

# 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

<b>Remerciements</b> . . . . .	<b>1</b>
<b>Résumé</b> . . . . .	<b>3</b>
<b>Abstract</b> . . . . .	<b>5</b>
<b>Chapitre 1 : Introduction</b> . . . . .	<b>7</b>
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 . . . . .	17
1.2.1 Anatomie du spermatozoïde . . . . .	17
La tête . . . . .	17
Le flagelle . . . . .	19
1.2.2 Fonction du spermatozoïde . . . . .	21
1.3 L'infertilité masculine . . . . .	22
1.3.1 Les différents phénotypes d'infertilité masculine . . . . .	22
Liée à la quantité . . . . .	22
Liée à la morphologie . . . . .	23
Liée à la mobilité . . . . .	25
1.3.2 La génétique de l'infertilité . . . . .	25
Les causes fréquentes . . . . .	25
Les nouveaux gènes . . . . .	27
1.4 Les techniques d'analyses génétiques . . . . .	30
1.4.1 Approche “gènes candidats” . . . . .	30
1.4.2 Les puces . . . . .	30
Les puces à SNP, le génotypage... (titre à revoir) . . . . .	31
Du tissu au transcriptome, le différentiel d'expression . . . . .	31
1.4.3 Le séquençage NGS . . . . .	31
La capture des parties à séquencer, avantage et inconvenants . . . . .	32
L'amplification . . . . .	33
La réaction de séquence . . . . .	35
1.5 L'analyse bioinformatique des données de NGS . . . . .	39

1.5.1	Les données fournies par le NGS . . . . .	39
	Un <i>read</i> c'est quoi ? . . . . .	39
	Le format FASTQ . . . . .	40
1.5.2	L'alignement . . . . .	40
1.5.3	L'appel des variants . . . . .	41
1.5.4	L'annotation des variants, filtrage et priorisation . . . . .	43
1.5.5	Conclusion NGS . . . . .	45
<b>Chapitre 2 : Investigation génétique et physiologique de la globozoospermie</b>		<b>47</b>
<b>Chapitre 3 : MutaScript</b>		<b>49</b>
3.1	Introduction . . . . .	49
3.2	Matériel & Méthodes . . . . .	51
	3.2.1 Récupération et filtrage des données . . . . .	51
	3.2.2 Validation du score . . . . .	52
3.3	Résultats . . . . .	52
	3.3.1 Résultat de l'annotation . . . . .	52
	3.3.2 Détermination de la formule du score . . . . .	52
	3.3.3 Analyse du score . . . . .	53
3.4	Comparaison avec RVIS et pLI . . . . .	53
3.5	Conclusion . . . . .	53
<b>Conclusion</b>		<b>55</b>
<b>Annexe A : The First Appendix</b>		<b>57</b>
<b>Annexe B : The Second Appendix, for Fun</b>		<b>59</b>
<b>References</b>		<b>61</b>

# Liste des tableaux

1.1 Durée de vie moyenne des cellules germinales humaines . . . . .	7
---	---



# 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 . . . . .	14
1.7	Principales étapes et modifications structurales lors de la spermogénèse	16
1.8	Anatomie simplifiée du spermatozoïde . . . . .	17
1.9	Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes . . . . .	19
1.10	Structure simplifiée de l'axonème . . . . .	20
1.11	Structure du flagelle d'un spermatozoïde . . . . .	21
1.12	Différentes anomalies morphologiques du spermatozoïde selon la classification de David modifiée adapté... TABLEAU à adapter et à insérer!!!!	24
1.13	Représentation schématique du chromosome Y adapté . . . . .	26
1.14	Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée . . . . .	32
1.15	Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS . . . . .	35
1.16	Exemple de séquençage CRT tel qu'il est effectué par Illumina . . . .	36
1.17	Exemple de séquençage SNA tel qu'il est effectué par Ion Torrent . .	37
1.18	Exemple de séquençage SBL tel qu'il est effectué par SOLiD . . . .	38
1.19	présentation d'un fichier FASTQ (FIGURE A CHANGER) . . . . .	40
1.20	Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel . . . . .	42
1.21	Diagramme de Venn des prédictions de pathogénicités de six logiciels	44



# 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 continue 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 est d'environ 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**).

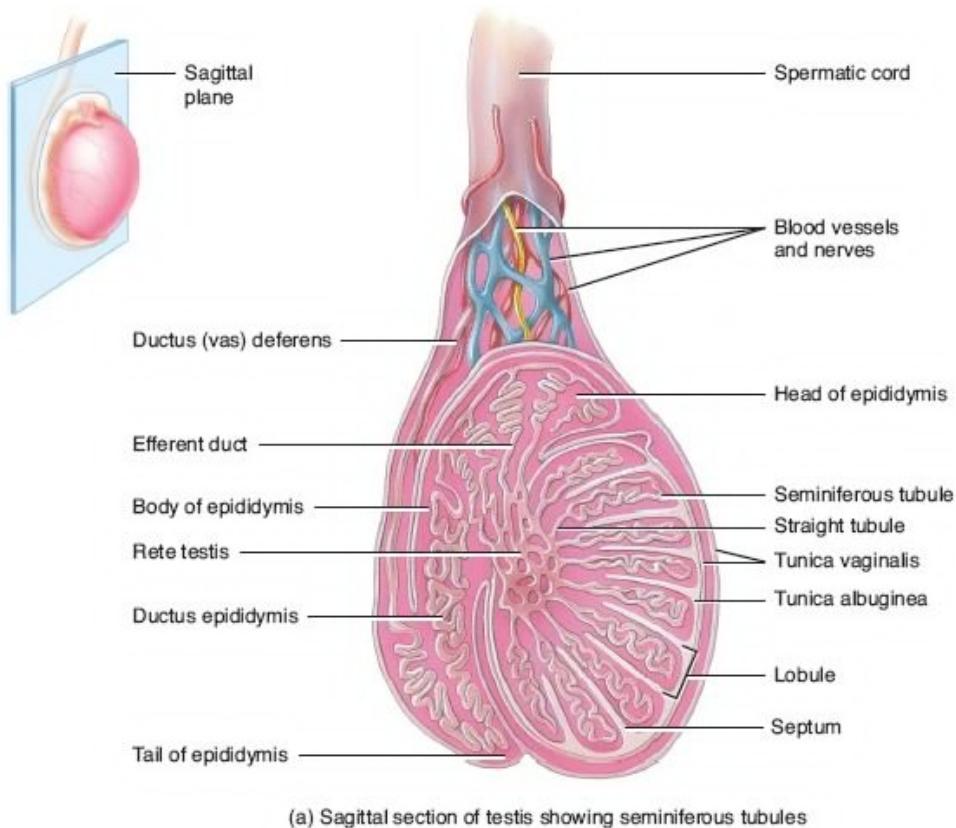
**Table 1.1 – Durée de vie moyenne des cellules germinales humaines**

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 cm<sup>3</sup>. 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.



