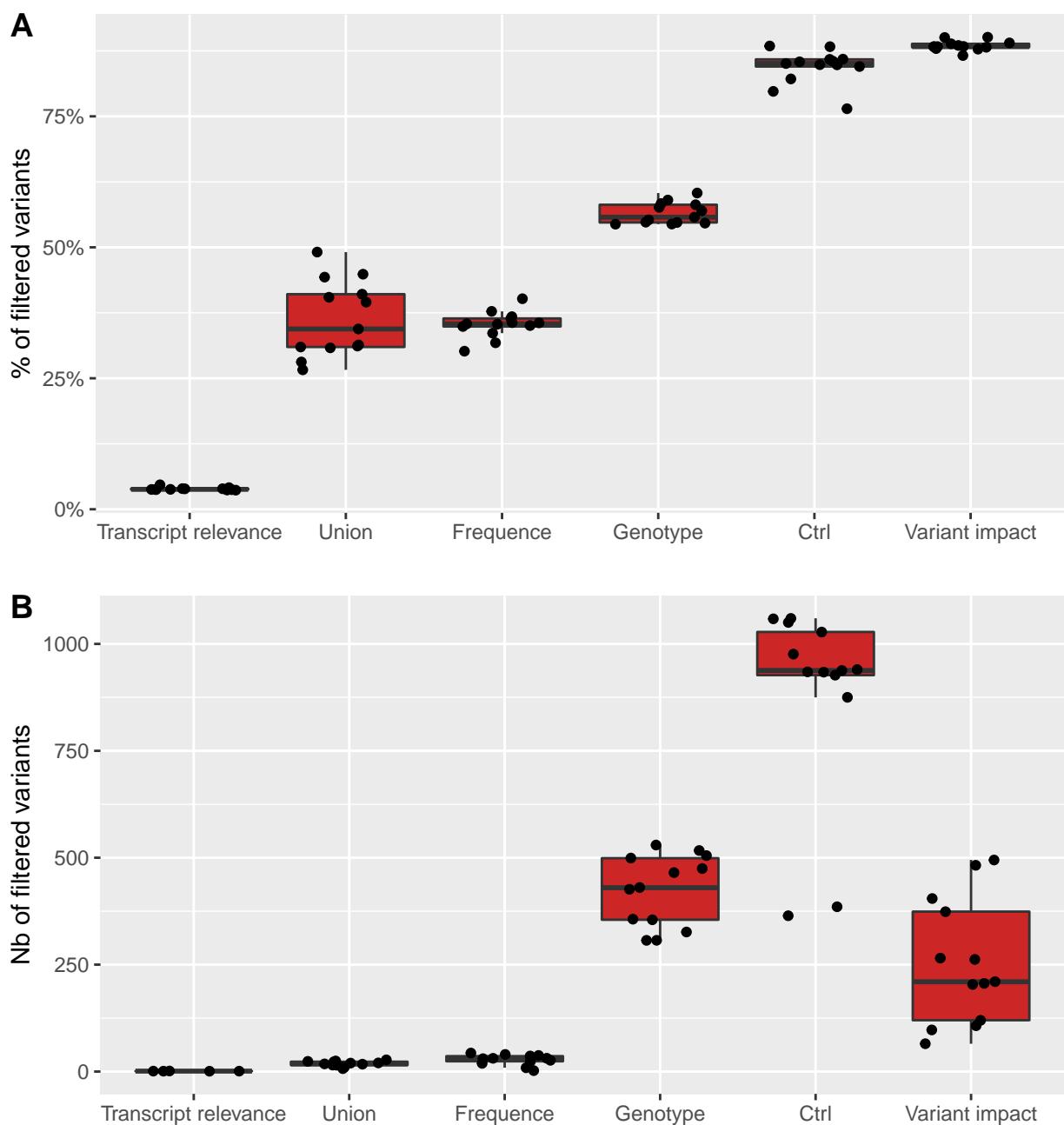


Figure 3.7 – Comparaison de l'efficacité de chacun des six filtres utilisés : **A** : Comparaison du pourcentage de variants filtrés par chacun des six filtres indépendamment les uns des autres pour chaque patient (représenté par les points. Dès lors, un même variants peut être filtré par plusieurs filtres. **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 : 3.3**) 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 (TODO fig). De plus, des mutations du gène *Spink2* chez la souris avait déjà été identifiée comme induisant des défauts de la spermatogenèse (B. Lee et al., 2011).
2. **Famille FF** : Pour cette famille, le gène *PLCζ1* a passé l'ensemble des filtres. Nos connaissances sur la fonction de ce gène et notamment son rôle dans l'activation ovocytaire (TODO : REF) on fait de ce gène le candidat idéal pour expliquer le phénotype de ces deux frères.
3. **Famille MMAF1** : L'analyse bibliographique des 2 gènes ayant passé l'ensemble des filtres n'a ici 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, il y a quelques années 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 5 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.
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* localise spécifiquement dans le cytoplasme du spermatide, ainsi que dans le corps résiduel et la pièce intermédiaire du flagelle du spermatozoïde mature (Y. Hu, Yu, Shaw, Renfree, & Pask, 2011). Ces données pourraient corrélérer avec le phénotype MMAF de ces 3 frères.

Table 3.3 – Liste des gènes ayant passé l’ensemble des filtres pour chaque famille

AZ	FF	MMAF1	MMAF2	MMAF3	MMAF4
SPINK2	PLCZ1	PLA2G4B JMJD7-PLA2G4B	MYH11 DNAH1	PCSK5 WEE2 GBP2 FCGR3A ZFYVE28	TGIF2
GUF1					

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 : 3.8**) 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 aussi séquencer en Sanger les séquences codantes du gène *SPINK2* pour une partie des 611 individus analyser 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 effectués 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 effectués 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 : 3.8**).

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 supposé 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 vu des résultats il apparaît 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.

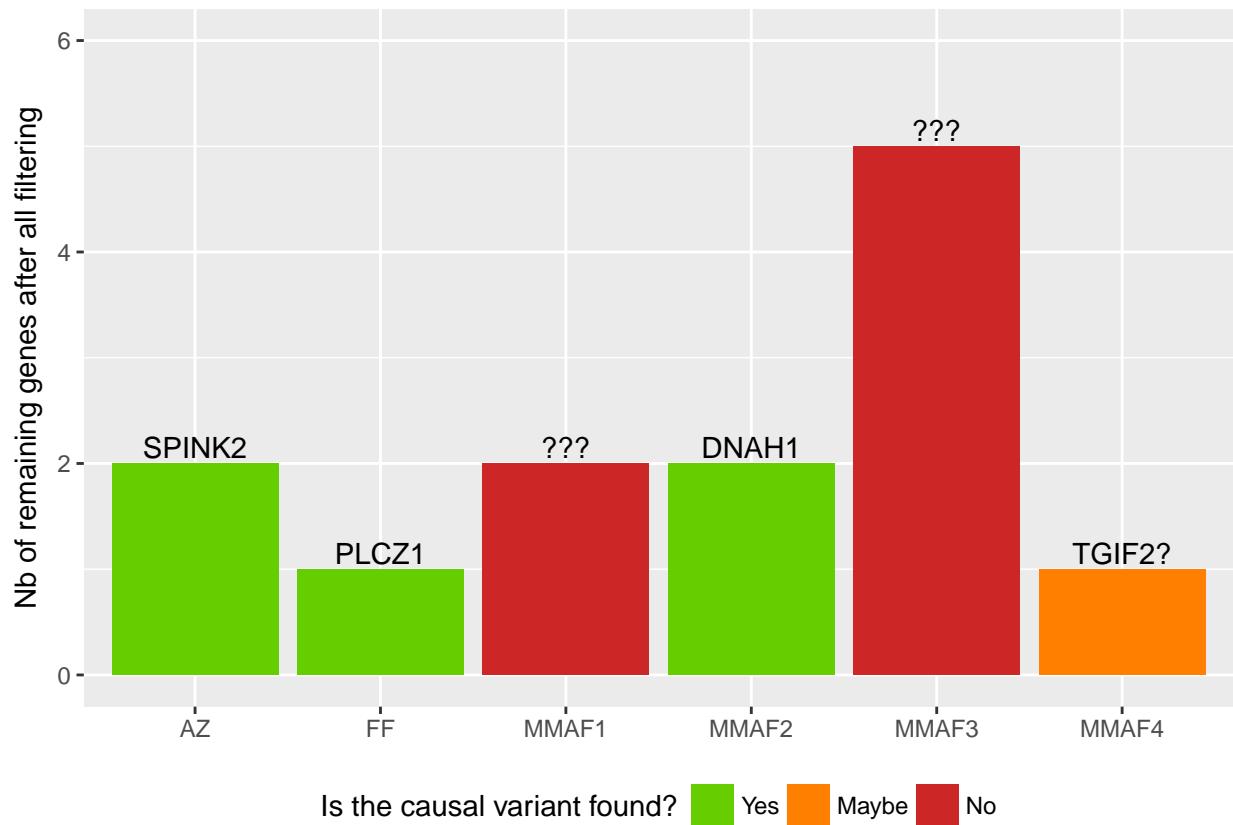


Figure 3.8 – Nombre de gènes passant l’ensemble des filtres par famille : Chaque barre représente une des familles analysées. La hauteur de cette barre correspond au nombre de gènes ayant passé l’ensemble des filtres pour chaque famille. Les barres vertes caractérisent les familles pour lesquelles le 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 suivi 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)

