

THÈSE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

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

Arrêté ministériel : 25 mai 2016

Présentée par **Thomas Karaouzene**

Dirigée par **Pierre Ray**

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

Préparée au sein des laboratoires **Génétique, Epigénétique et Thérapies de l'Infertilité (GETI)** et **Techniques de l'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications de Grenoble (TIMC-IMAG)**
Et de l'école doctorale "**Ingénierie de la Santé, de la Cognition et Environnement**" (**EDISCE**)

Bioinformatique et infertilité : analyse des données de séquençage haut-débit et caractérisation moléculaire du gène DPY19L2

Thèse soutenue publiquement le **29 novembre 2017**, devant le jury composé de :

Pr Jacques VAN HELDEN

Professeur des Universités, Université d'Aix-Marseille, Rapporteur

Dr Michaël MITCHELL

Directeur de Recherches INSERM, Université d'Aix-Marseille, Rapporteur

Pr Christel THAUVIN

Professeur des Universités–Praticien Hôpitalier, Université de Bourgogne, Examinateur

Dr Julien THÉVENON

Assistant Hospitalier Universitaire, Université Grenoble-Alpes, Examinateur

Pr Pierre RAY

Professeur des Universités–Praticien Hôpitalier, Université Grenoble Alpes, Directeur de thèse

Dr Nicolas THIERRY-MIEG

Chargé de Recherches CNRS, Université Grenoble Alpes, Co-directeur de thèse



Table des matières

Chapitre 1 : If you are creating a PDF you'll need to write your preliminary content here or	1
Remerciements	4
Résumé	6
Chapitre 2 : Introduction	7
2.1 La spermatogenèse	8
2.1.1 Rappels sur le testicule	9
2.1.2 La phase de multiplication	10
2.1.3 La méiose	12
2.1.4 La spermioïgenèse	16
2.2 Structure et fonction du spermatozoïde	18
2.2.1 La tête	19
2.2.2 Le flagelle	21
2.3 L'infertilité masculine	23
2.3.1 Les différents phénotypes d'infertilité masculine	23
Anomalies liées à la quantité spermatique	24
Anomalies liées à la morphologie	24
Anomalies liées à la mobilité	26
2.3.2 La génétique de l'infertilité	26
Les causes fréquentes	26
Les nouveaux gènes	29
2.4 Généralités sur l'ovogenèse et l'ovocyte	32
2.5 Les techniques d'analyses génétiques	34
2.5.1 Approche "gènes candidats"	34
2.5.2 Les puces	36
Les puces à expression	37
Les puces à SNP, plateforme génotypage	38
Les puces à indels	39
Limitation	39
2.5.3 Le séquençage NGS	40
La capture des parties à séquencer, avantages et inconvénients	41
L'amplification	42
La réaction de séquence	45
2.6 L'analyse bioinformatique des données de NGS	48
2.6.1 Les données fournies par le NGS	48
Un <i>read</i> , c'est quoi ?	48
Le format FASTQ	49
2.6.2 L'alignement	50
2.6.3 L'appel des variants	52
2.6.4 L'annotation des variants	55
2.6.5 Le filtrage des variants	58
2.6.6 Conclusion NGS	59

2.7 Problématique : Un patient, 50.000 variants. “ <i>There can be only one</i> ”.	
Et après ?	61

Chapitre 3 : Mise en place d'une stratégie pour l'analyse des données exomiques – application en recherche clinique	63
3.1 Méthode : Description d'ExSQLibur	64
3.1.1 L'alignement des <i>reads</i>	64
3.1.2 L'appel des variants	64
3.1.3 L'annotation	64
3.1.4 Le filtrage des variants	64
3.2 Résultats 1 : Analyse de 3 phénotypes par des cas familiaux	64
3.2.1 Résultats des différentes étapes de l'analyse	64
Résultat de l'alignement	64
L'appel des variants	64
L'annotation des variants	64
Le filtrage des variants	64
3.2.2 Article n°1	64
Contexte et objectifs	64
Principaux résultats	64
3.2.3 Article n°2	64
Contexte et objectifs	64
Principaux résultats	65
3.2.4 Article n°3	65
Contexte et objectifs	65
Principaux résultats	65
3.3 Résultats 2 : Étude d'une cohorte de femmes infertiles	65
3.3.1 Article n°4	65
Contexte et objectifs	65
Principaux résultats	65
3.4 Résultats 3 : Étude d'une large cohorte de patients MMAF	65
3.4.1 Article n°5	65
Contexte et objectifs	65
Principaux résultats	65
Chapitre 4 : Investigation génétique et physiologique de la globozoospermie	67
4.1 Introduction sur la globozoospermie	68
4.2 Résultats 1 : Les mécanismes mutationnels entraînant la délétion au locus de <i>DPY19L2</i> chez l'humain	68
4.2.1 Article n°6 :	68
Contexte et objectifs	68
Principaux résultats	68
4.3 Résultat 2 : La transcriptomique	68
4.3.1 Article n°7 :	68
Contexte et objectifs	68

Principaux résultats	68
Chapitre 5 : This chunk ensures that the thesisdown package is . . .	69
Chapitre 6 : Article annexe 1	71
References	74

Liste des tableaux

2.1 Durée de vie moyenne des cellules germinales humaines	8
---	---

Table des figures

2.1	Schéma anatomique du testicule humain	9
2.2	Les différentes phases de la spermatogenèse	11
2.3	Les différentes étapes de la méiose gamétique masculine	12
2.4	Schéma simplifié d'un enjambement chromosomique (crossing-over) .	14
2.5	Les différentes étapes de la première division méiotique masculine .	15
2.6	Les différentes étapes de la deuxième division méiotique masculine .	15
2.7	Principales étapes et modifications structurales lors de la spermogénèse	17
2.8	Anatomie simplifiée du spermatozoïde	18
2.9	Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes	20
2.10	Structure simplifiée de l'axonème	21
2.11	Structure du flagelle d'un spermatozoïde	22
2.12	Classification morphologique de spermatozoïdes humains normaux et anormaux	25
2.13	Représentation schématique du chromosome Y	27
2.14	Les différents types de translocation	28
2.15	La fécondation, liaison spermatozoïde-ovocyte et sortie de la méiose .	33
2.16	Représentation schématique des méthodes d'analyse d'expression génique par puce à ADN	37
2.17	Méthode de génotypage par discrimination allélique par hybridation .	38
2.18	Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée	41
2.19	Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS	44
2.20	Séquençage CRT tel qu'il est effectué par Illumina	45
2.21	Séquençage SNA tel qu'il est effectué par Ion Torrent	46
2.22	Séquençage SBL tel qu'il est effectué par SOLiD	47
2.23	Présentation d'un fichier FASTQ	49
2.24	Représentation schématique de l'alignement de reads paired-end . . .	51
2.25	Illustration schématique du processus d'appel des variants	52
2.26	Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel	54
2.27	Diagramme de Venn des prédictions de pathogénicité de variants de six logiciels	56
2.28	Représentation simplifiée du processus d'annotation	57
2.29	Représentation simplifiée du processus de filtrage des variants	58
2.30	Récapitulatif des différentes étapes du séquençage NGS dans le cadre d'une étude phénotype-génotype	60

CHAPITRE 1

If you are creating a PDF you'll need to write your preliminary content here or

Remerciements

Résumé

CHAPITRE 2

Introduction

2.1 La spermatogenèse

La spermatogenèse des mammifères est un processus long et complexe contrôlé par plusieurs mécanismes étroitement liés [1–3]. C'est au cours de celle-ci, qu'à partir de cellules germinales, seront produits les spermatozoïdes matures. Ce processus est divisé en trois phases principales : la phase de multiplication, la phase de division (appelée la méiose) et la phase de maturation. Chez les hommes, ces étapes se déroulent en continu dans la paroi des tubes séminifères du testicule depuis la puberté jusqu'à la mort et impliquent 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 spermatozoïdes s'élève environ à 45 millions par testicule [4]. 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 : 2.1**).