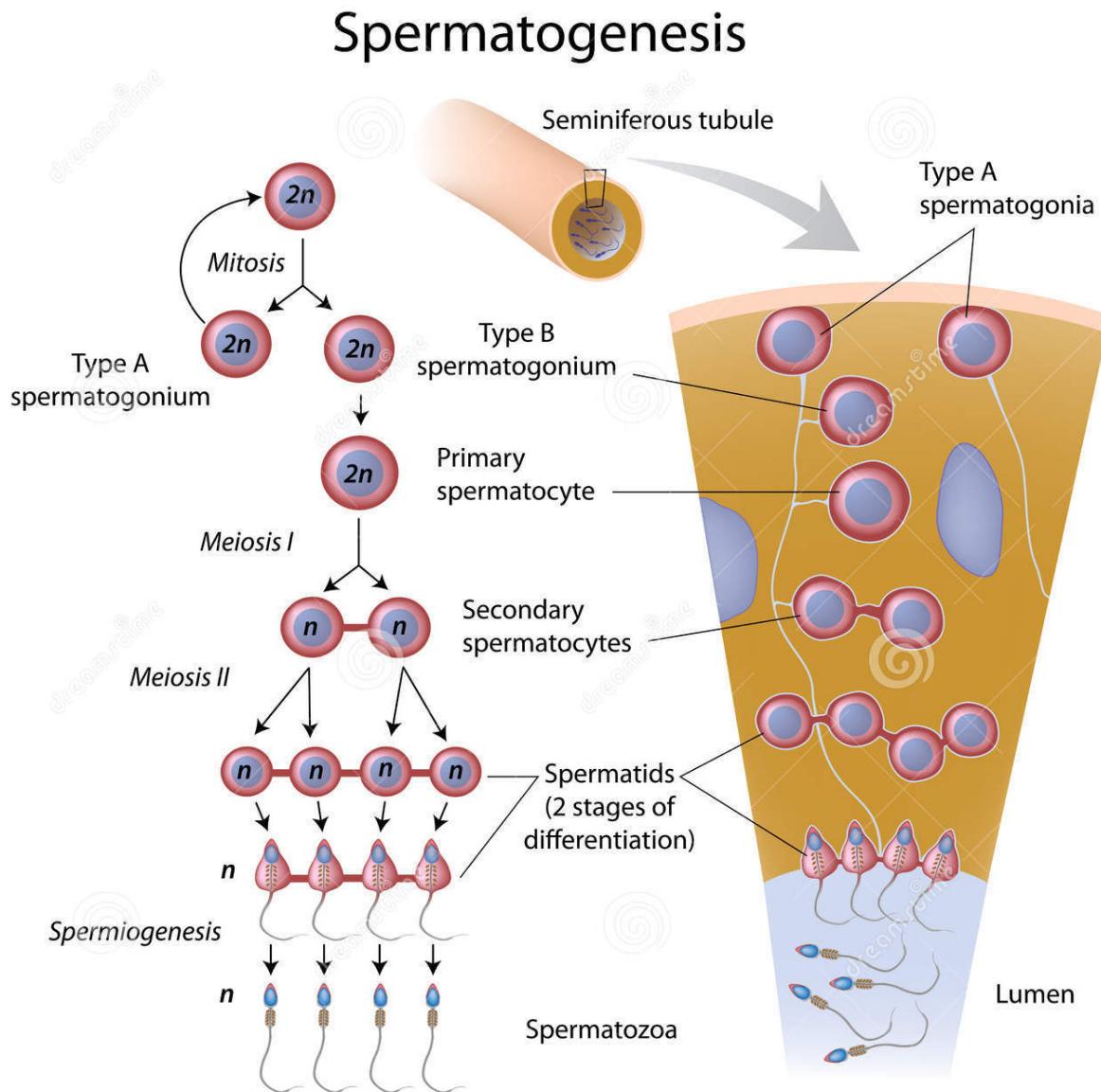
**Figure 1.1** – Schéma anatomique du testicule humain :

### 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 mitose successive 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 cellulaire 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**).

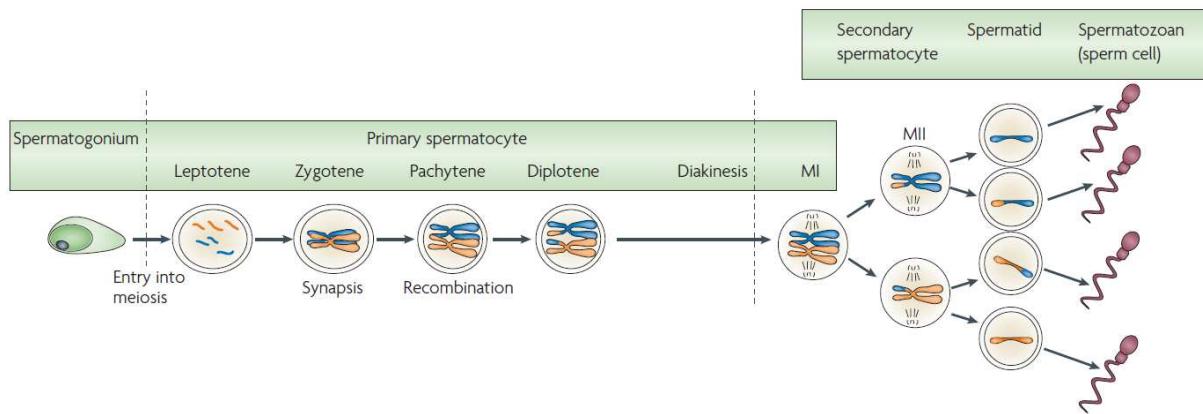


**Figure 1.2** – Les différentes phases de la spermatogénèse d'après [medizin-kompakt](<http://www.medizin-kompakt.de/spermatogenese>) : description à écrire!!!

### 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 secondaire (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**).



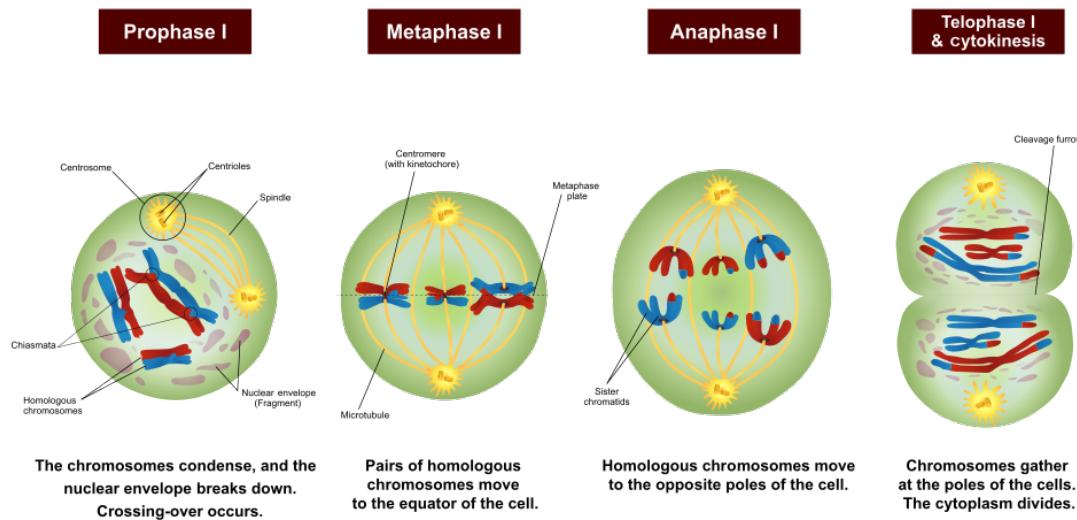
**Figure 1.3 – Les différentes étapes de la méiose gamétique masculine d'après Sasaki et Matsui, 2008**

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 à 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 reliés 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. Zygote :** Appariement des chromosomes homologues par paires appelées bivalents grâce l'intermédiaire d'une structure multi-protéique : le complexe synaptonémal.

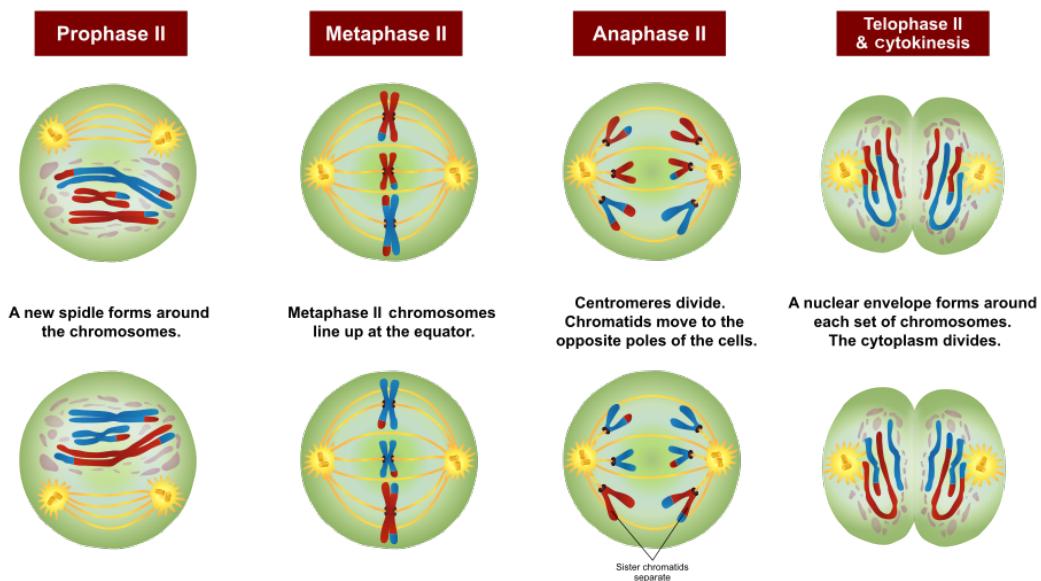
- iii. **Pachytène** : Ce stade dure 16 jours et est le plus long de la prophase I. C'est au cours de celui-ci qu'à lieu l'échange de matériel génétique par le biais des crossing-over (**Figure** : 1.6) entre les chromatides non-sœurs appelés nODULES de recombinaison.
  - iv. **Diplotène** : La dissociation du complexe synaptonémal va permettre aux chromosomes homologues d'initier leur séparation. Certains sites d'appariement étroits nommés chiasmas demeurent néanmoins 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>)