Table 3.4 – 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

3.2.3 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 ... venaient de et de ... ces séquençage ont été effectué dans ... centres différents que sont (listes des centre de séquençage) et sur ... plateforme : liste des plateformes

Application de la pipeline - Résultats

Après avoir appelé les variants de nos ... patients, nous avons obtenu une total de 677228 variants différents comprenant 628793 SNVs et 48435. Ces variants étant répartis entre chaque patients qui portaient environnements chacun 81618 SNV et 5148 indels dont 0 % 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 kit de capture dans une table) en revanche nous pouvons écarté un effet dus à la plateforme de séquençage ou encore le modèle de séquenceur puisque tout ces projets ont été réalisés sur des Illumina HiSeq2000 (**Table : 3.4**) (**Figure : 3.9 - 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 meilleur qualité que les appels SS. Cette augmentation des appels DS au cours du temps pourrait s'expliquer par une amélioration des protocole de séquençage qinsi que des kit 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 environnements ... % en (... Année) avec un projet réalisé au Génoscope. Ces derniers appels étant systématiquement filtrés, ils n'altèreront en rien les résultats obtenus en aval ormis le fait qu'ils réduisent la quantité des données utilisées (**Figure : 3.9 - B et C**).

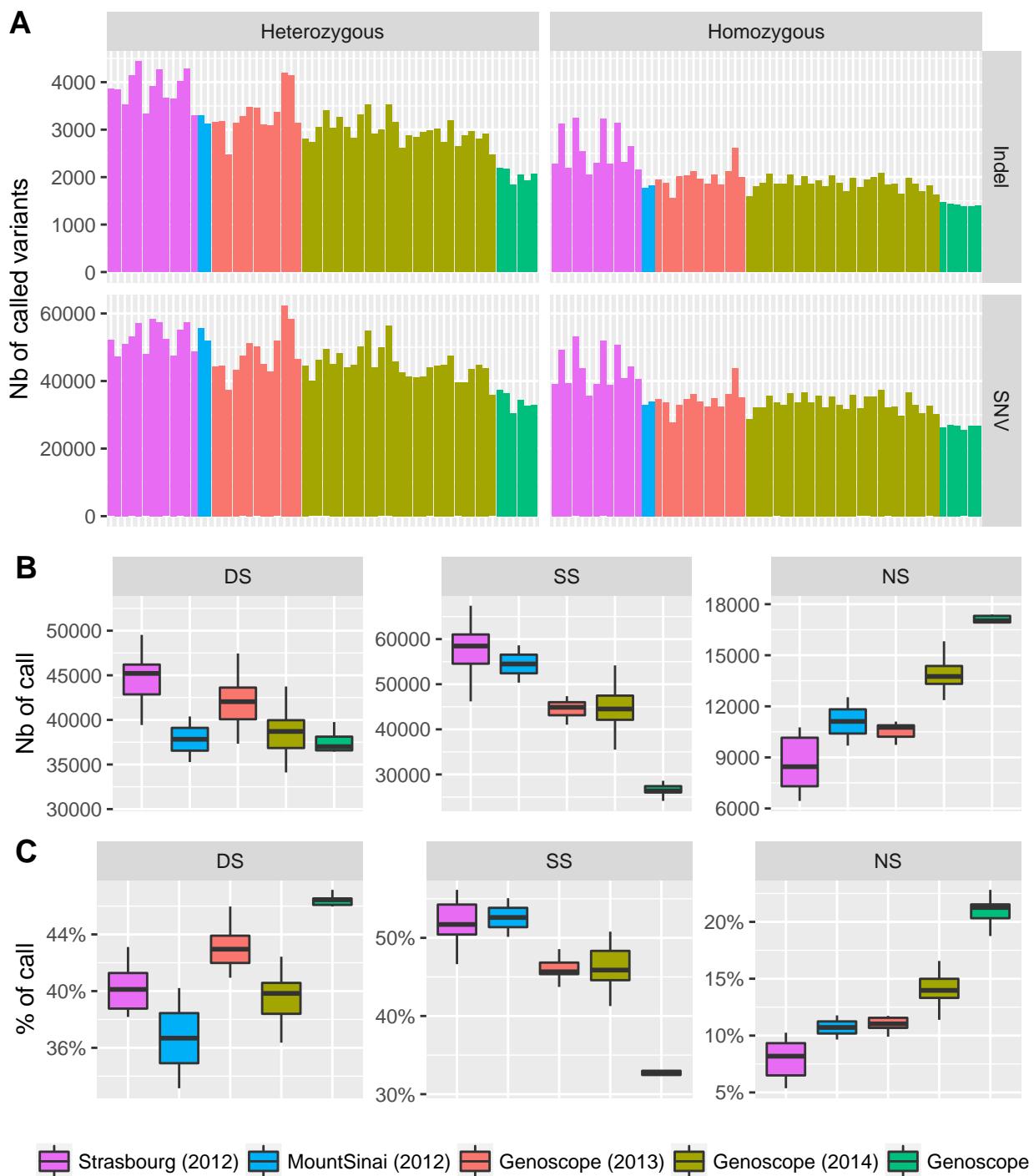


Figure 3.9 – Résultats de l'appel des variantes par individus et par projet de séquençage : Chaque couleur définit un projet de séquençage caractérisé par un centre de séquençage et une année. ****A**** : Quantification pour chaque individus (représentés par les barres) du nombre de variantes (SNVs et 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

Après avoir appliqué les mêmes filtres que ceux décrit précédemment à l'exception du filtre n°... Union puisqu'ici nous avons uniquement des individus non apparentés, nous avons pu obtenir une liste de 1568 variants différents composés de 1359 SNVs et 209 indels et impactant un total de 1306 gènes distincts. Ces variants étant répartis sur l'ensemble de nos ... patients ceux-ci portaient en moyenne 25 SNVs et 4 indels, de sorte que chacun d'entre eux avaient entre 1 et 73 gènes impactés par au moins un variant (**Figure : 3.10**).

Parmis l'ensemble de nos patients, 3 révélèrent porter au moins un variant passant l'ensemble des filtres sur le gène *DNAH1*, candidat évident pour ce phénotype. Ainsi, le patient Ghs90 porte 3 variants successifs induisant 3 variations faux-sens, le patient Ghs95 porte lui un seul variant entraînant lui aussi un faux-sens et le patient Ghs122 porte un indel entraînant un décalage du cadre de lecture (**Table : 3.5**). La cause génétique responsable du phénotype MMAF de ces ... (3) patients ayant été identifié, leurs variants furent ensuite retirer de notre liste réduisant ainsi celle-ci à ... variants différents impactant ... gènes.