Table 2.1 – 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

2.1.1 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^3 . 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 : 2.1**). 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 : 2.1**). C'est le long des parois du tube séminifère que se déroulera l'ensemble des étapes de la spermatogenèse.

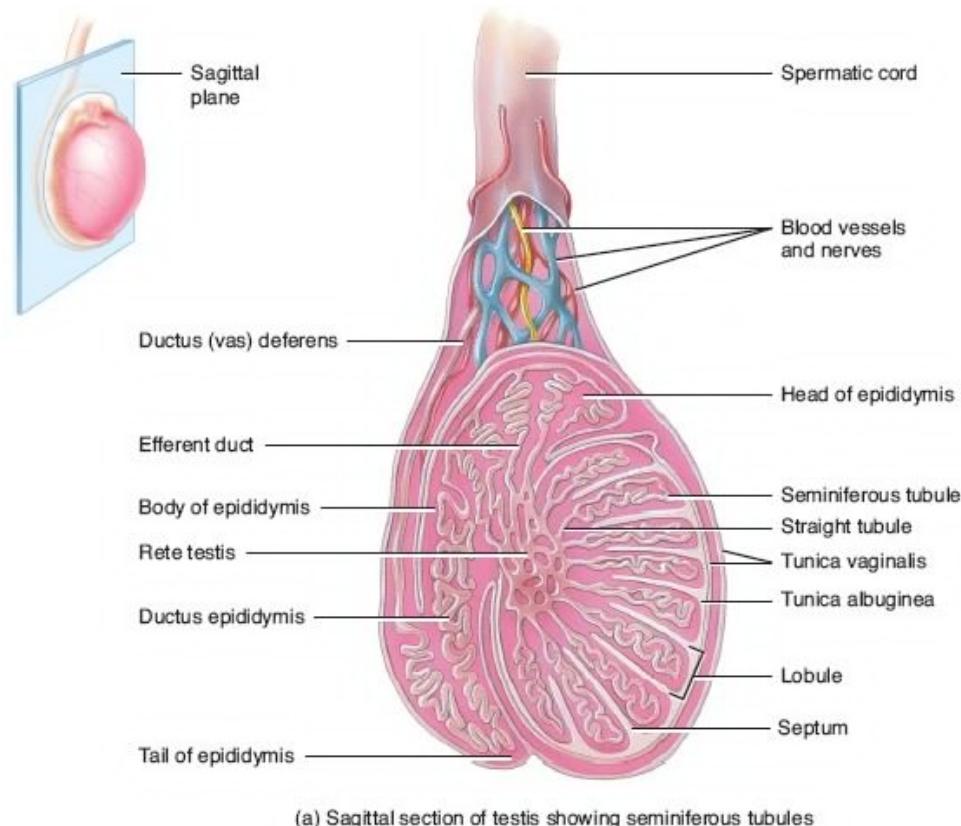


Figure 2.1 – Schéma anatomique du testicule humain.

2.1.2 La phase de multiplication

La phase de multiplication est la phase au cours de laquelle les spermatogonies se divisent par mitoses pour aboutir au stade de spermatocytes primaires. Les spermatogonies sont des cellules diploïdes à l'origine de l'ensemble des autres cellules germinales humaines. Pour cela, elles vont s'auto-renouveler par mitoses successives afin de maintenir une production continue de spermatozoïdes tout au long de la vie de l'individu. Ces cellules sont localisées dans le compartiment basal des tubes séminifères. Les analyses histologiques ont permis de distinguer trois types de spermatogonies en fonction de leur contenu en hétérochromatine [5–7] : Les spermatogonies de type A dark (ou Ad), les spermatogonies de type A pale (ou Ap) et les spermatogonies de type B.

Chez l'Homme, les spermatogonies Ad ont une activité mitotique au cours de la spermatogenèse et servent de réserve. Elles vont au cours d'une première mitose former une spermatogonie Ad et un spermatogonie Ap (**Figure : 2.2**). Cette propriété permet à la fois de se différencier en spermatocytes tout en constituant un compartiment de réserve de spermatogonies Ad pour la régénération de la population de cellules germinales au sein de l'épithélium séminifère. L'entrée en division des spermatogonies Ap se fait par groupes cellulaires tous les 16 jours. Les cellules d'une même génération maintiennent entre elles des ponts cytoplasmiques jusqu'à la spermiogenèse 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 former, lorsqu'elle se divise par mitose, deux spermatogonies B qui elles-mêmes se diviseront en deux spermatocytes primaires diploïdes (**Figure : 2.2**).