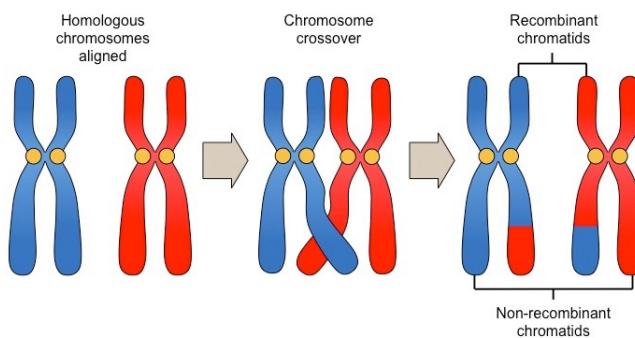
2. Méiose équationnelle : (Figure : 1.5) La MII est similaire à une division mitotique,

- 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 centromère de chaque chromosome se rompent 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

### 1.1.4 La spermiogénèse

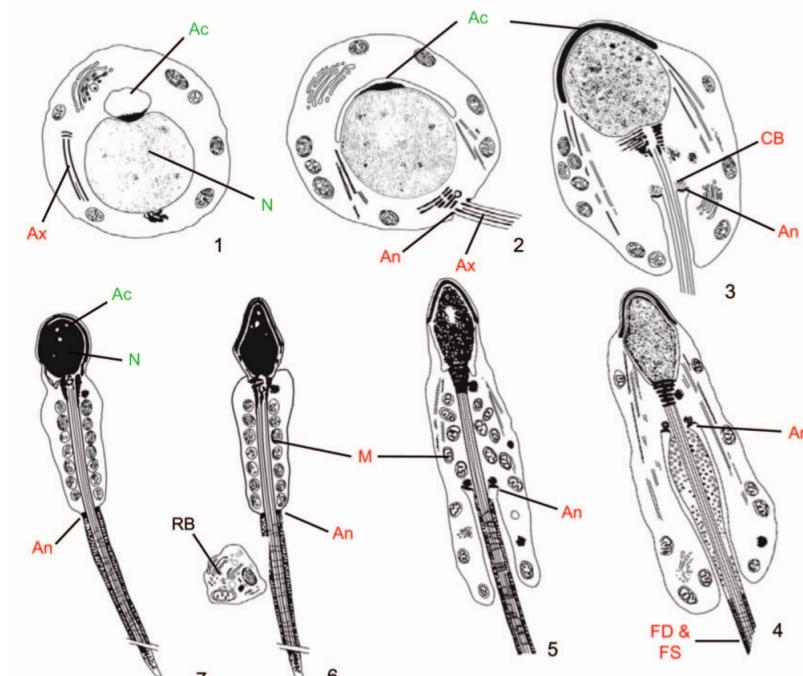
La spermiogénèse est la phase finale de la spermatogénèse. Elle dure environ 23 jours chez l'humain et peut être subdivisée en sept étapes (**Figure : 1.7**). La spermiogénèse définit la cytodifférenciation des spermatides en spermatozoïdes. C'est au cours de cette phase que les caractéristiques morphologique 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 commence 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, (à 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**) : peuvent aussi donner naissance avec un meilleur taux que les spermatides rondes et engendrerai 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 différenciation 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 spermiation.



**Figure 1.7 – Principales étapes et modifications structurales lors de la spermiogénèse d'après Touré et al., 2011 :** 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

## 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 déterminée. 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).

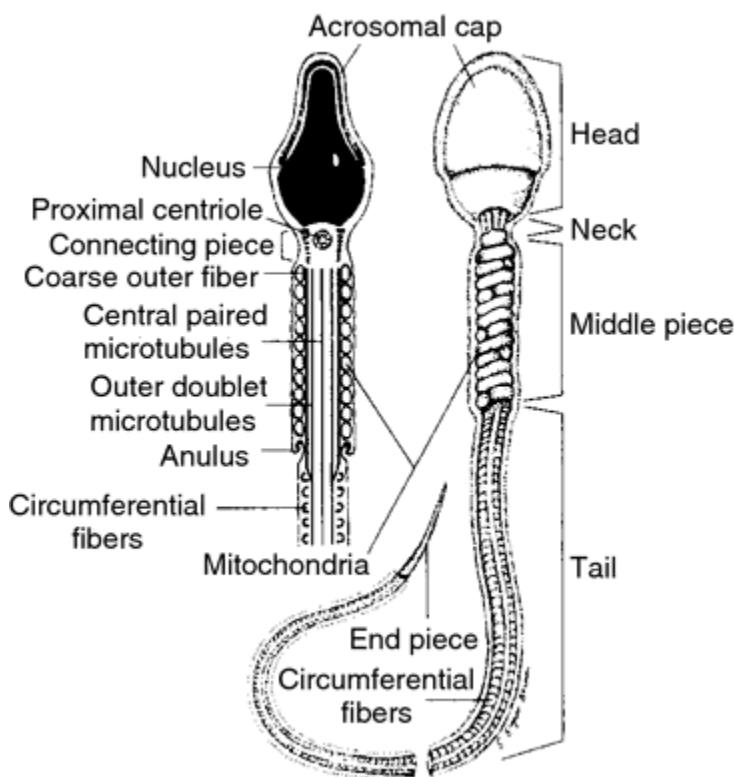


Figure 1.8 – Anatomie simplifiée du spermatozoïde

### 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'excréter notamment l'acrosine, une enzyme de digestion permettant au spermatozoïde de pénétrer 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 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 chromatinnienne plus importante au cours de la spermatogénèse afin d'augmenter sa compaction. Ainsi, les octamères d'histones présents dans les cellules somatiques sont remplacées par deux protéines riches en arginine et en cystéine PRM1 et PRMM2. Ces protéines sont appelées des protamines (**Figure : 1.9**). L'intégrité des deux protéines composant ce dimère est nécessaire pour la procréation (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 chromatinnienne 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)

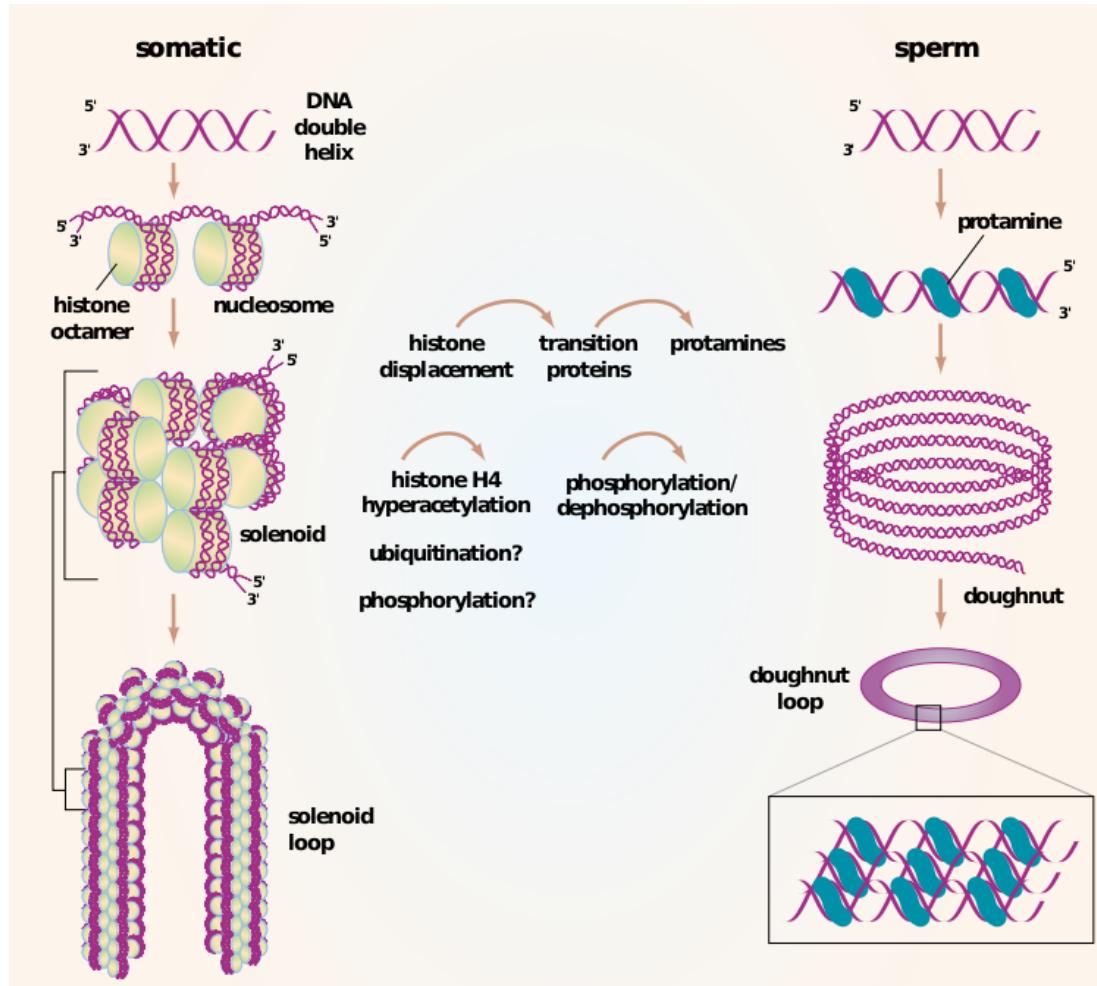
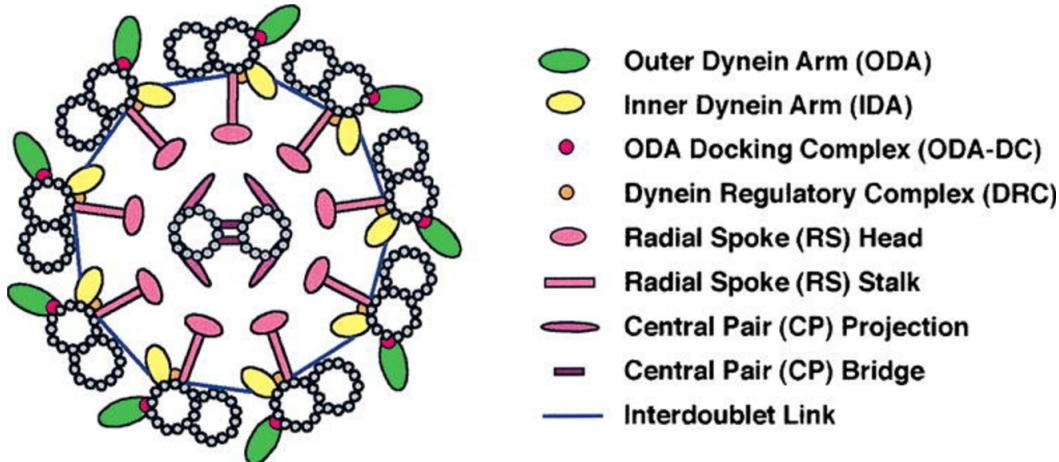