Table 3.5 – liste des variants passant les filtres et chevauchant le gène *DNAH1*

Run ID	Variant coordinates	Consequence	Gene symbol
Ghs90	3-52382919-A-C	missense	DNAH1
Ghs90	3-52382920-T-C	missense	DNAH1
Ghs90	3-52382922-T-C	missense	DNAH1
Ghs95	3-52422540-C-G	missense	DNAH1
Ghs122	3-52414073-CC-C	frameshift	DNAH1

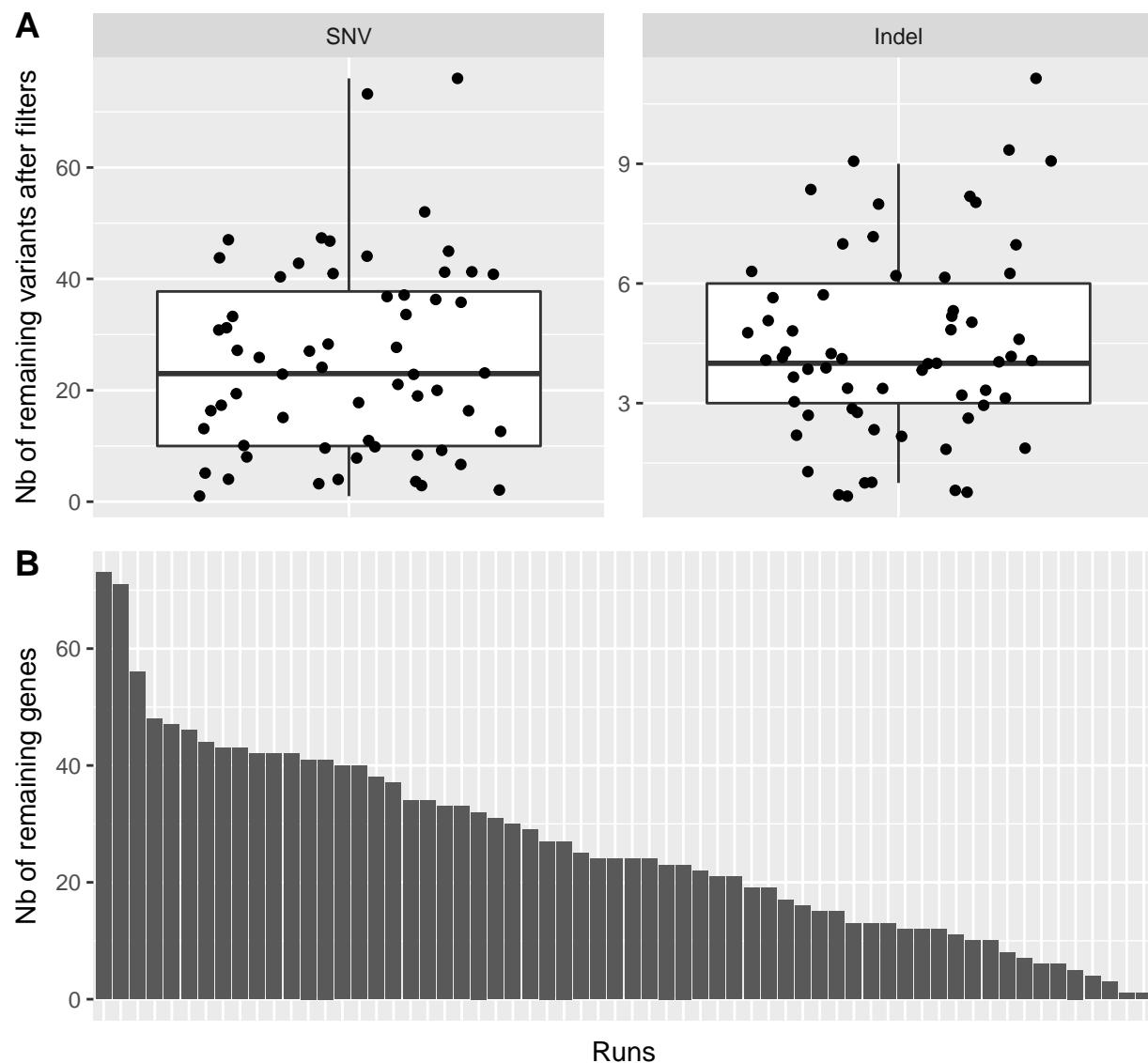


Figure 3.10 – TODOOOOOOOOOOOOOOOO : **A** : Quantification du nombre de SNVs et indels ayant passé l'ensemble des filtres pour chaque patients. **B** : Nombre de gènes impactés par au moins un variant ayant passé les filtres pour chaque individus représentés par les barres

Analyse des listes de gènes

Chapitre 4

MutaScript

4.1 Introduction

Il y a quelques années, le séquençage Sanger était encore massivement utilisé en recherche clinique. Cette technique était extrêmement couteuse en temps et en argent freinant considérablement la progression des recherches du type phénotype-génotype de sorte qu'en 2011, les causes de plus de 3.500 pathologies à transmission Mendélienne restaient inconnus (Stitzel, Kiezun, & Sunyaev, 2011). L'émergence du séquençage haut-débit a immédiatement initié une nouvelle ère dans le domaine de la recherche clinique et permettant dans un temps record et à cout raisonnable d'obtenir la séquence de l'intégralité du génome ou bien des régions exomiques. Ce bond technologique est à l'origine de grandes avancées permettant de lier plus de ... variants génétiques à une pathologie mendélienne [citation].

Cependant, de par sa masse, les données produites créent de nouvelles problématiques pour les généticiens qui se retrouvent désormais face au "déluge de données génétiques" (Schatz & Langmead, 2013). En effet, un génome humain typique compte en moyenne 3,5 millions de variants différents et plus de 1000 variations du nombre de copies (CNVs) (Gonzaga-Jauregui et al., 2012) après comparaison avec le génome de référence. Parmi ceux-ci, 20.000-25.000 d'entre eux impactent des régions codant pour une protéine avec environs 10.000 variants impliquant un changement d'acide aminé et 50-100 prédict comme tronquant la protéine (Gonzaga-Jauregui et al., 2012). Ainsi, les analyses fastidieuses permettant de mettre en évidence le variant responsable de la pathologie font désormais partie du quotidien des généticiens. Appliquer cette tâche est d'autant plus laborieux qu'elles nécessitent entre autres des compétences en informatique et en statistiques qui sont assez éloignées des compétences "traditionnelles" des généticiens. De manière générale, ces analyses se découpent en trois étapes principales. La première est l'étape d'alignement qui basiquement consiste à aligner les reads générés lors de l'étape de séquençage le long d'un génome de référence. Une fois cela fait, l'étape d'appel des variants consiste à recenser l'ensembles des "différences" observées entre les données de l'individu séquencé et le génome de référence permettant ainsi d'établir

une liste de SNVs et de petites insertions / délétions (indels) avec leur génotype associé. Comme dit précédemment, cette liste peut atteindre 25.000 variants différents par individus. La dernière des étapes regroupe l'annotation et le filtrage des variants. Elle représente souvent la faiblesse des analyses phénotype-génotype puisque dans une grande partie des cas, le pouvoir filtrant n'est pas assez puissant pour obtenir une liste de variants suffisamment petite pour qu'elle soit interprétable par l'homme, ainsi le variant causal se retrouve bien souvent noyé parmi d'autre variant rendant l'analyse et l'interprétation moins efficaces.