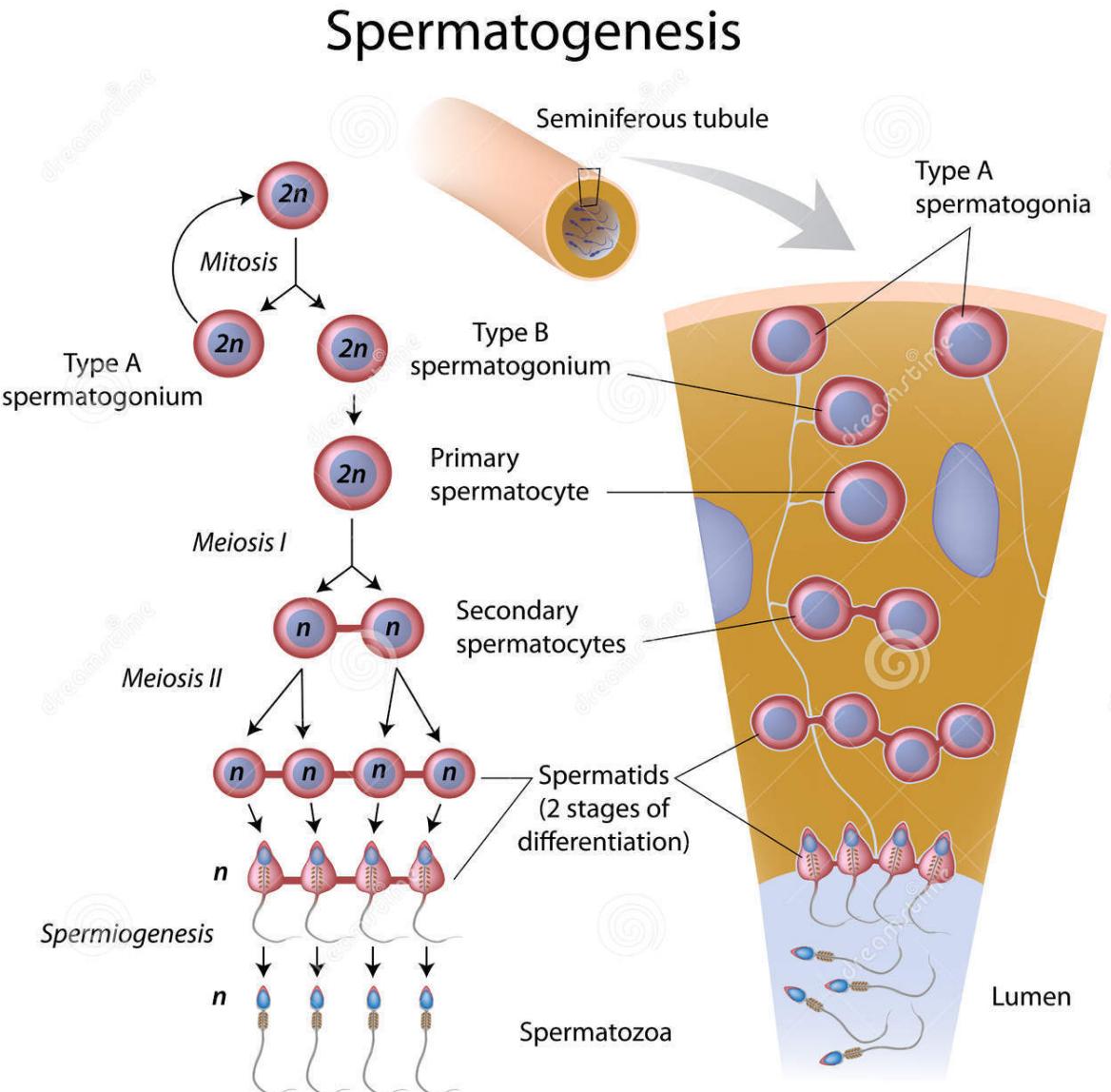


Figure 2.2 – Les différentes phases de la spermatogenèse
d'après medizin-kompakt : L'évolution de la spermatogenèse est strictement coordonnée, tant dans le sens transversal des tubules, que dans le sens longitudinal. Le sens transversal correspond au sens unique de la différenciation germinale, les cellules souches sont situées à la base du tube séminifère alors que les spermatozoïdes aboutissent à la lumière. Dans le sens longitudinal de ces tubules, on retrouve l'ensemble des cellules d'un même cycle donnant ainsi l'onde spermatique.

2.1.3 La méiose

La méiose, ou phase de maturation, est l'étape au cours de laquelle, à partir de cellules diploïdes (les spermatogonies B) vont se former des cellules haploïdes, les spermatocytes secondaires (spermatocytes II). Ce résultat est le fruit de deux divisions successives (**Figure : 2.3**) appelées 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 nom *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 : 2.4**).

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'elles vont tout d'abord subir une interphase (stade préleptotène) durant entre 2 et 4 jours. Au cours de cette phase a lieu la réplication de l'ADN. Cette réplication se fait lorsque l'ADN est à l'état de chromatine, pendant la phase S (pour synthèse) de l'interphase. À l'issue de cette phase, chaque chromosome sera composé de deux chromatides reliées entre elles par le centromère, le matériel génétique de chaque cellule ayant donc été multiplié par deux. Par la suite, ces cellules vont subir deux divisions méiotiques, chacune composée de quatre étapes distinctes (**Figure : 2.3**) :

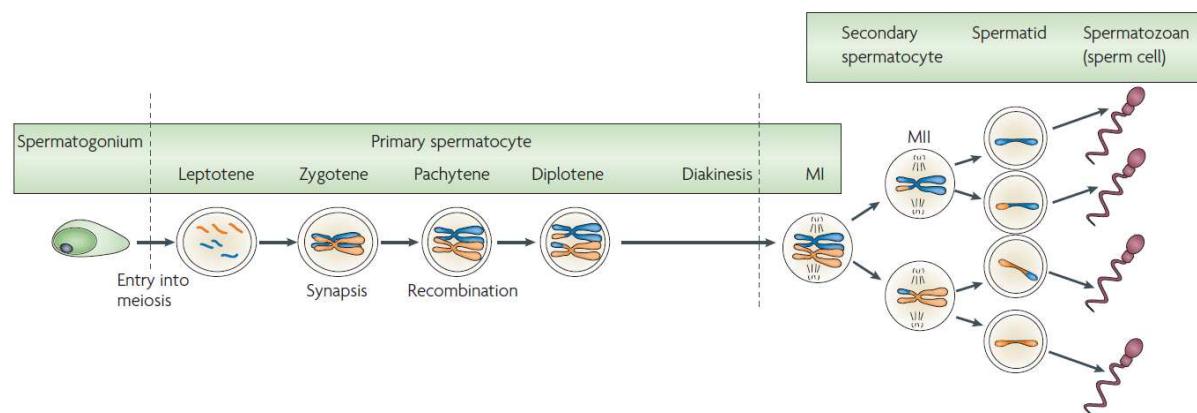


Figure 2.3 – Les différentes étapes de la méiose gamétique masculine d'après [8] : Les spermatogones B servent de point d'initiation et se différencient en spermatocyte primaire. Les cinq étapes de la prophase I sont suivies d'une première division cellulaire, donnant deux cellules haploïdes au sein desquelles les chromosomes sont composés de deux chromatides. Celles-ci seront séparées au cours de la méiose II donnant ainsi quatre spermatides qui après plusieurs étapes de maturation donneront les spermatozoïdes.

1. Méiose réductionnelle : (Figure : 2.5)

- a. **La prophase I** : cette longue étape dure 23 jours chez l'homme et peut être subdivisée en cinq phases successives : leptotène, zygotène, pachytène, diplotène et diacinèse.
 - i. **Leptotène** : condensation de la chromatine et formation des chromosomes.
 - ii. **Zygotène** : appariement des chromosomes homologues par paires appelés bivalents grâce à l'intermédiaire d'une structure multi-protéique : le complexe synaptonémal.
 - iii. **Pachytène** : ce stade dure 16 jours. Il est le plus long de la prophase I. C'est au cours de celui-ci, qu'a lieu l'échange de matériel génétique par le biais des crossing-over entre les chromatides non-sœurs appelées nodules de recombinaison (Figure : 2.4).
 - iv. **Diplotène** : la dissociation du complexe synaptonémal va permettre aux chromosomes homologues d'initier leur séparation. Certains sites d'appariement étroits nommés chiasmas demeurent néanmoins liés permettant une séparation plus progressive des chromosomes et réduisant ainsi le risque d'aneuploïdies (nombre anormal de chromosomes) [9].
 - 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.

2. Méiose équationnelle : (Figure : 2.6) la MII est similaire à une division mitotique et peut se décomposer en quatre parties distinctes :

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

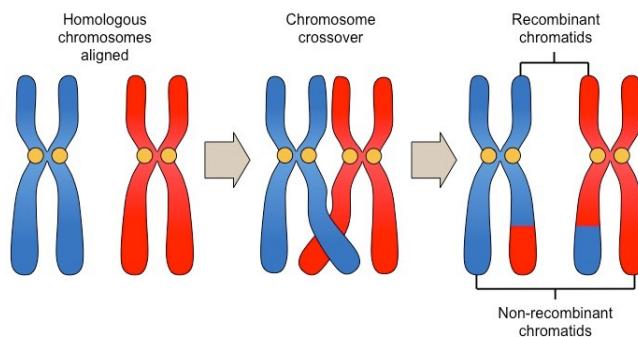


Figure 2.4 – Schéma simplifié d'un enjambement chromosomal (crossing-over).