Figure 1.9 – Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes : D'après Braun (2001)

## Le flagelle

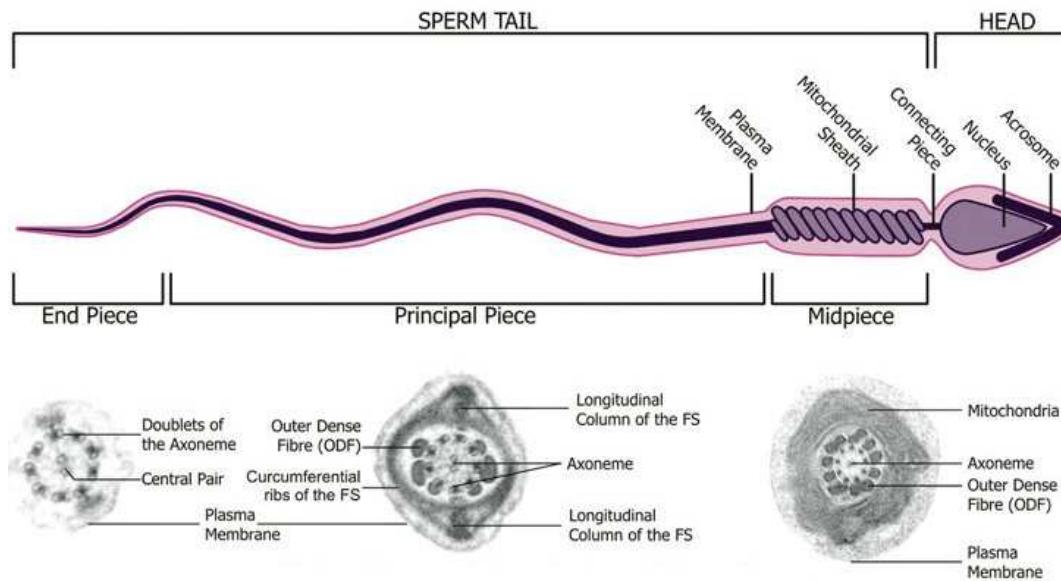
Le flagelle représente la queue du spermatozoïde. Celui-ci permet, par mouvement d'oscillation à haute vitesse, le déplacement du spermatozoïde. Cette mobilité est générée par un cytosquelette interne extrêmement conservé durant l'évolution appelée l'axonème. Celui-ci est composé de neuf doublets de 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.



**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 d’un doublet central relié aux doublets périphériques par des ponts radiaux

Le flagelle su spermatozoïde peut être divisé en trois partie 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, appelées colonnes longitudinales sur lesquelles s’insère les fibres denses 3 et 8. C’est le long de la gaine fibreuse qu’est produit 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 médicale 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 10-20% des cas. Comme nous avons pu le voir, la spermatogénèse est une succession de processus complexes qui s'effectue de manière synchrone, de fait la moindre altération génétique affectant une seule de ces étapes est susceptible d'entraîner 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 tel 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 épithéliales de l'urètre, appelées aussi cellules rondes, se retrouvent à la fois dans l'éjaculat des individus ayant une quantité de spermatozoïdes “normal” (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, Dubinskaya, & Gaeva, 1993). Cependant, leur nombre augmente tandis que la quantité de spermatozoïde diminu (SPERLING & KADEN, 1971).

#### Liée à la quantité

Chez l'humain, l'arrêt de la spermatogénèse est défini comme l'incapacité des cellules spermatogénétique à devenir des spermatozoïdes matures. Elle peut survenir à n'im-

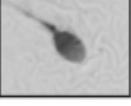
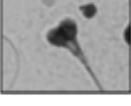
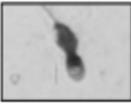
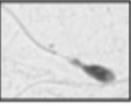
porte qu'elle étape de la formation des cellules germinales. Les arrêts au stade de spermatocyte I sont les plus fréquents, suivis par l'arrêt au niveau des spermatides et moins fréquemment au niveau des spermatogonies (Girgis, Etriby, Ibrahim, & Kahil, 1969).

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 observée dans 4 à 30% des biopsie testiculaire 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 considérés comme sans espoir pour les couples désirant concevoir, jusqu'à l'émergence de *intracytoplasmic sperm injection* (ICIS) (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 total de spermatozoïde 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.

## Liée à la morphologie

Ces anomalies sont observables en effectuant un spermocytogramme. Plusieurs classifications ont été établie, cependant, c'est la classification de David modifiée (**Table : 1.12**) qui est la plus rependue 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 spermatozoïdes.

Table 1. Morphological abnormalities<sup>a</sup> and sperm pathology.

Sperm defect <sup>a</sup> (Light microscopy, final magnification [ $\times 1000$ ])	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 cap 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]

## Liée à la mobilité

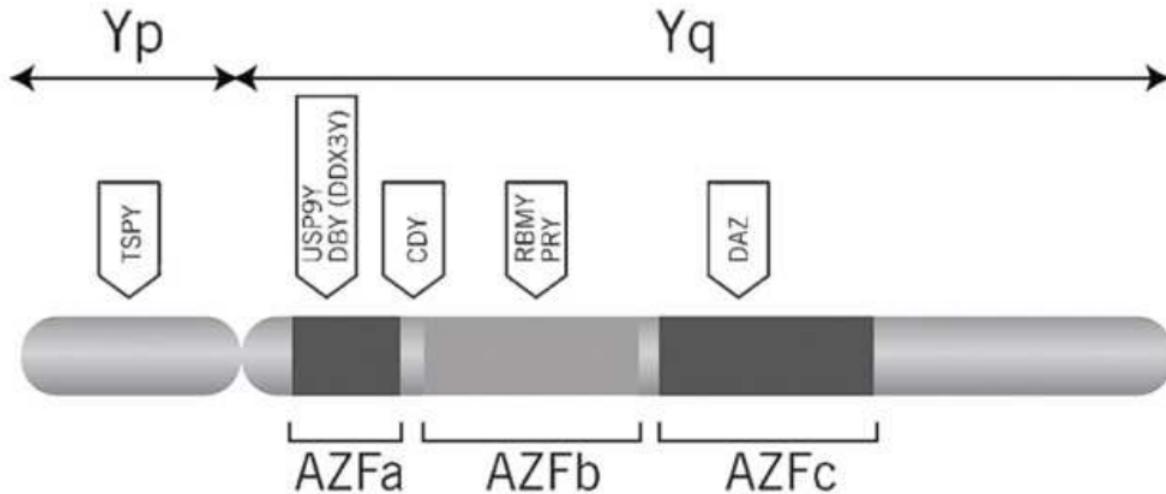
Le succès du passage du spermatozoïdes le long du tractus génital féminin dépend en grande partie de la mobilité et de la vitesse du spermatozoïde (Lindholmer, 1974, Björndahl (2010)). La vitesse moyenne d'un spermatozoïde étant de 25 µm/s. Une mauvaise mobilité observée dans plus de 50% des spermatozoïdes 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 inexpliquées 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 délétion invisibles au caryotype dans la région du facteur AZF (*Azoospermia Factor*). Cette région peut être subdivisée en trois sous-partie, AZFa, AZFb et AZFc (**Figure : 1.13**). Depuis plusieurs années, de nombreuses séries de patients azoospermiques ou oligozoospermiques ont été publiées et tendent à montrer que les microdélétions du chromosome Y seraient responsables de 10% des cas d'azoospermie non-obstructives et chez 5% des cas d'oligozoospermie sévères (<5 millions de spermatozoïdes/ml) (Hotaling & Carrell, 2014).