Améliorer la qualité de l'annotation et le filtrage des variants dans les analyses phénotype-génotype se révèle donc être une des clés permettant d'améliorer l'efficience de ces analyses, c'est pourquoi nous avons développé le score MutaScript. Ce score a pour but de classer l'ensemble des transcrit codant en fonction de leur charge mutationnelle avec l'idée sous-jacente que les transcrits les plus mutés dans la population générale ne sont probablement pas impliqués dans des pathologies sévères à transmission Mendélienne, *a contrario* ceux retrouvés comme n'étant pas / peu mutés le sont probablement. Pour ce faire, le score MutaScript repose sur trois (...). La première étant le jeu de transcrit fournit par Ensembl (B. L. Aken et al., 2017) qui comporte ... transcrits codants. Afin de connaître la charge mutationnelle de ces transcrits, nous nous sommes basées sur les variants mis à disposition par *the Exome Aggregate Consortium* (ExAC) (Lek et al., 2016) qui réunit les données d'exome de 60.706 individus non apparentés que nous avons ensuite annoté grâce au logiciel *variant effect predictor* (VEP) (W. McLaren et al., 2016) afin de prédire l'impact de chaque variant sur l'ensemble des transcrits qu'ils chevauchent de sorte à ce que les variants ayant un impact prédit comme étant délétère aient une plus grosse contribution au score MutaScript que ceux ayant un impact faible. À l'heure actuelle, plusieurs logiciels tel que SIFT (P. Kumar et al., 2009) ou encore PolyPhen-2 (I. A. Adzhubei et al., 2010). Cependant, ces logiciels donnent un score pour un variant et n'extrapolent pas leurs prédictions au niveau du gènes et/ou du transcrits. D'autres logiciels tel que Exomiser (Robinson et al., 2014) et Endeavour (Tranchevent et al., 2016) cependant, pour pouvoir fonctionner, ces logiciels nécessitent d'avoir des connaissances génétiques sur la pathologie étudiée. Plus récemment, favorisé par l'émergence de gros jeux de données exomiques comme ExAC, d'autres scores ont vu le jour tel que le *residual variance intolerance score* (RVIS) (Petrovski et al., 2013) ou encore *the Probability of loss-of-function Incoherency* (pLI) (Lek et al., 2016). MutaScript se présente comme une alternative à ces derniers scores et, bien que sa fonction soit similaire, il diffère de ceux-ci sur de nombreux points. Tout d'abord, MutaScript donne un score à l'ensemble des transcrits codant pour une protéine là où pLI donne un score seulement au transcrit consensus de chaque gène et RVIS qui agrège les séquences codantes de l'ensemble des transcrits d'un même gène créant ainsi un transcrit "chimérique". Ce procédé, bien qu'il facilite l'interprétation du score, engendre une perte d'information puisque l'on se retrouve avec un seul score par gène et non par transcrits. De plus, dans la conception de leur score, RVIS et pLI ne considère que les variants dit *loss-of-function* (LoF), c'est à dire les variants impactant l'épissage, engendrant un codon stop ou un décalage du cadre de lecture. Cependant, ces variants ne représentent que ... % des variants fournis par la base de données ExAC. C'est pourquoi, MutaScript prend en compte l'ensemble

des variants, peu importe leur impact sur les différents transcrits qu'ils chevauchent, et leur attribue un poids en fonction de cet impact de sorte à ce que les variants délétères contribuent plus au score d'un transcrits que les autres. Aussi, l'étude des scores RVIS et pLI nous a permis de mettre en évidence une corrélation forte entre le score qu'ils attribuent à un gène et la taille de la séquence codante (CDS) de ce même gène. Cette corrélation étant principalement due à un biais causé par leur manière de calculer leur score et non à une réalité biologique, MutaScript fut construit de sorte à éviter cette corrélation qui peut mener à des erreurs d'interprétations. Afin d'évaluer le score MutaScript nous l'avons confrontés au RVIS (Petrovski et al., 2013) ainsi qu'à pLI (Lek et al., 2016) afin de comparer à la fois leur capacité à prédire les gènes intolérant aux variation en se basant sur des listes de gènes fournit par *the human phenotype ontology* (HPO) (Köhler et al., 2014) mais aussi en testant sa capacité à prédire les gènes considérés comme "dispensables" pour la vie et la reproduction humaine en se basant sur... .

4.2 Matériel & Méthodes

4.2.1 Récupération et filtrage des données

1. **Le jeu de transcrits Ensembl :** Pour cette étude, nous nous sommes basés sur la version 75 du jeu de transcrits fourni par Ensembl (B. L. Aken et al., 2017). Le fichier gtf contenant les données est téléchargeable ici. Cette version bien qu'elle ne soit pas la dernière publiée par Ensembl, est la dernière à se basée sur la version GRCH37/hg19 qui est la version du génome qu'a choisi ExAC pour effectuer l'alignement de ses données. À partir du fichier gtf, seul les transcrit tagués comme codant pour une protéine furent conservés, de même, l'ensemble des transcrits ayant une couverture médiane <15 sur plus de 30% de leur séquence codante dans les données ExAC furent filtrés.
2. **Filtrage des variants :**
 - a. Lien pour télécharger le vcf ExAC
 - b. L'ensemble des variants n'ayant pas la mention "PASS" dans la colonne FILTER du fichier VCF fourni par ExAC furent filtrés.
 - c. L'ensemble des variant n'ayant pas une couverture médiane ≥ 15 furent filtrés
 - d. L'ensemble des variants intronique (sauf ceux proches des sites d'épissage) et les variants situés dans les régions *upstream* et *downstream* furent filtrés

3. **Réannotassions des données ExAC** : Afin d'utiliser une version plus récente de VEP, l'annotation fut effectuée avec le logiciel VEP version 81 en utilisant la version 75 des transcrits Ensembl (INSÉRER LA COMMANDE)

4.2.2 Validation du score

1. Les gènes HPO :
2. Les gènes dispensables :

4.2.3 Les figures

L'ensemble des graphiques ont été générées en R version 3.3.3 (R Core Team, 2017) grâce au package ggplot2 (Wickham, 2009).

4.3 Résultats

4.3.1 Résultat de l'annotation

1. Tableau avec l'ensemble des csq et l'impact vep associée
2. Fréquence de ces impactes
3. Bar plot des poids

4.3.2 Détermination de la formule du score

Le SLAC et le WSLAC

Pour chaque transcrit nous avons défini deux métriques. La première est le *SLAC* (4.1) qui se définit comme étant pour transcrit, la somme du log des comptages allélique de chaque variant chevauchant ce transcrit. La seconde est le *WSLAC* (4.2) qui elle se définit comme étant la la somme **pondérée** du log des comptages allélique des variants chevauchant le transcrit.

\forall transcript $T \in \{\text{transcrit codant Ensembl}\}$:

$$SLAC_T = \sum_{v = \text{variant chevauchant } T} \log(\text{allele count}_v) \quad (4.1)$$

\forall transcript $T \in \{\text{transcrit codant Ensembl}\}$:

$$WSLAC_T = \sum_{I=Impact} \sum_{v=variant} Poid_I \cdot \log(\text{allele count}_v) \quad (4.2)$$

1. Le SLAC et le WSLAC
 - a. Formule du SLAC et du WSLAC
 - b. Graphique SLAC x WSLAC avec régression linéaire
 - c. Discussion sur la forme du graphique
2. Calcule de l'offset (décalage de l'origine)
 - a. But de l'offset
 - b. Graphique montrant l'évolution de la corrélation CDS-score en fonction de l'offset

4.3.3 Analyse du score

1. distribution du ratio (histo)
2. Analyse du top / bottom 50
 - a. Pie chart contribution moyenne des 4 impacts
 - b. Analyse panther (les pathway + expression différentielle)
3. Variance entre les différents transcrits d'un même gène
 - a. Histo de la variance
 - b. Discussion des gènes ayant la plus haute variance (intérêt de regarder le score par transcript plutôt que par gène)

4.4 Comparaison avec RVIS et pLI

1. Corrélation score~size
2. Hpo
3. Gene dispensable
 - a. Liste des 240 gènes
 - b. Recepteurs olfactifs

4.5 Conclusion

Conclusion

Annexe A

The First Appendix

In the main Rmd file

In Chapter ?? :

Annexe B

The Second Appendix, for Fun

References

- Adelman, M. M., & Cahill, E. M. (1989). *Atlas of sperm morphology* (p. 123). ASCP Press.
- 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>
- Aitken, R. J., Sutton, M., Warner, P., & Richardson, D. W. (1985). Relationship between the movement characteristics of human spermatozoa and their ability to penetrate cervical mucus and zona-free hamster oocytes. *Journal of Reproduction and Fertility*, 73(2), 441–9. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3989795>
- Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., ... Fllice, P. (2017). Ensembl 2017. *Nucleic Acids Research*, 45(D1), D635–D642. <http://doi.org/10.1093/nar/gkw1104>
- Alkan, C., Kidd, J. M., Marques-bonet, T., Aksay, G., Hormozdiari, F., Kitzman, J. O., ... Eichler, E. E. (2010). Personalized Copy-Number and Segmental Duplication Maps using Next-Generation Sequencing. *Nature Genetics*, 41(10), 1061–1067. <http://doi.org/10.1038/ng.437>. Personalized
- 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>
- Amdani, S. N., Jones, C., & Coward, K. (2013). Phospholipase C zeta (PLC ζ) : Oocyte activation and clinical links to male factor infertility. *Advances in Biological Regulation*, 53(3), 292–308. <http://doi.org/10.1016/j.jbior.2013.07.005>
- Amiri-Yekta, A., Coutton, C., Kherraf, Z.-E., Karaouzène, T., Le Tanno, P., Sanati, M. H., ... Ray, P. F. (2016). Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new *DNAH1* mutations. *Human Reproduction*, 31(12), 2872–2880. <http://doi.org/10.1093/humrep/dew280>

//doi.org/10.1093/humrep/dew262

Asimakopoulos, B. (2003). Is There a Place for Round and Elongated Spermatids Injection in, 1(1), 1–6.