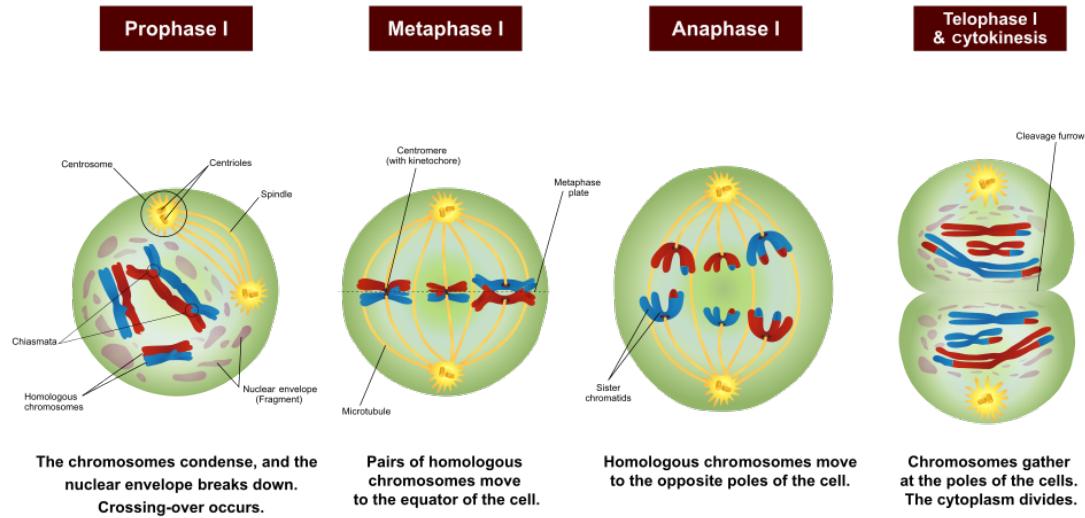


Figure 2.5 – *Les différentes étapes de la première division méiotique masculine* adapté d'après [10].

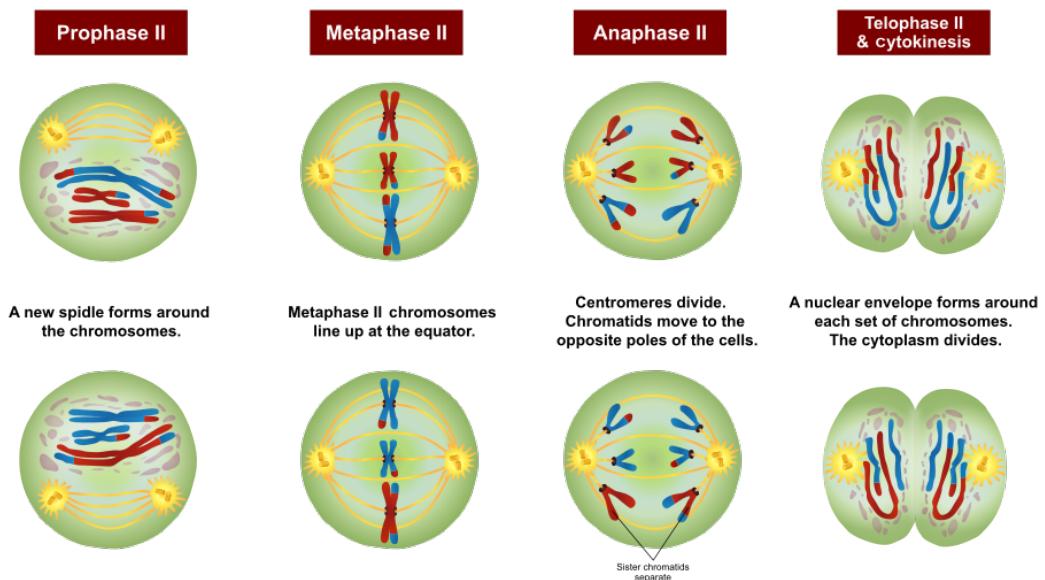


Figure 2.6 – *Les différentes étapes de la deuxième division méiotique masculine* adapté d'après [10].

2.1.4 La spermiogenèse

La spermiogenèse est la phase finale de la spermatogenèse. Elle dure environ 23 jours chez l'humain et peut être subdivisée en sept étapes (**Figure : 2.7**). La spermiogenèse définit la cytodifférentiation des spermatides en spermatozoïdes. C'est au cours de cette phase que les caractéristiques morphologiques et fonctionnelles du spermatozoïde seront déterminées [11]. Elle est caractérisée par trois événements majeurs : la formation de l'acrosome, la compaction de l'ADN nucléaire et la formation du flagelle. Le développement de l'acrosome et la formation du flagelle commencent au niveau des spermatides rondes [12]. Pendant l'élongation de la spermatide, le noyau se condense et devient hautement polarisé [13]. 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 : 2.7 - 1 et 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 [5, 14–17]. Plusieurs études animales ont pu démontrer le potentiel des spermatides rondes à donner la vie à des individus sains et fertiles, [18–20], la même chose ayant été également observée plus récemment chez l'homme [21] bien que le taux de fécondation et d'implantation soit extrêmement faible [22]. 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 : 2.7 - 3 et 4**) : à ce stade, l'acrosome va s'étendre le long du noyau lui donnant une forme plus allongée et la chromatine devient plus sombre. Un réseau de microtubules se forment autour du noyau créant ainsi la manchette qui participera également à l'allongement de la tête du spermatozoïde et permettra la migration des mitochondries vers la pièce intermédiaire du flagelle pour former le manchon de mitochondries [23]. Les spermatides en élongation peuvent aussi permettre la fécondation et d'initier des grossesses avec un meilleur taux de réussite que les spermatides rondes. De plus, ils engendreraient théoriquement moins de risques d'anomalies génétiques [22].

3. Les spermatides en condensation (Figure : 2.7 - 5 et 7) : c'est le stade final de la différenciation du 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 [24].

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.

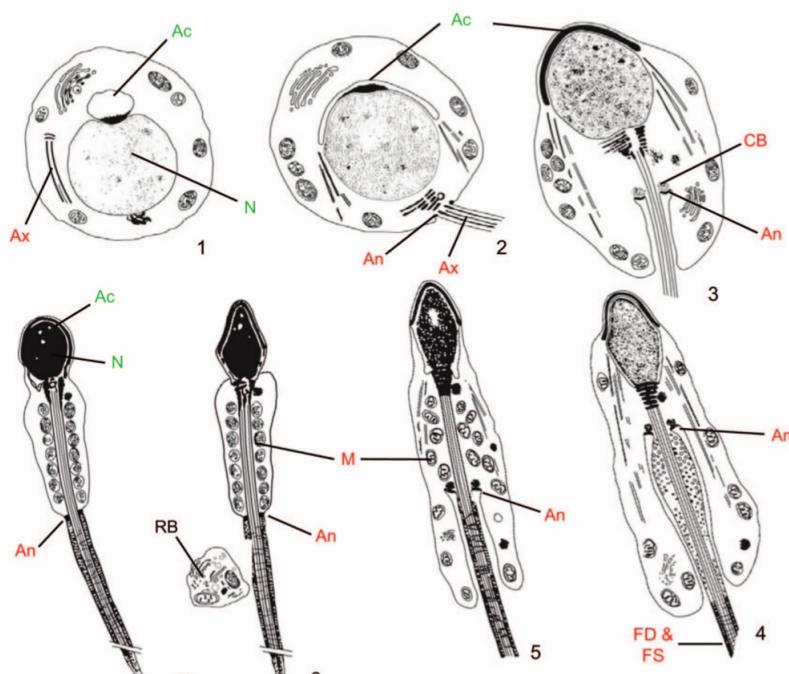


Figure 2.7 – Principales étapes et modifications structurales lors de la spermiogenèse d'après [25] : 1. La spermatide immature avec un gros noyau arrondi. La vésicule acrosomale est attachée au noyau, l'ébauche du flagelle n'atteint pas le noyau. 2. La vésicule acrosomale a augmenté de taille et apparaît 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.

2.2 Structure et fonction du spermatozoïde

Le spermatozoïde est une cellule hautement différenciée dont la taille, l'orientation et la symétrie sont déterminées. La morphologie générale du spermatozoïde éjaculé est similaire à celle du spermatozoïde testiculaire. Le spermatozoïde humain normal mature mesure environ 60 µm de long et est essentiellement constitué de deux parties : la tête et le flagelle (**Figure : 2.8**). 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. La fécondation d'un ovocyte par un spermatozoïde formera un zygote diploïde qui pourra se développer ensuite en embryon dans l'utérus féminin.