**Figure 1.13** – Représentation schématique du chromosome Y adapté d'après [O'Flynn O'Brien 2010] : 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 chromosomique de nombre ou de structure impliquant les autosomes ou, le plus souvent, les gonoosomes, peuvent être impliqués dans des cas d'infertilité masculine. Le pourcentage d'individu concerné 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 translocation 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 translocation réciproques et les translocation robertsonniennes. 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 de structures chromosomique responsables d'infertilité masculine, on peut par exemple citer les hommes de formule 46,XX bien qu'elle soit moins fréquente que les translocations et les inversions. 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.
- 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 tout phénotype confondus. Cette observation suggère fortement l'implication d'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 spermatozoïdes :

- 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 responsable 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 dysfonctionnement de celle-ci menant à des spermatozoïdes polyploïdes, c'est à dire, portant une quantité de matériel génétique trop importante [INSERT REF].

- 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 d'une majorité de spermatozoïde dépourvu d'acrosome dans l'éjaculat empêchant ainsi le spermatozoïde de franchir la zone pellucide de l'ovocyte comprenant 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 reporté 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 tête / flagelle du spermatozoïde (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 et les flagelles spermatiques présentent de multiples anomalies morphologiques. Ce phénotype a été décrit plusieurs fois dans la littérature et revêt plusieurs formes (C. Coutton, Escoffier, Martinez, Arnoult, & Ray, 2015). 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 et al., 2015, Ben Khelifa et al. (2014)). Recemment, 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 retrouvée des mutation 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 échec 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  $\text{Ca}^{2+}$ . Ce processus est médié par une protéine spécifique du spermatozoïde : *PLC $\zeta$  1* (Nomikos, Kashir, Swann, & Lai, 2013, Amdani, Jones, & Coward (2013)). Plusieurs cas d'échec d'activation ovocitaire ont été liés à l'absence ou la mauvaise localisation de la protéine *PLC $\zeta$  1* (*phospholipase C Zeta 1*). Malgré cela, aucune preuve génétique directe n'avait été reporté jusque récemment où deux mutations au sein du gène *PLC $\zeta$  1* furent retrouvés 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 génétique discrètes” que sont les gènes fut suggéré dès la deuxième moitié du XIX<sup>e</sup> siècle grâce aux travaux de Gregor Mendel portant sur l'hérédité de certains traits chez le poïds. Depuis, de nombreuses méthodes permettant de lier le phénotype d'un individu à son génotype on 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 à la grande partie de 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, quelques cette méthode compte à son actif plusieurs succès retentissants [INSERT PETITE LISTE].

### 1.4.2 Les puces

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

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

### Du tissu au transcriptome, le différentiel d'expression

#### 1.4.3 Le séquençage NGS

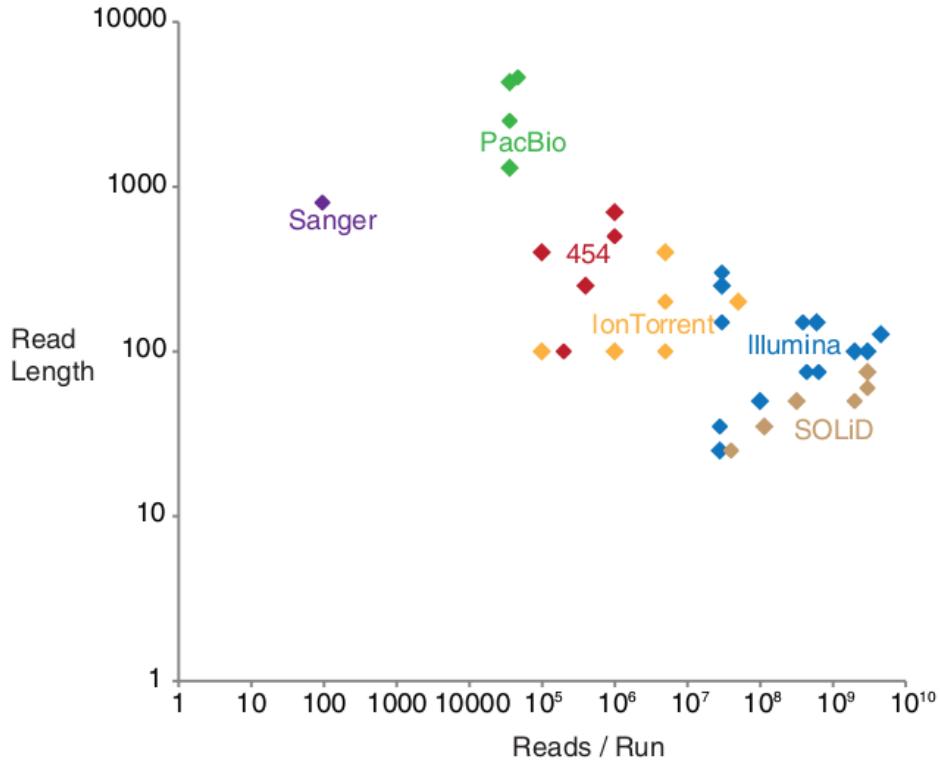
Le terme séquençage de l'ADN fait référence à l'ensemble des techniques permettant de déterminer l'ordre des nucléotides A, T, C et G de l'intégralité ou d'une partie d'une molécule d'ADN. Avant de parler des nouvelles technologies de séquençage (NGS) faisons un bref historique du séquençage de l'ADN. 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'autre méthode 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 de commerciale et de recherche (Wikipédia). 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 cout 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 ... 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 majeurs 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 du 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 *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] : Sequencing space based on read length (in bases) and number of reads per run. Points represent official platform/chemistry combination releases and are color-coded based on the platform family. To see this illustration in color, the reader is referred to the web version of this article at [www.liebertpub.com/wound](http://www.liebertpub.com/wound)

## La capture des parties à séquencer, avantage et inconvenants

Pour de nombreuse application, 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 associé, l'ensembles 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 cout 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. Pour ces raisons, le WES considéré comme le standard dans le cadre de recherche sur des pathologies génétiques et se 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 dépendant 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 couts de séquençage et de stockage des données, il est possible que le WGS remplace totalement le WES ainsi que l'ensemble des techniques impliquant la capture de séquences ciblées (Meienberg et al., 2016).