Auffray, C., Chen, Z., & Hood, L. (2009). Systems medicine : the future of medical genomics and healthcare. *Genome Medicine*, 1(1), 2. <http://doi.org/10.1186/gm2>

Baes, C. F., Dolezal, M. A., Koltes, J. E., Bapst, B., Fritz-Waters, E., Jansen, S., ... Gredler, B. (2014). Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. *BMC Genomics*, 15(1), 948. <http://doi.org/10.1186/1471-2164-15-948>

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., ... Eichler, E. E. (2002). Recent Segmental Duplications in the Human Genome. *Science*, 297(5583), 1003–1007. <http://doi.org/10.1126/science.1072047>

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>

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., & Song, Y.-Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. *Journal of Human Genetics*, 56(May), 406–414. <http://doi.org/10.1038/jhg.2011.62>

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>

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics and Development*, 16(6), 545–552. <http://doi.org/10.1016/j.gde.2006.10.009>

Björndahl, L. (2010). The usefulness and significance of assessing rapidly progressive spermatozoa. *Asian Journal of Andrology*, 12(1), 33–5. <http://doi.org/10.1038/aja.2008.50>

Boer, P. de, Vries, M. de, & Ramos, L. (2015). A mutation study of sperm head shape and motility in the mouse : lessons for the clinic. *Andrology*, 3(2), 174–202. <http://doi.org/10.1111/andr.300>

Boivin, J., Bunting, L., Collins, J. A., & Nygren, K. G. (2007). International estimates of infertility prevalence and treatment-seeking : potential need and demand for infertility medical care. *Human Reproduction*, 22(6), 1506–1512. <http://doi.org/10.1093/humrep/dem046>

Bojesen, A., & Gravholt, C. H. (2011). Morbidity and mortality in Klinefelter syndrome

- (47,XXY). *Acta Paediatrica*, 100(6), 807–813. <http://doi.org/10.1111/j.1651-2227.2011.02274.x>
- Carson, A. R., Cheung, J., & Scherer, S. W. (2006). Duplication and relocation of the functional DPY19L2 gene within low copy repeats. *BMC Genomics*, 7, 45. <http://doi.org/10.1186/1471-2164-7-45>
- 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>
- Chemes, H. E., & Rawe, V. Y. (2010). The making of abnormal spermatozoa : cellular and molecular mechanisms underlying pathological spermiogenesis. *Cell and Tissue Research*, 341(3), 349–357. <http://doi.org/10.1007/s00441-010-1007-3>
- Chemes, H. E., Carizza, C., Scarinci, F., Brugo, S., Neuspiller, N., & Schwarsztein, L. (1987). Lack of a head in human spermatozoa from sterile patients : a syndrome associated with impaired fertilization. *Fertility and Sterility*, 47(2), 310–6. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3545911>
- Cheung, J., Estivill, X., Khaja, R., MacDonald, J. R., Lau, K., Tsui, L.-C., & Scherer, S. W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biology*, 4(4), R25. <http://doi.org/10.1186/gb-2003-4-4-r25>
- Cho, C., Willis, W. D., Goulding, E. H., Jung-Ha, H., Choi, Y. C., Hecht, N. B., & Eddy, E. M. (2001). Haploinsufficiency of protamine-1 or -2 causes infertility in mice. *Nature Genetics*, 28(1), 82–6. <http://doi.org/10.1038/88313>
- Choi, Y., Sims, G. E., Murphy, S., Miller, J. R., & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, 7(10). <http://doi.org/10.1371/journal.pone.0046688>
- 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>
- Clermont, Y. (1963). The cycle of the seminiferous epithelium in man. *American Journal of Anatomy*, 112(1), 35–51. <http://doi.org/10.1002/aja.1001120103>
- Clermont, Y. (1966). Renewal of spermatogonia in man. *American Journal of Anatomy*, 118(2), 509–524. <http://doi.org/10.1002/aja.1001180211>
- Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, 38(6), 1767–1771. <http://doi.org/10.1093/nar/gkp1137>
- Colgan, T. J., Bedard, Y. C., Strawbridge, H. T., Buckspan, M. B., & Klotz, P. G. (1980). Reappraisal of the Value of Testicular Biopsy in the Investigation of

- Infertility. *Fertility and Sterility*, 33(1), 56–60. [http://doi.org/10.1016/S0015-0282\(16\)44479-1](http://doi.org/10.1016/S0015-0282(16)44479-1)
- Collins, F. S., Morgan, M., & Patrinos, A. (2003). The Human Genome Project : Lessons from Large-Scale Biology. *Science*, 300(5617), 286–290. <http://doi.org/10.1126/science.1084564>
- Cooper, T. G., Noonan, E., Eckardstein, S. von, Auger, J., Baker, H. W. G., Behre, H. M., ... Vogelsong, K. M. (2010). World Health Organization reference values for human semen characteristics. *Human Reproduction Update*, 16(3), 231–245. <http://doi.org/10.1093/humupd/dmp048>
- Coutton, C., Escoffier, J., Martinez, G., Arnoult, C., & Ray, P. F. (2015). Teratozoospermia : spotlight on the main genetic actors in the human. *Human Reproduction Update*, 21(4), 455–485. <http://doi.org/10.1093/humupd/dmv020>
- Dam, A. H. D. M., Koscinski, I., Kremer, J. A. M., Moutou, C., Jaeger, A.-S., Oudakker, A. R., ... Viville, S. (2007). Homozygous mutation in SPATA16 is associated with male infertility in human globozoospermia. *American Journal of Human Genetics*, 81(4), 813–20. <http://doi.org/10.1086/521314>
- Dam, A. H., Koscinski, I., Kremer, J. A., Moutou, C., Jaeger, A.-S., Oudakker, A. R., ... Viville, S. (2007). Homozygous Mutation in SPATA16 Is Associated with Male Infertility in Human Globozoospermia. *The American Journal of Human Genetics*, 81(4), 813–820. <http://doi.org/10.1086/521314>
- Dam, A., Feenstra, I., Westphal, J., Ramos, L., Golde, R. van, & Kremer, J. (2006). Globozoospermia revisited. *Human Reproduction Update*, 13(1), 63–75. <http://doi.org/10.1093/humupd/dml047>
- 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>
- Dieterich, K., Soto Rifo, R., Faure, A. K., Hennebicq, S., Ben Amar, B., Zahi, M., ... Ray, P. F. (2007). Homozygous mutation of AURKC yields large-headed polyplloid spermatozoa and causes male infertility. *Nature Genetics*, 39(5), 661–5. <http://doi.org/10.1038/ng.2027>
- Ding, X., Xu, R., Yu, J., Xu, T., Zhuang, Y., & Han, M. (2007). SUN1 Is Required for Telomere Attachment to Nuclear Envelope and Gametogenesis in Mice. *Developmental Cell*, 12(6), 863–872. <http://doi.org/10.1016/j.devcel.2007.03.018>
- Eddy, E. M. (2007). The scaffold role of the fibrous sheath. *Society of Reproduction and Fertility Supplement*, 65, 45–62. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/17644954>
- ElInati, E., Kuentz, P., Redin, C., Jaber, S., Vanden Meerschaut, F., Makarian,