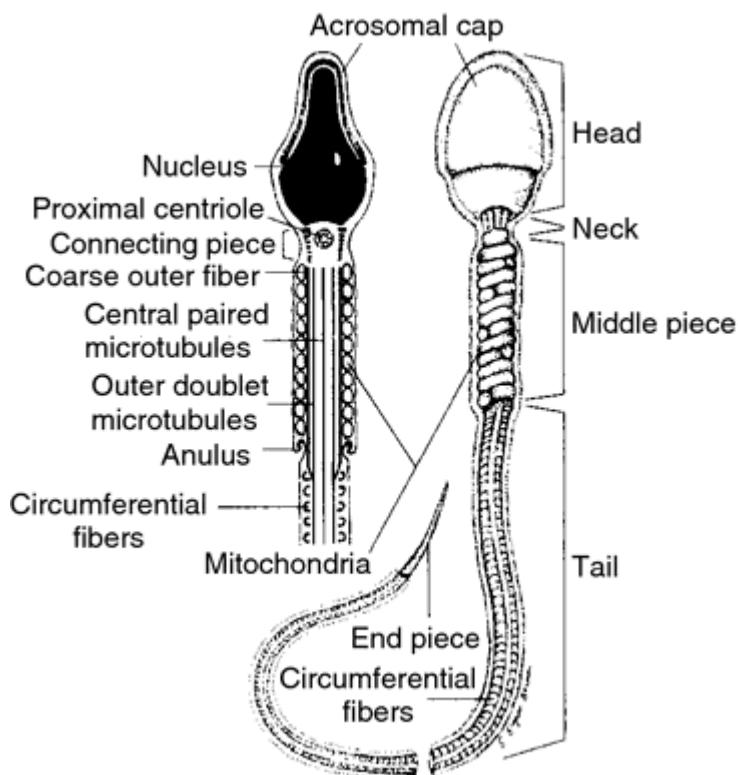


Figure 2.8 – *Anatomie simplifiée du spermatozoïde* d'après medical-dictionary.

2.2.1 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 spermiogenèse. 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 traverser la zone pellucide qui entoure les ovocytes. Ce processus de relargage est appelé réaction acrosomale.
2. **L'acropaxome** : l'acropaxome 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 [26]. C'est une structure transitoire qui disparaît pour être remplacée par la thèque périnucléaire dans le spermatozoïde mature.
3. **Le noyau** : c'est une structure cellulaire présente dans la majorité des cellules eucaryotes. Il contient l'essentiel du matériel génétique. Le noyau du spermatozoïde est caractérisé par une compaction extrêmement importante de l'ADN. Dans les cellules somatiques l'ADN est enroulé par unité de 146 paires de bases autour d'un octamère d'histones dit de cœur (H2A, H2B, H3 et H4) afin d'organiser les trois milliards de paires de bases du génome humain dans un noyau de quelques microns (**Figure** : 2.9). L'ADN des spermatides va subir une réorganisation chromatinnienne plus importante au cours de la spermatogenèse afin d'augmenter sa compaction. Ainsi, les octamères d'histones présents dans les cellules somatiques sont remplacés par les protéines de transition (TPN1, TPN2) puis par les protamines (PRM1, PRM2), deux protéines riches en arginine et en cystéine (**Figure** : 2.9). L'intégrité des deux protéines composant ce dimère est nécessaire pour la procréation [27]. 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 [28]. Le noyau du spermatozoïde est donc un noyau au repos, transcriptionnellement inactif [29].

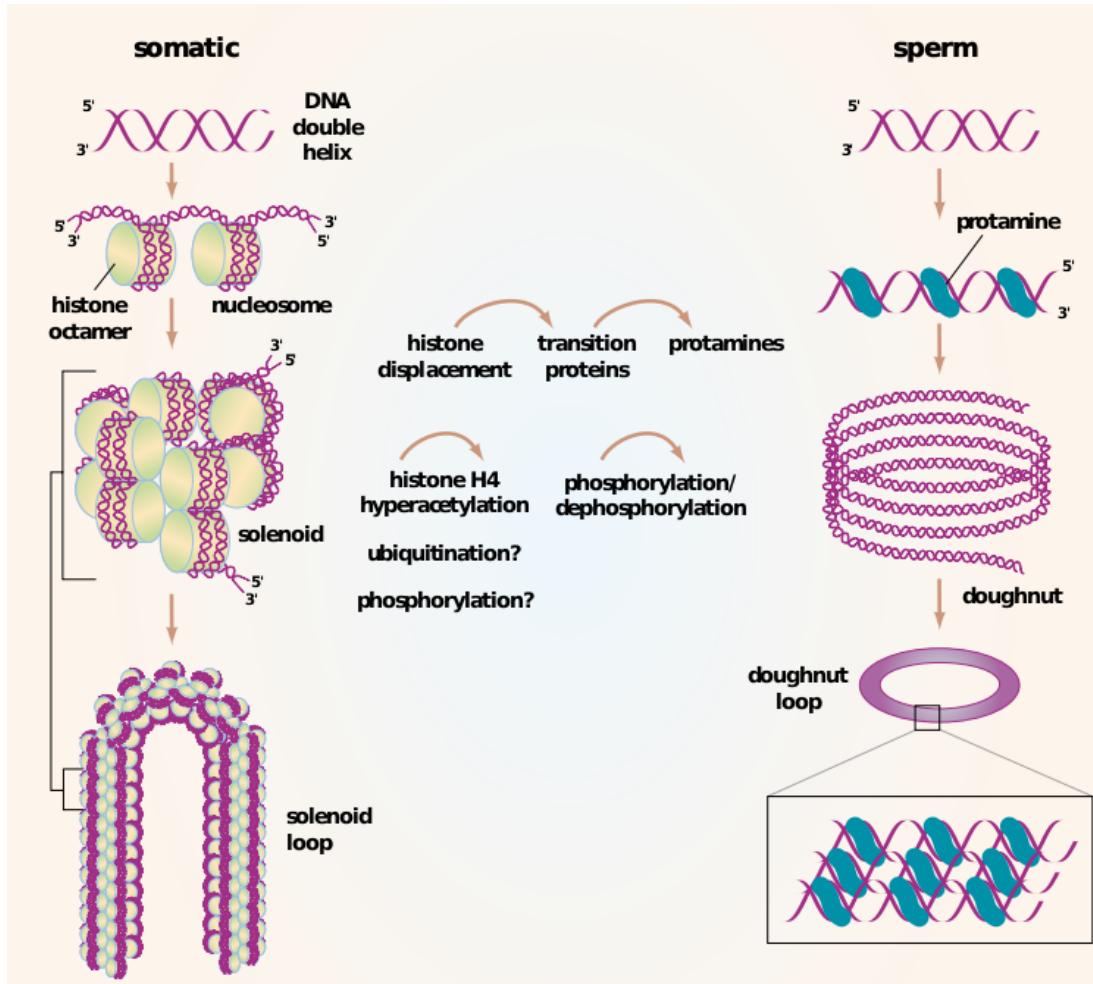


Figure 2.9 – Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes d'après [30] : Dans les cellules somatiques, l'ADN est enroulé sous forme de nucléosome. Les nucléosomes vont s'agencer entre eux pour former un solénoïde qui sera attaché à la matrice nucléaire par sa base. Dans le noyau spermatique les nucléosomes sont remplacés par des protamines qui vont compacter l'ADN sous forme de “doughnut”. Le remplacement des histones est facilité par des acétylations, des ubiquitinisations et des phosphorylations.