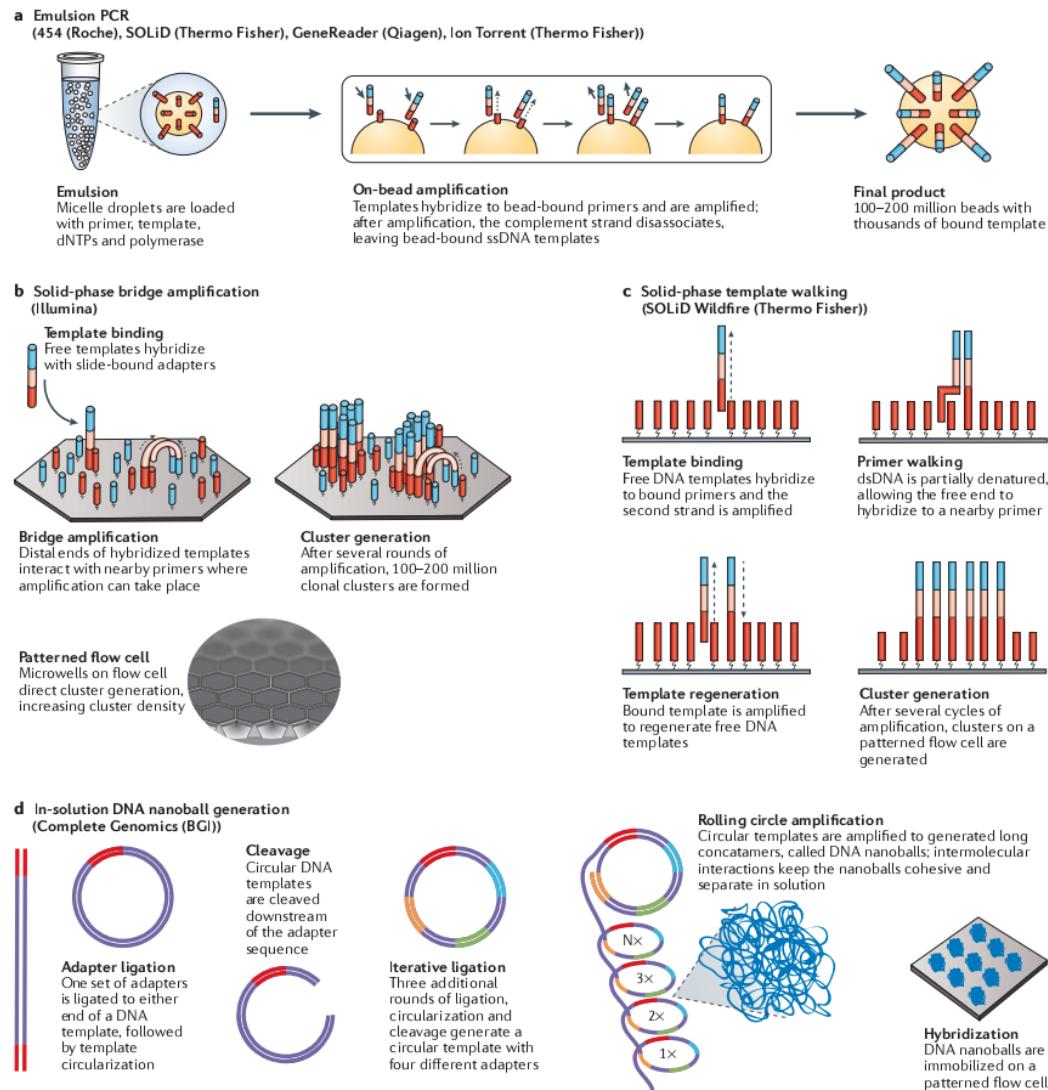
## L'amplification

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 excepté pour la PCR en émulsion qui s'effectue en phase aqueuse. Elle permet d'obtenir dans une région définie plusieurs milliers de copie 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 et sera ensuite séquencé parallèlement aux autres *spots*. Une plateforme de séquençage pouvant gérer plusieurs millions de ces centres de réactions simultanément, séquençant ainsi plusieurs millions de molécules d'ADN en parallèle, donnant ainsi le nom à ces techniques qualifiées de séquençage massif en parallèle. 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 amplifié. 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 composant nécessaire à la réaction de PCR. En respectant un ratio

nombre de molécule 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 liée 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 initiale un massif de molécules fixées 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 fixée 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\*\* :

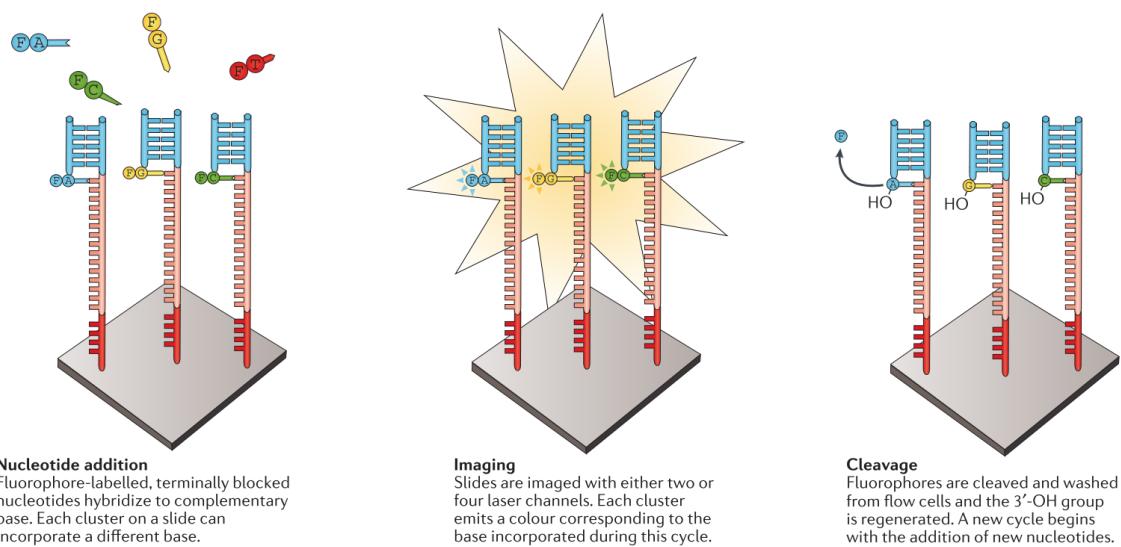
## La réaction de séquence

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) :

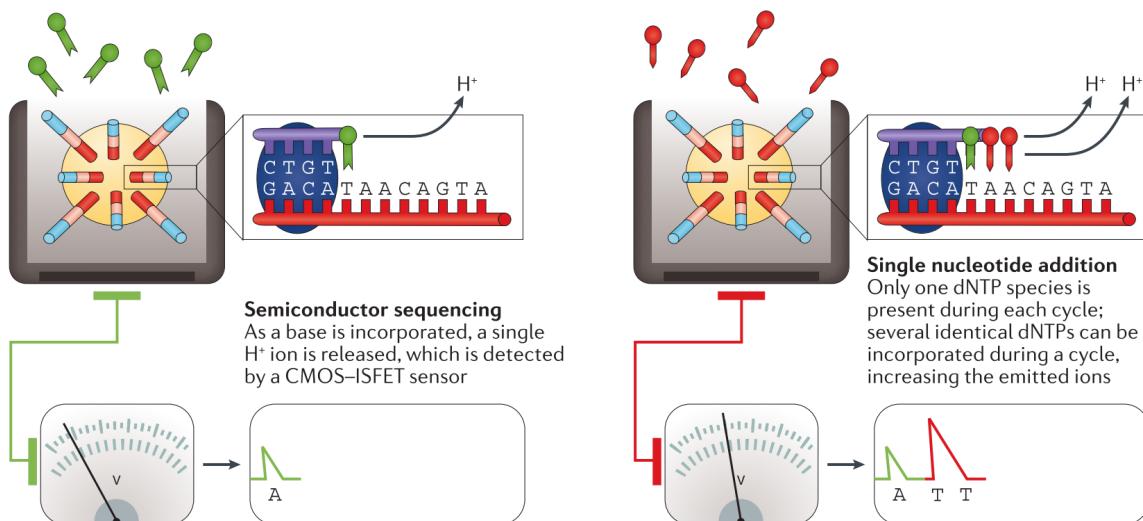
- Terminaison par cycle réversible, cyclic reversible termination (CRT)** (**Figure : 1.16**) : Cette méthode est caractérisée par son utilisation de molécules d'acides terminatrices auxquelles le groupement 3' – OH est modifié de sorte à éviter l'élongation (J. Guo et al., 2008), on parlera de groupement 3' – bloqué. Le processus est initialisé une amorce est liée au fragment d'ADN et permettra l'initialisation de la polymerisation. À chaque cycle, un mélange comprenant l'ensemble des quatre d'oxynucléotides (dNTPs), préalablement étiquetés par un fluorophore et 3' – bloqué sont mis en contact avec le fragment. Après l'incorporation d'un seul dNTP au fragment, les dNTP 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'élongation