- J., ... Viville, S. (2012). Globozoospermia is mainly due to DPY19L2 deletion via non-allelic homologous recombination involving two recombination hotspots. *Human Molecular Genetics*, 21(16), 3695–3702. <http://doi.org/10.1093/hmg/dds200>
- Elliott, D. J., & Cooke, H. J. (1997). The molecular genetics of male infertility. *BioEssays*, 19(9), 801–809. <http://doi.org/10.1002/bies.950190910>
- Escalier, D., Gallo, J. M., Albert, M., Meduri, G., Bermudez, D., David, G., & Schrevel, J. (1991). Human acrosome biogenesis : immunodetection of proacrosin in primary spermatocytes and of its partitioning pattern during meiosis. *Development (Cambridge, England)*, 113(3), 779–788. Retrieved from <http://dev.biologists.org/content/develop/113/3/779.full.pdf>
- Escoffier, J., Lee, H. C., Yassine, S., Zouari, R., Martinez, G., Karaouzène, T., ... Arnoult, C. (2016). Homozygous mutation of PLCZ1 leads to defective human oocyte activation and infertility that is not rescued by the WW-binding protein PAWP. *Human Molecular Genetics*, 25(5), 878–91. <http://doi.org/10.1093/hmg/ddv617>
- Flicek, P., & Birney, E. (2009). Sense from sequence reads : methods for alignment and assembly. *Nature Methods*, 6(11 Suppl), S6–S12. <http://doi.org/10.1038/nmeth0610-479b>
- Gekas, J., Thepot, F., Turleau, C., Siffroi, J. P., Dadoune, J. P., Briault, S., ... Association des Cytogeneticiens de Langue Francaise. (2001). Chromosomal factors of infertility in candidate couples for ICSI : an equal risk of constitutional aberrations in women and men. *Human Reproduction (Oxford, England)*, 16(1), 82–90. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11139542>
- Girgis, S. M., Etriby, A. N., Ibrahim, A. A., & Kahil, S. A. (1969). Testicular biopsy in azoospermia. A review of the last ten years' experiences of over 800 cases. *Fertility and Sterility*, 20(3), 467–77. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/5769396>
- Gnessi, L., Fabbri, A., & Spera, G. (1997). Gonadal peptides as mediators of development and functional control of the testis : An integrated system with hormones and local environment. *Endocrine Reviews*, 18(4), 541–609. <http://doi.org/10.1210/er.18.4.541>
- Gonzaga-Jauregui, C., Lupski, J. R., & Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annual Review of Medicine*, 63, 35–61. <http://doi.org/10.1146/annurev-med-051010-162644>
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age : ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6), 333–351. <http://doi.org/10.1038/nrg.2016.49>
- Goossens, E., & Tournaye, H. (2013). Adult stem cells in the human testis. *Semi-*

- nars in Reproductive Medicine, 31(1), 39–48. <http://doi.org/10.1055/s-0032-1331796>
- Guffanti, A., Iacono, M., Pelucchi, P., Kim, N., Soldà, G., Croft, L. J., ... De Bellis, G. (2009). A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics*, 10(1), 163. <http://doi.org/10.1186/1471-2164-10-163>
- Guo, J., Xu, N., Li, Z., Zhang, S., Wu, J., Kim, D. H., ... Ju, J. (2008). Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 105(27), 9145–9150. <http://doi.org/10.1073/pnas.0804023105>
- Hamilton, D. W., Waites, G. M. H. (1990). *Cellular and Molecular Events in Spermiogenesis* (p. 334). Cambridge University Press. Retrieved from <http://www.cambridge.org/us/academic/subjects/medicine/obstetrics-and-gynecology-reproductive-medicine/cellular-and-molecular-events-spermiogenesis>
- Handyside, A. H. (2012). Molecular origin of female meiotic aneuploidies. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1822(12), 1913–1920. <http://doi.org/10.1016/j.bbadi.2012.07.007>
- Harbuz, R., Zouari, R., Pierre, V., Ben Khelifa, M., Kharouf, M., Coutton, C., ... Ray, P. F. (2011). A recurrent deletion of DPY19L2 causes infertility in man by blocking sperm head elongation and acrosome formation. *American Journal of Human Genetics*, 88(3), 351–61. <http://doi.org/10.1016/j.ajhg.2011.02.007>
- Hermo, L., Pelletier, R. M., Cyr, D. G., & Smith, C. E. (2010). Surfing the wave, cycle, life history, and genes/proteins expressed by testicular germ cells. Part 3 : Developmental changes in spermatid flagellum and cytoplasmic droplet and interaction of sperm with the zona pellucida and egg plasma membrane. *Microscopy Research and Technique*, 73(4), 320–363. <http://doi.org/10.1002/jemt.20784>
- Heytens, E., Parrington, J., Coward, K., Young, C., Lambrecht, S., Yoon, S.-Y., ... De Sutter, P. (2009). Reduced amounts and abnormal forms of phospholipase C zeta (PLCzeta) in spermatozoa from infertile men. *Human Reproduction (Oxford, England)*, 24(10), 2417–28. <http://doi.org/10.1093/humrep/dep207>
- Holstein, A. F., Schirren, C., & Schirren, C. G. (1973). Human spermatids and spermatozoa lacking acrosomes. *Journal of Reproduction and Fertility*, 35(3), 489–91. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/4760149>
- Horner, D. S., Pavese, G., Castrignano', T., Meo, P. D. O. de, Liuni, S., Sammeth, M., ... Pesole, G. (2009). Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. *Briefings in Bioinformatics*, 11(2),

- 181–197. <http://doi.org/10.1093/bib/bbp046>
- Hotaling, J., & Carrell, D. T. (2014). Clinical genetic testing for male factor infertility : current applications and future directions. *Andrology*, 2(3), 339–350. <http://doi.org/10.1111/j.2047-2927.2014.00200.x>
- 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>
- Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5(December), 17875. <http://doi.org/10.1038/srep17875>
- 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>
- Inaba, K. (2003). Molecular Architecture of the Sperm Flagella : Molecules for Motility and Signaling. *Zoological Science*, 20(9), 1043–1056. <http://doi.org/10.2108/zsj.20.1043>
- JOHNSON, L., PETTY, C. S., & NEAVES, W. B. (1980). A Comparative Study of Daily Sperm Production and Testicular Composition in Humans and Rats. *Biol Reprod*, 22(5), 1233–1243. Retrieved from <http://www.biolreprod.org/content/22/5/1233.short>
- KIERSZENBAUM, A. L. (1994). Mammalian Spermatogenesis *<in Vivo>* and *<in Vitro>* : A Partnership of Spermatogenic and Somatic Cell Lineages*. *Endocrine Reviews*, 15(1), 116–134. <http://doi.org/10.1210/edrv-15-1-116>
- Kierszenbaum, A. L., & Tres, L. L. (1978). RNA transcription and chromatin structure during meiotic and postmeiotic stages of spermatogenesis. *Federation Proceedings*, 37(11), 2512–6. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/357185>
- Kierszenbaum, A. L., & Tres, L. L. (2004). The acrosome-acroplaxome-manchette complex and the shaping of the spermatid head. *Archives of Histology and Cytology*, 67(4), 271–84. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/15700535>
- Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., ... Snyder, M. (2009). Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *October*, 318(5849), 420–426. <http://doi.org/10.1126/science.1149504.Paired-End>
- Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., ... Robinson, P. N. (2014). The Human Phenotype Ontology project : linking

- molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(Database issue), D966–74. <http://doi.org/10.1093/nar/gkt1026>
- Krausz, C., & Forti, G. (2000). Clinical aspects of male infertility. *Results and Problems in Cell Differentiation*, 28, 1–21. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/10626292>
- 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>
- Kurilo, L. F., Liubashevskaya, I. A., Dubinskaia, V. P., & Gaeva, T. N. (1993). [Karyological analysis of the count of immature germ cells in the ejaculate]. *Urologii I Nefrologii*, (2), 45–7. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/7941145>
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. <http://doi.org/10.1038/nmeth.1923>
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), R25. <http://doi.org/10.1186/gb-2009-10-3-r25>
- 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>
- Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. a., & Gilissen, C. (2015). Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human Mutation*, 36(8), 815–22. <http://doi.org/10.1002/humu.22813>
- Levin, H. S. (1979). Testicular biopsy in the study of male infertility. *Human Pathology*, 10(5), 569–584. [http://doi.org/10.1016/S0046-8177\(79\)80100-8](http://doi.org/10.1016/S0046-8177(79)80100-8)
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <http://doi.org/10.1093/bioinformatics/btp352>
- Li, H., Ruan, J., Durbin, R., Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Mapping short DNA sequencing reads and calling variants using mapping quality scores*,