2.2.2 Le flagelle

Le flagelle représente la queue du spermatozoïde. Celui-ci permet, par mouvements d'oscillation à haute vitesse, le déplacement du spermatozoïde. Cette mobilité est générée par un cytosquelette interne extrêmement conservé durant l'évolution appelé l'axonème. Celui-ci est composé de neuf doublets de microtubules périphériques et de deux doublets internes [31] (**Figure : 2.10**), 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 radiaux. Les doublets externes sont également reliés entre eux par les complexes protéiques qui forment les dynéines externes et internes. Ce sont ces protéines qui en exerçant une contraction alternée permettent le mouvement du spermatozoïde.

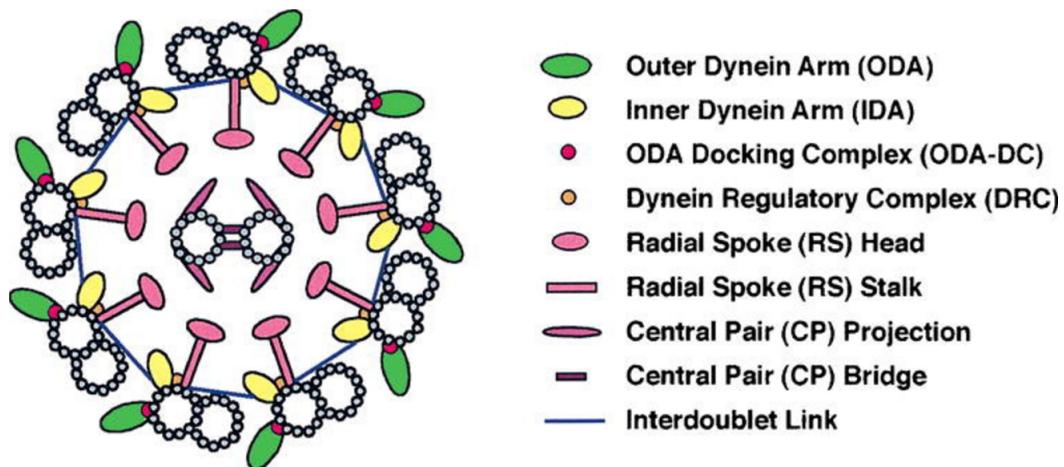


Figure 2.10 – Structure simplifiée de l'axonème d'après [31] : L'axonème est constitué de neuf doublets de microtubules périphériques reliés entre eux par des liens de nexine et d'un doublet central relié aux doublets périphériques par des ponts radiaux.

Le flagelle du spermatozoïde peut être divisé en trois parties distinctes (**Figure : 2.11**) :

1. **La pièce intermédiaire** : elle fait jonction avec la tête du spermatozoïde. Elle 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 est un ensemble de neuf faisceaux de fibres denses.

2. **La pièce principale** : ici, la gaine de mitochondrie a disparu ainsi que deux des faisceaux de fibres denses présents dans la pièce intermédiaire. On note cependant la présence d'une structure supplémentaire, la gaine fibreuse. Cette gaine entoure l'axonème et comporte deux épaississements diamétralement opposés, appelés colonnes longitudinales sur lesquelles s'insèrent les fibres denses 3 et 8. C'est le long de la gaine fibreuse qu'est produit la majorité de l'énergie nécessaire au glissement des microtubules [32].
3. **La pièce terminale** : elle est située au niveau de l'extrémité distale du flagelle et ne contient que l'axonème [31].

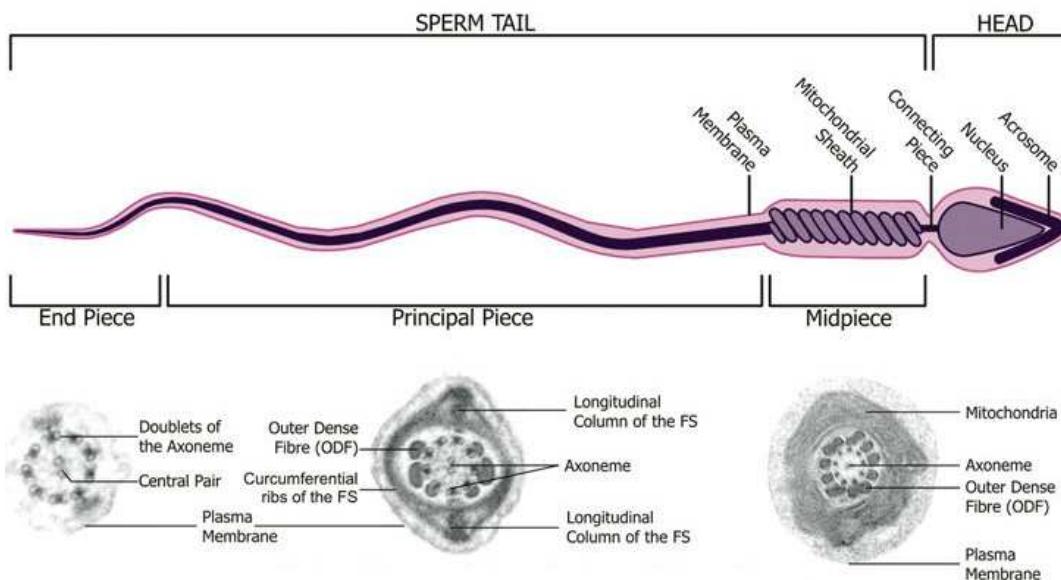


Figure 2.11 – Structure du flagelle d'un spermatozoïde
d'après [33] : Coupes transversales en microscopie électronique. Le flagelle se compose de trois parties : la pièce intermédiaire, contenant les mitochondries, la pièce principale et la pièce terminale. L'axonème, en position centrale, parcourt tout 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.

2.3 L'infertilité masculine

L'organisation mondiale de la santé définit l'infertilité comme étant : “*une pathologie du système reproductif définie par l'échec d'une grossesse clinique après 12 mois ou plus de rapports sexuels réguliers non protégés*” (Who.int. 2013-03-19. Retrieved 2013-06-17). L'étude de l'infertilité représente un des enjeux scientifique et médical majeur de ces dernières années. On estime qu'environ 10 à 15% des couples humains font face à des problèmes d'infertilité soit plus de 70 millions de personnes dans le monde [34]. Dans la moitié des cas, la cause sous-jacente serait masculine. On estime que les facteurs causaux sous-jacents de l'infertilité masculine peuvent être attribués à des toxines environnementales, des troubles systémiques tels que la maladie hypothalamo-hypophysaire, les cancers testiculaires et l'aplasie des cellules germinales. Les facteurs génétiques, y compris les aneuploïdies et les mutations de gènes uniques, contribuent également à l'infertilité masculine. Cependant, aucune cause n'est identifiée dans près de la moitié des cas. Comme nous avons pu le voir, la spermatogenèse est une succession de processus complexes qui s'effectue de manière coordonnée ; de fait la moindre altération génétique affectant une seule de ces étapes est susceptible d'entraîner un phénotype d'infertilité [35].

2.3.1 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 peut être détecté et quantifié dans des laboratoires spécialisés, par réalisation d'un spermogramme. Au cours de celui-ci, plusieurs critères tels que le volume de sperme sécrété, son pH, la quantité et la vitalité des spermatozoïdes qu'il contient seront évalués. La proportion de cellules immatures sera elle aussi analysée. Ces cellules rondes, se retrouvent à la fois dans l'éjaculat des individus ayant une quantité de spermatozoïdes “normale” [36], chez les individus présentant une quantité basse de spermatozoïdes [38, 39] ou en étant dépourvu [40]. Cependant, leur nombre augmente tandis que la quantité de spermatozoïde diminue [41].