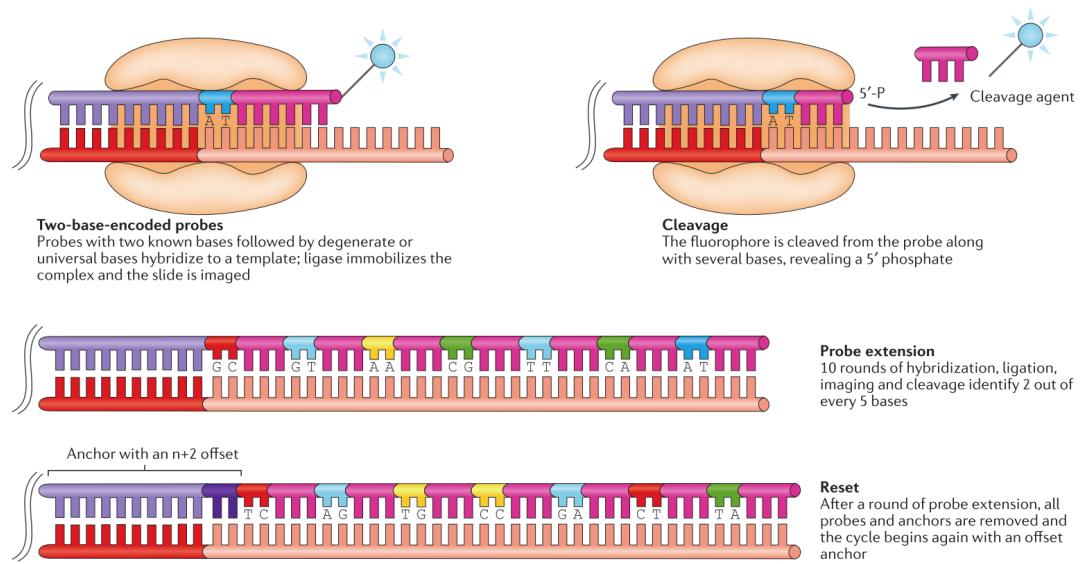
- Addition de nucléotide unique, single nucleotide addition (SNA)** (**Figure : 1.17**) : L'initialisation de la méthode SNA est identique à celle de la méthode CRT. La différence se fait donc au moment de la phase d'élongation. Contrairement à la méthode CRT, le mélange contenant les dNTPs ne contient qu'un seul type de dNTP. Quatre mélanges 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 mixe, vérifie si un dNTP s'est lié au fragment. Lors des séquences homopolymériques (plusieurs nucléotides identiques successifs dans la séquence), plusieurs dNTP sont donc lié simultanément, cela sera détecté car le signal émis sera 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 d'homopolymère, plusieurs nucléotides identiques successifs, autant de proton sont relâché que de constituant 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 (Tomkinson, Vijayakumar, Pascal, & Ellenberger, 2006) d'une sonde liée à un fluorophore. Ce processus utilise les caractéristiques de la ligase, une enzyme qui a pour fonction de catalyser la liaison de deux brins d'ADN par des liaison 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* court 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 couts raisonnable. Ainsi, l'émergence du séquençage haut débit et notamment du WGS et du WES a permis de réunir une quantité jusqu'à présent inégalé d'information sur les variations génétiques, et d'une manière plus générale, 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**). Il existe deux types de *reads* :

1. ***Read single-end*** :
2. ***Read paired-end*** : Avec le séquençage de *paired-end reads* les deux extrémités (les *ends*) du fragment d'ADN est désormais séquencée. 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 le format de donnée le plus couramment retourné par les séquenceur haut-débit à l'heure actuelle. Sa création est cependant antérieure à l'émergence du NGS puisqu'il fut inventé à la fin du XX<sup>ième</sup> par Jim Mullikin au Wellcome Trust Sanger Institute alors que le séquençage commençait à prendre de l'ampleur grâce à des projets tel que le Projet Génome Humain. La quantité de données générées par ces programmes à nécessité une analyse automatisé, c'est ainsi que chaque base séquencée s'est vue associé 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 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
ACACATACGCACTGGCGTAAAGGGCGCGAGCGCGTCAGAGCGTCGGTCAAAGTCCACCGCTAACGGTGGAGGCCTG
+HC9D00P01AN1VB
FFFFFFFFFFFGD554A6911144442AAABDFFIIIIIIIIIIIIIIHHHFFFFFFFA@CFDFDFDFC???CFFFFFFFFFF
@HC9D00P01AWYAE rank=0000402 x=258.0 y=772.0 length=373
ACACATACGCACTGGGCATAAAAGGGCACGTAGGCCGATTGTAAGTCAGGGGTGAATCCGGGCGTCAACCTCGGAACTGCCT
+HC9D00P01AWYAE
IIIIIIIIIIIIIIHHHII;666MHIIIIIIIIIIICCIIEEEFDC2//.<-//93.....---9CCCCFECCCCIIIIDI
@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
IIIIHHHIIIIHHHII;;IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@HC9D00P01AU18Y rank=0000952 x=230.0 y=2656.0 length=372
ACACATACGCACTGGCATAAAAGAGCGCGTAGGCCGCTTGTAGTCAGGTGAAGGCCCTGGCTTAACCCGGGAAGCGCGC
+HC9D00P01AU18Y
IIIIIIIIIIIIHHHII;;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

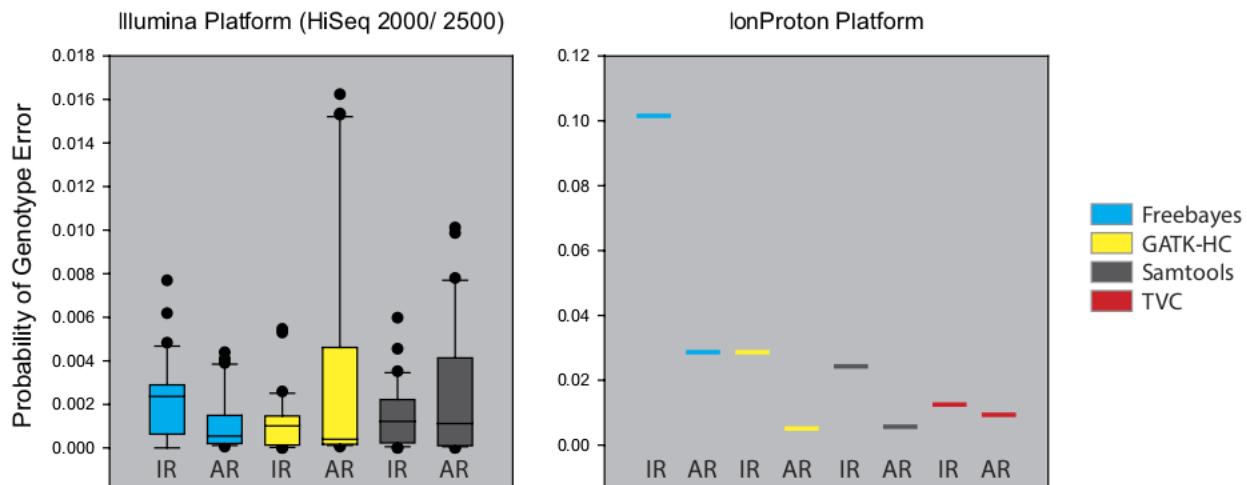
censé représenter la séquence consensus d'une espèce donnée. Par exemple, la séquence de référence humaine GRCh37 (*Genome Reference Consortium human build 37*) a été créés à partir de 13 volontaires anonymes New-Yorkais. Dès lors, cette référence servira de patron aux aligneurs afin qu'ils replacent 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 reposent l'ensemble des étapes effectuées en aval, notamment l'appel des variants (R. Nielsen, Paul, Albrechtsen, & Song, 2011). Cependant, l'étape d'alignement peut être sujette à de nombreuses erreurs dont certaines proviennent directement des erreurs de survenues lors de l'étape de séquençage, d'autres, sont dues aux caractéristiques des régions séquencées comme par exemple les séquence 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 émergé 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 nombreuse études ont cependant montré de grandes différences entre ces aligneurs, au niveau du temps de calcul, de leur cout 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 dépendant 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 variants toutes différences de séquence observées entre un individu et la séquence de référence utilisée. Pour reprendre la comparaison avec la construction d'un puzzle, cette étape consiste à détecter quels sont les pièces qui présentent des différences avec le modèle. De nombreux logiciels d'appel 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 préfère les appels variant-homozygote tandis que GATK-HC et Samtools préfèrent les appels hétérozygotes. Pour la plateforme Ion Proton, les 4 logiciels ont une préférence pour les 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 caller montrait que seulement 57,4% des variants étaient appelés par les 5 caller 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 *caller* (O'Rawe et al., 2013). Ces résultats sont cependant à pondérer avec une étude de 2015 comparant 4 *caller* et montrant que 91,7% des SNVs séquencés sur une plateforme Illumina étaient appelés par 3 *caller*, cependant, pour les variants séquencés sur Ion Proton, seulement 27,3% des variants étaient appelés par au moins 3 *caller* et 57,4% des variants n'étaient appelés que par un seul des *caller* (Hwang et al., 2015).