- 1851–1858. <http://doi.org/10.1101/gr.078212.108>
- Lindholmer, C. (1974). The importance of seminal plasma for human sperm motility. *Biology of Reproduction*, 10(5), 533–42. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/4142752>
- Lu, L., Lin, M., Xu, M., Zhou, Z.-M., & Sha, J.-H. (2006). Gene functional research using polyethylenimine-mediated in vivo gene transfection into mouse spermatogenic cells. *Asian Journal of Andrology*, 8(1), 53–59. <http://doi.org/10.1111/j.1745-7262.2006.00089.x>
- 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>
- MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., & Scherer, S. W. (2014). The Database of Genomic Variants : a curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(Database issue), D986–92. <http://doi.org/10.1093/nar/gkt958>
- MacLeod, J. (1970). The Significance of Deviations in Human Sperm Morphology. In (pp. 481–494). Springer US. http://doi.org/10.1007/978-1-4615-9008-8_35
- Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133–141. <http://doi.org/10.1016/j.tig.2007.12.007>
- McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. a, Gaulton, K., Cazier, J.-B., & Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3), 26. <http://doi.org/10.1186/gm543>
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., ... DePristo, M. A. (2010). The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297–303. <http://doi.org/10.1101/gr.107524.110>
- 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>
- McPherson, J. D. (2009). Next-generation gap. *Nature Methods*, 6(11s), S2–S5. <http://doi.org/10.1038/nmeth.f.268>
- Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11s), S13–S20. <http://doi.org/10.1038/nmeth.1374>
- Meienberg, J., Bruggmann, R., Oexle, K., & Matyas, G. (2016). Clinical sequencing : is WGS the better WES ? *Human Genetics*, 135(3), 359–362. <http://doi.org/>

10.1007/s00439-015-1631-9

- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, 11(1), 31–46. <http://doi.org/10.1038/nrg2626>
- Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2017). PANTHER version 11 : expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1), D183–D189. <http://doi.org/10.1093/nar/gkw1138>
- Michael, M., & Joel, K. (1937). Zellformen in normalen und pathologischen Ejakulaten und ihre klinische Bedeutung. *Schweiz. Med. Wsch.* Retrieved from <https://scholar.google.com/scholar?cluster=6307038842480257282&hl=en&oi=scholarr>
- Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Abigail, W., Lee, C., ... Shendure, J. (2010). Targeted Capture and Massicely Parallel Sequencing of twelve human exomes. *Nature*, 461(7261), 272–276. <http://doi.org/10.1038/nature08250>. Targeted
- 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>
- Nistal, M., Paniagua, R., & Herruzo, A. (1978). Multi-tailed spermatozoa in a case with asthenospermia and teratospermia. *Virchows Archiv B*, 26(1), 111–118. <http://doi.org/10.1007/bf02889540>
- Nomikos, M., Kashir, J., Swann, K., & Lai, F. A. (2013). Sperm PLC ζ : From structure to Ca $²⁺$ oscillations, egg activation and therapeutic potential. *FEBS Letters*, 587(22), 3609–3616. <http://doi.org/10.1016/j.febslet.2013.10.008>
- Ogura, a., Matsuda, J., & Yanagimachi, R. (1994). Birth of normal young after electrofusion of mouse oocytes with round spermatids. *Proceedings of the National Academy of Sciences of the United States of America*, 91(16), 7460–7462. <http://doi.org/10.1073/pnas.91.16.7460>
- Ogura, A., Matsuda, J., Asano, T., Suzuki, O., & Yanagimachi, R. (1996). Mouse oocytes injected with cryopreserved round spermatids can develop into normal offspring. *Journal of Assisted Reproduction and Genetics*, 13(5), 431–434. <http://doi.org/10.1007/BF02066177>
- Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg : Springer Berlin Heidelberg. <http://doi.org/10.1007/978-3-642-86659-3>
- O'Flynn O'Brien, K. L., Varghese, A. C., & Agarwal, A. (2010). The genetic causes of male factor infertility : A review. *Fertility and Sterility*, 93(1), 1–12. <http://doi.org/10.1016/j.fertnstert.2009.10.045>
- O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., ... Lyon, G. J. (2013).

- Low concordance of multiple variant-calling pipelines : practical implications for exome and genome sequencing. *Genome Medicine*, 5(3), 28. <http://doi.org/10.1186/gm432>
- Palermo, G., Joris, H., Devroey, P., & Van Steirteghem, A. C. (1992). Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. *Lancet (London, England)*, 340(8810), 17–8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/1351601>
- Panidis, D., Rousso, D., Kourtis, A., Gianoulis, C., Papathanasiou, K., & Kalachanis, J. (2001). Headless spermatozoa in semen specimens from fertile and subfertile men. *The Journal of Reproductive Medicine*, 46(11), 947–50. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11762149>
- Papic, Z., Katona, G., & Skrabalo, Z. (1988). The cytologic identification and quantification of testicular cell subtypes. Reproducibility and relation to histologic findings in the diagnosis of male infertility. *Acta Cytologica*, 32(5), 697–706. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3421018>
- Pedersen, H., & Rebbe, H. (1974). Fine structure of round-headed human spermatozoa. *Journal of Reproduction and Fertility*, 37(1), 51–4. <http://doi.org/10.1530/JRF.0.0370051>
- Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., Goldstein, D. B., Davydov, E., ... Lisacek, F. (2013). Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics*, 9(8), e1003709. <http://doi.org/10.1371/journal.pgen.1003709>
- Pierre, V., Martinez, G., Coutton, C., Delaroche, J., Yassine, S., Novella, C., ... Arnoult, C. (2012). Absence of Dpy19l2, a new inner nuclear membrane protein, causes globozoospermia in mice by preventing the anchoring of the acrosome to the nucleus. *Development*, 139(16), 2955–2965. <http://doi.org/10.1242/dev.077982>
- Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., ... Lipman, D. (2009). The consensus coding sequence (CCDS) project : Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Research*, 19(7), 1316–1323. <http://doi.org/10.1101/gr.080531.108>
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, S., Manichanh, C., ... Yang, H. (2010). A human gut microbial gene catalog established by metagenomic sequencing. *Nature*, 464(7285), 59–65. <http://doi.org/10.1038/nature08821.A>
- R Core Team. (2017). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org/>
- Ravel, C., Berthaut, I., Bresson, J. L., Siffroi, J. P., & Genetics Commission of the French Federation of CECOS. (2006). Prevalence of chromosomal abnormalities in

- phenotypically normal and fertile adult males : large-scale survey of over 10 000 sperm donor karyotypes. *Human Reproduction*, 21(6), 1484–1489. <http://doi.org/10.1093/humrep/de1024>
- Ray, P. F., & Arnoult, C. (2011). La délétion homozygote du gène <i>DPY19L2</i> est responsable de la majorité des cas de globozoospermie. *Médecine/Sciences*, 27(8-9), 692–693. <http://doi.org/10.1051/medsci/2011278004>
- 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>
- Rosenfeld, J. A., Mason, C. E., Smith, T. M., Wallin, C., & Diekhans, M. (2012). Limitations of the Human Reference Genome for Personalized Genomics. *PLoS ONE*, 7(7), e40294. <http://doi.org/10.1371/journal.pone.0040294>
- Ruffalo, M., Laframboise, T., & Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20), 2790–2796. <http://doi.org/10.1093/bioinformatics/btr477>
- 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>
- Sasagawa, I., & Yanagimachi, R. (1997). Spermatids from mice after cryptorchid and reversal operations can initiate normal embryo development. *Journal of Andrology*, 18(2), 203–209. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9154515>
- Schatz, M. C., & Langmead, B. (2013). The DNA Data Deluge : Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. *IEEE Spectrum*, 50(7), 26–33. <http://doi.org/10.1109/MSPEC.2013.6545119>
- Schenck, U., & Schill, W. B. (n.d.). Cytology of the human seminiferous epithelium. *Acta Cytologica*, 32(5), 689–96. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3421017>
- Sen, C. G. S., Holstein, A. F., & Schirren, C. (1971). über die Morphogenese rundköpfiger Spermatozoen des Menschen. *Andrologia*, 3(3), 117–125. <http://doi.org/10.1111/j.1439-0272.1971.tb02106.x>
- Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage : key considerations in genomic analyses. *Nature Reviews. Genetics*, 15(2), 121–32. <http://doi.org/10.1038/nrg3642>
- Singh, G. (n.d.). Ultrastructural features of round-headed human spermatozoa. *International Journal of Fertility*, 37(2), 99–102. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/3421017>