Anomalies liées à la quantité spermatique

Chez l'humain, l'arrêt de la spermatogenèse est défini comme l'incapacité des cellules spermatogénétiques à devenir des spermatozoïdes matures. Elle peut survenir à n'importe quelle étape de la formation des cellules germinales. Les blocages méiotiques, au stade de spermatocyte I, sont les plus fréquents, suivis par l'arrêt au niveau des spermatides et moins fréquemment au niveau des spermatogonies [42].

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 [43]. Un arrêt de la spermatogenèse a été observé dans 4 à 30% des biopsies testiculaires des hommes présentant une oligospermie sévère [44–47]. Cet arrêt a longtemps été considéré comme sans espoir pour les couples désirant concevoir, jusqu'à l'émergence de l'injection mécanique d'un spermatozoïde dans l'ovocyte appelé *intracytoplasmic sperm injection* (ICSI) [48].
2. **L'azoospermie** : comme l'oligozoospermie, l'azoospermie est un phénotype d'infertilité masculine cette fois-ci caractérisé par l'absence totale de spermatozoïdes dans l'éjaculat. On distingue des causes excrétoires empêchant l'excrétion des spermatozoïdes, on parle alors d'azoospermie obstructive et des causes sécrétoires, les plus fréquentes, accompagnées d'un défaut de la spermatogenèse, on parle alors d'azoospermie non-obstructive.

Anomalies liées à la morphologie

Ces anomalies sont observables en effectuant un spermocytogramme. Plusieurs classifications ont été établies. Cependant, c'est la classification de David modifiée (**Table** : 2.12) qui est la plus répandue en France. Pour ce faire, on procède généralement à une observation de 100 spermatozoïdes au cours de laquelle l'ensemble des anomalies observées est relevé et quantifié permettant ainsi de définir un index d'anomalies multiples (nombre total d'anomalies/nombre de spermatozoïdes anormaux) révélant le nombre moyen d'anomalies par spermatozoïdes.

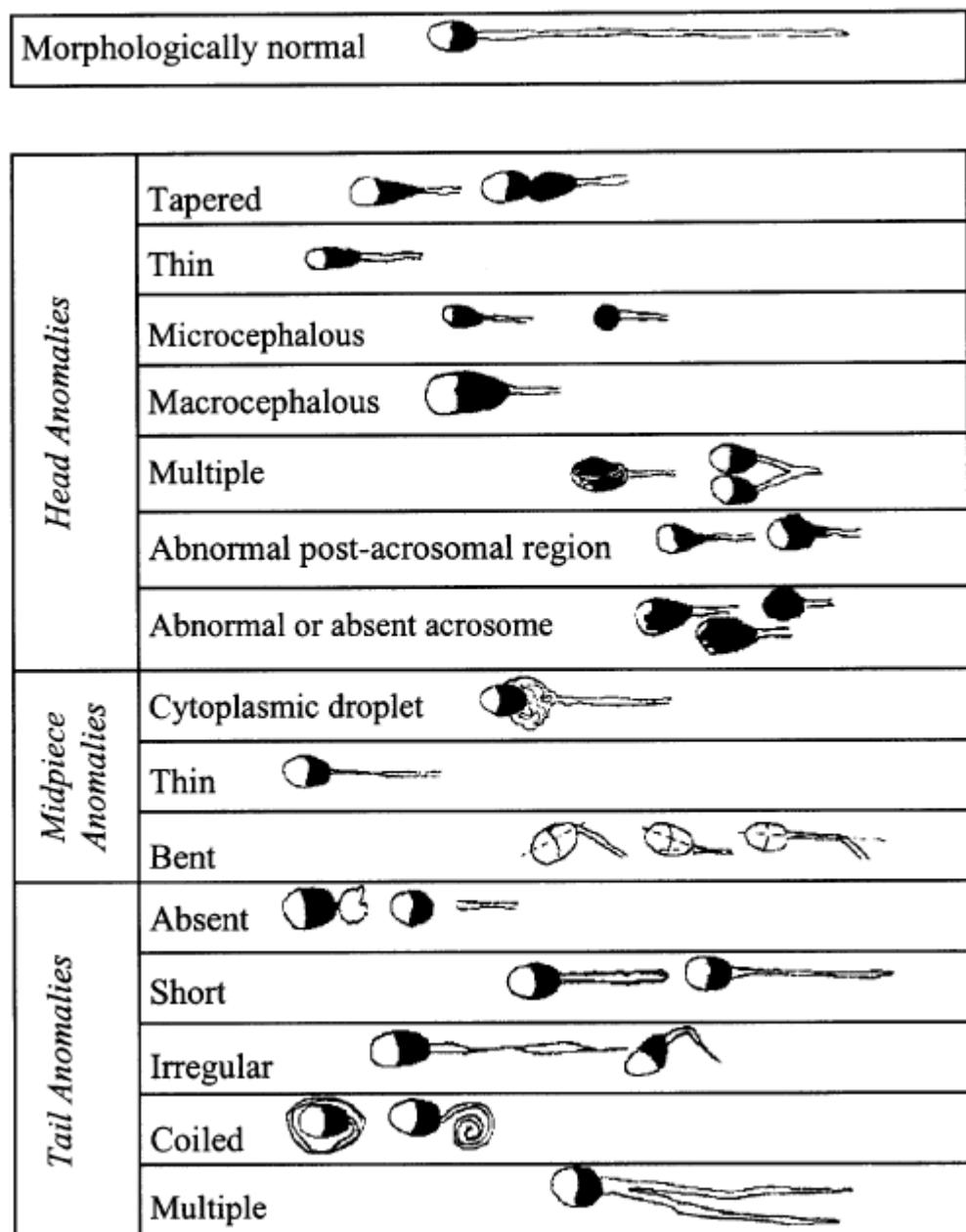


Figure 2.12 – *Classification morphologique de spermatozoïdes humains normaux et anormaux d'après [49].*

Anomalies liées à la mobilité

Le succès du passage du spermatozoïde le long du tractus génital féminin dépend en grande partie de la mobilité et de la vitesse du spermatozoïde [50, 51]. 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 [52].

2.3.2 La génétique de l'infertilité

Comme il a déjà été dit, il est estimé que 10 à 15% des couples humains font face à des problèmes d'infertilité. Par ailleurs, 30% des infertilités restent inexplicables et près de 40% ont des causes incertaines. Ainsi, l'infertilité masculine d'origine génétique pourrait concerner près d'un homme sur quarante [53].

Les causes fréquentes

1. **Les microdélétions du chromosome Y** : le chromosome Y est un petit chromosome atteignant une taille d'environ 53 Mb. Il est porteur de 78 gènes principalement impliqués dans la différentiation sexuelle masculine et la spermatogénèse [54]. De fait, le chromosome Y représente une région d'intérêt évidente dans l'étude de facteurs génétiques liés à l'infertilité masculine. L'évolution des technologies a permis de mettre en évidence des délétions invisibles au caryotype dans la région du facteur AZF (*Azoospermia Factor*). Cette région peut être subdivisée en trois sous-parties, AZFa, AZFb et AZFc (**Figure : 2.13**). Depuis plusieurs années, de nombreuses séries de patients azoospermiques ou oligozoospermiques ont été étudiées et publiées et tendent à montrer que les microdélétions du chromosome Y seraient responsables de 10% des cas d'azoospermie non-obstructive et chez 5% des cas d'oligozoospermie sévère (<5 millions de spermatozoïdes/ml) [55].