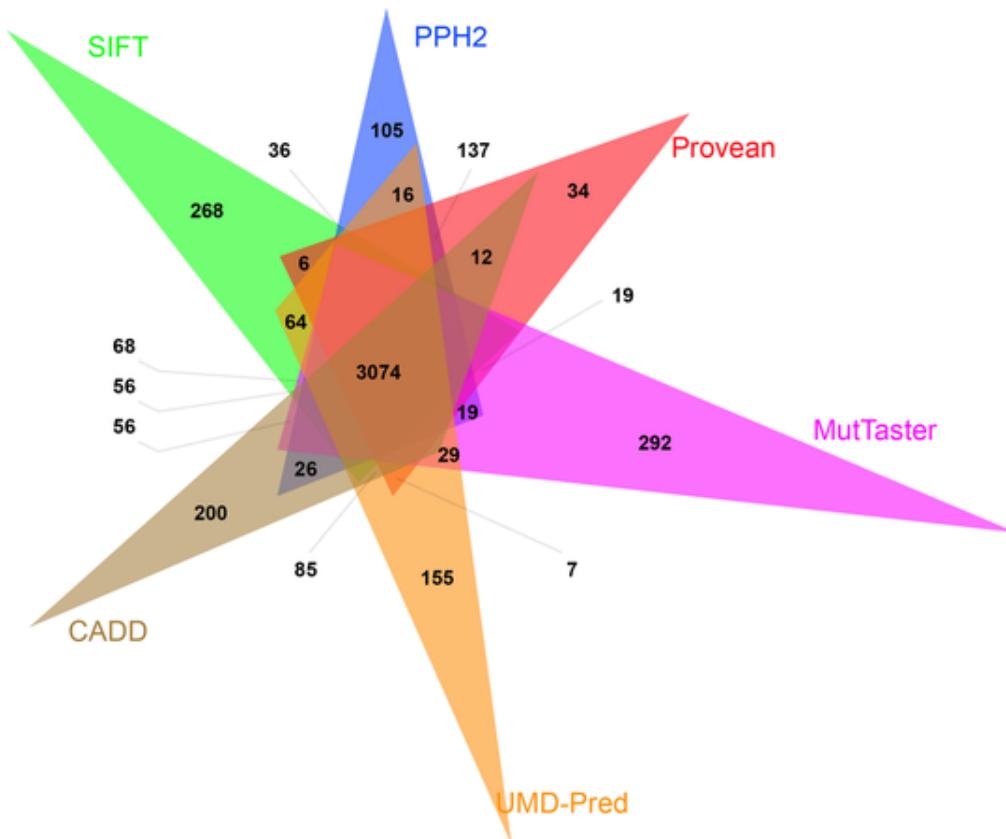
#### 1.5.4 L'annotation des variants, filtrage et priorisation

Traditionnellement, les scientifiques développaient leur expertise dans un nombre de pathologie et de gènes associés limité. L'émergence du NGS a totalement remis en cause cette pratique, dès lors qu'il est désormais 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 désormais indispensable d'annoter cet ensemble de variant, c'est à dire d'associer à ces variants l'ensembles 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 variant à considérer .... On peut généralement distinguer deux niveaux d'annotation 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. 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 on put 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 effectuer aucun séquençage mais qui à récupérer 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.

- c. **Son impact sur le transcrit** : Dans la plupart des analyses phénotype-génotype, les chercheurs se limitent au variant chevauchant des transcrits codant 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... 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'autre logiciel tel que SIFT (P. Kumar, Henikoff, & Ng, 2009), PROVEAN (Y. Choi, Sims, Murphy, Miller, & Chan, 2012), Polyphen2, ou encore CADD vont chercher à prédire la pathogénicité de ce variant, c'est à dire la probabilité que ce variant soit délétère pour l'individu qui le porte. 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 logiciel (**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,

transcrit). L'annotation au niveau de l'unité génétique consiste à récupérer l'ensemble des informations disponible 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 utile notamment lorsque peu d'information sont disponibles sur le variant lui-même. En pratique, la plupart des variants connues pour impliquer une pathologie sont des variants privés, c'est à dire spécifique à une famille ou un individu limitant ainsi la quantité d'information disponible sur ce variant. Élargir l'annotation au niveau des unités génétiques impactés 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 AAnalysis THrough Evolutionary Relationships* (PANTHER) (Mi et al., 2017) 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 tolérant aux variations que les autres.

Comme nous l'avons vu, l'accumulation de cette information est extrêmement importante puisqu'elle permet aux biologistes de faire face à la masse de données générées par le NGS l'aidant ainsi dans ses prises de décisions. Il est à noter que la plupart de ces informations sont extrêmes dépendantes du jeu de gènes utilisés, les prédictions seront donc différentes si l'on se base 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, & Brouard (2016)).

### 1.5.5 Conclusion NGS

En moins de 10 ans, les technologies NGS sont passées du séquençage de panel de gènes (environ 100 Mb pour le Roche GS FLX system) au séquençage de génome entiers (environ 1500 GB pour l'Illumina Hiseq 4000) et d'une utilisation exclusive à la recherche à la routine clinique. Le nombre croissant de 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 couts par génome / exomes séquencés laisse supposer que ces technologies

deviendront d'ici peut le fer de lance de la génétique clinique moderne. Cependant, cette quantité de donnée 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 pour palier la contrainte instaurée par les *reads* courts dans le cadre d'analyse génomique, les solutions informatique et bioinformatique proposée 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 retrouve désormais être un frein dans la compréhension et l'interprétation des réseaux de gènes et leurs implications dans des pathologies, la limitation de cette technologie n'étant 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 autres algorithmes d'analyses, bien que nécessaire, s'ajoute à la confusion.

Malgré les dizaines de milliers d'exomes et de génomes ayant été jusqu'à présent étudiés, notre compréhension des mécanismes moléculaire qui sous-tendent la variété génomique humaine reste limité, 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 caractérisée mais un nombre similaire ont toujours une cause inconnue (Amberger, Bocchini, & Hamosh, 2011). L'éluication de ces mystères passera probablement par une harmonisation de méthodes de production des données ainsi que par l'amélioration des techniques d'analyses.

## Chapitre 2

# Investigation génétique et physiologique de la globozoospermie



# Chapitre 3

## MutaScript

### 3.1 Introduction

Il y a quelques années, le séquençage Sanger était encore massivement utilisé en recherche clinique. Cette technique était extrêmement coûteuse 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 part sa masse, les données produites créées 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. Parmis ceux-ci, 20.000-25.000 d'entre eux impactent des régions codant pour une protéine avec environ 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 ces tâches 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'ensemble 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 associés. 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 Mendelienne et, *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 des 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'impacte de chaque variant sur l'ensemble des transcrits qu'ils chevauchent de sorte à ce que les variants ayant un impacte prédict commé étant délétère aient une plus grosse contribution au score MutaScript que ceux ayant un impacte 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 aux 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'emergence 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 aggrge les séquence codante 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 décallage du cadre de lecture. Cependant, ces variants ne représentent que ... % des variants fournit par la base de données ExAC. C'est pourquoi, MutaScript prend en compte l'ensemble

des variants, peut importe leur impacte sur les différents transcrits qu'ils chevauchent, et leur attribue un poids en fonction de cet impacte 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 variations en se basant sur des listes de gènes fournies 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...

## 3.2 Matériel & Méthodes

### 3.2.1 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 baser 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 transcrits taggés comme codants 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 variantes n'ayant pas une couverture médiane  $\geq 15$  furent filtrés
  - d. L'ensemble des variants introniques (sauf ceux proches des sites d'épissage) et les variants situés dans les régions *upstream* et *downstream* furent filtrés

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

### 3.2.2 Validation du score

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

## 3.3 Résultats

### 3.3.1 Résultat de l'annotation

1. tableau avec l'ensemble des csq et l'impacte vep associée
2. fréquence des ces impactes
- 3 bar plot des poids

### 3.3.2 Détermination de la formule du score

**Le SLAC et le WSLAC** Pour chaque transcrit nous avons défini deux métriques. La première est le SLAC (3.2) qui se définit comme étant pour transcrit, la somme du log des comptages allélique de chaque variant chevauchant ce transcrit.

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

$$SLAC_T = \sum_{v = \text{variant chevauchant } T} \log(\text{allele count}_v) \quad (3.1)$$

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

$$WSLAC_T = \sum_{I=Impact} \sum_{v=variant} Poid_I \cdot \log(\text{allele count}_v) \quad (3.2)$$

1. Le SLAC et le WSLAC
  - a. formule du SLAC et du WSLAC
  - b. graphique SLAC x WSLAC avec regression 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

### 3.3.3 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 transcrit plutôt que par gène)

## 3.4 Comparaison avec RVIS et pLI

- 1. corrélation score~size
- 2. hpo
- 3. gene dispensable
  - a. liste des 240 gènes
  - b. récepteurs olfactifs

## 3.5 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., ... Flieck, 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 $\zeta$ ) : 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/dew270>

//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>

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>

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>
- 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., Koscienski, 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 polyploid spermatozoa and causes male infertility. *Nature Genetics*, 39(5), 661–5. <http://doi.org/10.1038/ng.2027>
- 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>
- 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>.

- 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. *Seminars 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 (PLC<sub>Zeta</sub>) 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., Pavesi, 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>
- 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>
- 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 *< i>in Vivo</i>* and *< i>in Vitro</i>* : 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]. *Urologiia I Nefrologiia*, (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>
- 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>
- 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](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 Massively 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 $\zeta$  : From structure to Ca<sup>2+</sup> 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- 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>
- Stitziel, 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>
- Thankaswamy-Kosalai, S., Sen, P., & Nookaew, 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>
- 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>
- 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>
- 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](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>
- 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>