- [nlm.nih.gov/pubmed/1349598](http://www.ncbi.nlm.nih.gov/pubmed/1349598)
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., ... Page, D. C. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942), 825–837. <http://doi.org/10.1038/nature01722>
- Soderström, K. O., & Suominen, J. (1980). Histopathology and ultrastructure of meiotic arrest in human spermatogenesis. *Archives of Pathology & Laboratory Medicine*, 104(9), 476–82. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/6893401>
- SPERLING, K., & KADEN, R. (1971). Meiotic Studies of the Ejaculated Seminal Fluid of Humans with Normal Sperm Count and Oligospermia. *Nature*, 232(5311), 481–481. <http://doi.org/10.1038/232481a0>
- Stitzel, N. O., Kiezun, A., & Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biology*, 12(9), 227. <http://doi.org/10.1186/gb-2011-12-9-227>
- 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>
- Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., ... Yaspo, M.-L. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891), 956–960. <http://doi.org/10.1126/science.1160342>
- Tanaka, A., Nagayoshi, M., Takemoto, Y., Tanaka, I., Kusunoki, H., Watanabe, S., ... Yanagimachi, R. (2015). Fourteen babies born after round spermatid injection into human oocytes. *Proceedings of the National Academy of Sciences*, 112(March 2014), 201517466. <http://doi.org/10.1073/pnas.1517466112>
- Taylor, K. H., Kramer, R. S., Davis, J. W., Guo, J., Duff, D. J., Xu, D., ... Shi, H. (2007). Ultradeep Bisulfite Sequencing Analysis of DNA Methylation Patterns in Multiple Gene Promoters by 454 Sequencing. *Cancer Research*, 67(18), 8511–8518. <http://doi.org/10.1158/0008-5472.CAN-07-1016>
- Taylor, S., Yoon, S., Morshedi, M., Lacey, D., Jellerette, T., Fissore, R., & Oehninger, S. (2010). Complete globozoospermia associated with PLC ζ deficiency treated with calcium ionophore and ICSI results in pregnancy. *Reproductive BioMedicine Online*, 20(4), 559–564. <http://doi.org/10.1016/j.rbmo.2009.12.024>
- Thankaswamy-Kosalai, S., Sen, P., & Nookae, I. (2017). Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics. *Genomics*. <http://doi.org/10.1016/j.ygeno.2017.03.001>

2017.03.001

- Tomkinson, A. E., Vijayakumar, S., Pascal, J. M., & Ellenberger, T. (2006). DNA Ligases : Structure, Reaction Mechanism, and Function. *Chemical Reviews*, 106(2), 687–699. <http://doi.org/10.1021/cr040498d>
- Tomlinson, M. J., Barratt, C. L. R., & Cooke, I. D. (1993). Prospective study of leukocytes and leukocyte subpopulations in semen suggests they are not a cause of male infertility**Supported by the Infertility Research Trust, and the University of Sheffield, Sheffield, United Kingdom (M.J.T.). *Fertility and Sterility*, 60(6), 1069–1075. [http://doi.org/10.1016/S0015-0282\(16\)56412-7](http://doi.org/10.1016/S0015-0282(16)56412-7)
- Tomlinson, M., Barrati, C., Bolton, A., Lenton, E., Roberts, H., & Cooke, I. (1993). Round cells and sperm fertilizing capacity : The presence of immature germ cells but not seminal leukocytes are associated with reduced success of in vitro fertilization. *International Journal of Gynecology & Obstetrics*, 42(2), 223–224. [http://doi.org/10.1016/0020-7292\(93\)90672-J](http://doi.org/10.1016/0020-7292(93)90672-J)
- Tranchevent, L.-C., Ardesthirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., & Moreau, Y. (2016). Candidate gene prioritization with Endeavour. *Nucleic Acids Research*, 44(W1), W117–W121. <http://doi.org/10.1093/nar/gkw365>
- Treangen, T. J., & Salzberg, S. L. (2013). Repetitive DNA and next-generation sequencing : computational challenges and solutions. *Nat Rev Genet.*, 13(1), 36–46. <http://doi.org/10.1038/nrg3117.Repetitive>
- Tüttelmann, F., Simoni, M., Kliesch, S., Ledig, S., Dworniczak, B., Wieacker, P., & Röpke, A. (2011). Copy number variants in patients with severe oligozoospermia and Sertoli-cell-only syndrome. *PLoS One*, 6(4), e19426. <http://doi.org/10.1371/journal.pone.0019426>
- Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., ... Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, 5(3), 247–252. <http://doi.org/10.1038/nmeth.1185>
- Vorona, E., Zitzmann, M., Gromoll, J., Schüring, A. N., & Nieschlag, E. (2007). Clinical, Endocrinological, and Epigenetic Features of the 46,XX Male Syndrome, Compared with 47,XXY Klinefelter Patients. *The Journal of Clinical Endocrinology & Metabolism*, 92(9), 3458–3465. <http://doi.org/10.1210/jc.2007-0447>
- Walsh, B. (2003). Population-genetic models of the fates of duplicate genes. *Genetica*, 118(2-3), 279–94. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12868616>
- Wang, X., Jin, H., Han, F., Cui, Y., Chen, J., Yang, C., ... Gao, Z. (2017). Homozygous *< i>DNAH1</i>* frameshift mutation causes multiple morphological anomalies of the sperm flagella in Chinese. *Clinical Genetics*, 91(2), 313–321. <http://doi.org/>

- 10.1111/cge.12857
- Ward, W. S. (1994). The structure of the sleeping genome : implications of sperm DNA organization for somatic cells. *Journal of Cellular Biochemistry*, 55(1), 77–82. <http://doi.org/10.1002/jcb.240550109>
- Wickham, H. (2009). *ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from <http://ggplot2.org>
- Wold, B., & Myers, R. M. (2007). Sequence census methods for functional genomics. *Nature Methods*, 5(1), 19–21. <http://doi.org/10.1038/nmeth1157>
- WONG, T.-W., STRAUS, F. H. I., & WARNER, N. E. (1973). TESTICULAR BIOPSY IN THE STUDY OF MALE INFERTILITY : II. POST... : Obstetrical & Gynecological Survey. *Obstetrical & Gynecological Survey*, 28(9), 660–661. Retrieved from http://journals.lww.com/obgynsurvey/Citation/1973/09000/TESTICULAR{_}BIOPSY{_}IN
- World Health Organization. (1992). *WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction*. (3th ed, p. 128). Cambridge University Press.
- Yang, M. Q., Athey, B. D., Arabnia, H. R., Sung, A. H., Liu, Q., Yang, J. Y., ... Deng, Y. (2009). High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. *BMC Genomics*, 10 Suppl 1, I1. <http://doi.org/10.1186/1471-2164-10-S1-I1>
- Yassine, S., Escoffier, J., Abi Nahed, R., Nahed, R. A., Pierre, V., Karaouzene, T., ... Arnoult, C. (2015). Dynamics of Sun5 localization during spermatogenesis in wild type and Dpy19l2 knock-out mice indicates that Sun5 is not involved in acrosome attachment to the nuclear envelope. *PloS One*, 10(3), e0118698. <http://doi.org/10.1371/journal.pone.0118698>
- Yoon, S.-Y., Jellerette, T., Salicioni, A. M., Lee, H. C., Yoo, M.-S., Coward, K., ... Fissore, R. A. (2008). Human sperm devoid of PLC ζ fail to induce Ca(2+) release and are unable to initiate the first step of embryo development. *The Journal of Clinical Investigation*, 118(11), 3671–81. <http://doi.org/10.1172/JCI36942>
- Yu, J., Chen, Z., Ni, Y., & Li, Z. (2012). CFTR mutations in men with congenital bilateral absence of the vas deferens (CBAVD) : a systemic review and meta-analysis. *Human Reproduction*, 27(1), 25–35. <http://doi.org/10.1093/humrep/der377>
- Zhao, S., & Zhang, B. (2015). A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC Genomics*, 16(1), 97. <http://doi.org/10.1186/s12864-015-1308-8>
- Zhu, F., Wang, F., Yang, X., Zhang, J., Wu, H., Zhang, Z., ... Cao, Y. (2016). Biallelic SUN5 Mutations Cause Autosomal-Recessive Acephalic Spermatozoa

Syndrome. *The American Journal of Human Genetics*, 99(4), 942–949. <http://doi.org/10.1016/j.ajhg.2016.08.004>