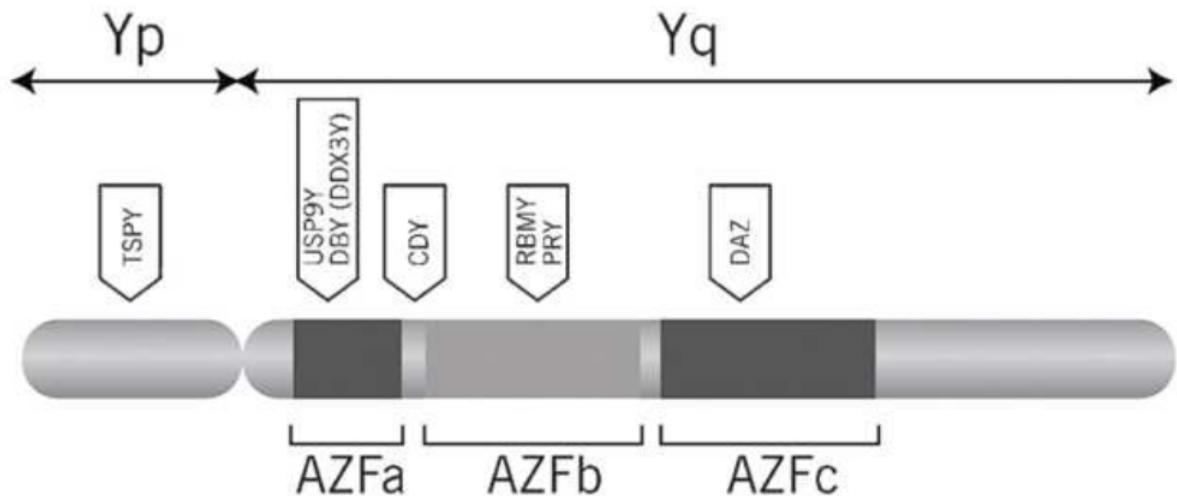


Figure 2.13 – Représentation schématique du chromosome Y adapté d'après [56] : Visualisation la région AZF ainsi que des trois sous-régions AZF a, b, c et des principaux gènes compris dans chacune des sous-régions.

2. **Anomalies chromosomiques** : des anomalies chromosomiques de nombre ou de structure impliquant les autosomes ou, le plus souvent, les gono-somes, peuvent être impliquées dans des cas d'infertilité masculine. Le pourcentage d'individus concernés varie entre 2 et 8% et peut atteindre 15% pour les patients azoospermiques soit 10 à 20 fois la fréquence retrouvée dans la population générale [57].
 - a. **Syndrome de Klinefelter** : le syndrome de Klinefelter (ou 47,XXY) fut décrit pour la première fois en 1942 par Harry F. Klinefelter. Il 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) [58] mais elle est environ 50 fois supérieure chez les patients infertiles azoospermiques [59].
 - b. **Les anomalies de structure** : les translocations et les inversions sont les anomalies de structure retrouvées le plus fréquemment chez les patients infertiles.
 - i. La translocation est définie comme l'échange de matériel génétique entre deux chromosomes non homologues. On en distingue deux types, les translocations réciproques et les translocations robertsonniennes. Les premières (**Figure : 2.14 - A**) décrivent un échange équilibré entre deux mêmes segments chromosomiques de deux chromosomes différents. Elles sont retrouvées 4 à 10 fois plus fréquemment chez les patients infertiles que dans la population générale [60]. Les secondes (**Figure : 2.14 - B**) impliquent deux chromosomes acrocentriques et sont caractérisées par la fusion entre les brins longs de deux chromosomes, les

brins courts étant perdus. Elles sont retrouvées chez 1.6% des patients oligozoospermiques et 0.09% des patients azoospermiques [56].

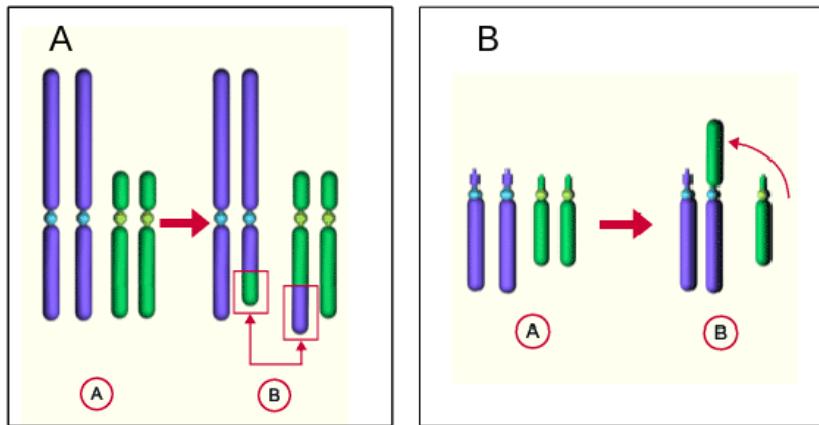


Figure 2.14 – Les différents types de translocation d'après embryology.ch : **A** : La translocation réciproque. **B** : La translocation robertsonniène

- ii. Les inversions chromosomiques caractérisent le mécanisme de cassure d'un fragment de chromosome suivi de son retour 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 [61].
 - c. **Autres anomalies chromosomiques** : parmi les anomalies chromosomiques responsables d'infertilité masculine, on peut par exemple citer les hommes de formule 46,XX. Ces patients sont généralement totalement infertiles et présentent une azoospermie par absence des sous- régions AZF a, b et c [62] bien qu'ils aient un phénotype masculin normal. Ces anomalies sont souvent le fait de la translocation du gène SRY sur un des chromosomes X du patient.
3. **Mutations du gène CFTR** : l'identification du gène *CFTR* (*Cystic Fibrosis Transmembrane conductance Regulator*) chez les patients atteints de mucoviscidose et présentant une agénésie bilatérale des canaux déférents (ABCD) a permis d'associer ce gène au phénotype d'azoospermie obstructive. Cette malformation serait responsable de 2% des cas d'infertilité masculine et de 25% des cas d'azoospermie obstructive [63].

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

Les nouveaux gènes

1. **Les anomalies quantitatives** : Ces dernières années, de nombreux gènes impliqués dans l'oligo/azoospermie ont été identifiés. En voici une liste non exhaustive :
 - a. **2014** : deux familles consanguines (non liées entre elles) ont fait l'objet d'une étude portant sur les gènes de l'azoospermie. À la suite de cette étude le variant c.1831C>T, p.R611X impactant le gène *AF4B* a été retrouvé à l'état homozygotes chez les trois frères azoospérmes de la famille 1. De même, la délétion homozygote c.1520_1523delAACAA, p.K507Sfs*3 sur le gène *ZMYND15* a été retrouvée chez les trois frères infertiles de la famille 2 [64].
 - b. **2015** : une analyse par CGHarray de huit patients azoospérmes a permis d'identifier une micro-duplication 20q11.22 englobant le gène *E2F1* chez un des patients. Dans cette même étude, des variations du nombre de copies (CNVs) dans ce gène ont été cherchées chez 102 autres patients. Parmi ceux-ci, sept portaient soit une délétion soit une duplication impliquant *E2F1* [65]. Ce gène était déjà considéré comme potentiellement associé à l'infertilité masculine puisque des études sur le modèle murin avaient déjà révélées que l'excès ou la déficience de la protéine E2f1, entraînait une atrophie testiculaire causée par une anomalies de la spermatogénèse [66]. De même, plusieurs variants dans le gène *TEX11* et *SYCE1* ont été décrits comme entraînant un arrêt de la méiose [67–69]. Une autre étude portant sur l'analyse de deux familles consanguines a permis d'identifier deux variants dans le gène *MCM8*. Le premier, c.1954-1G>A, entraîne un défaut d'épissage tandis que le second, c.1469-1470insTA, entraîne un décalage du cadre de lecture [70]. Une autre analyse familiale, cette fois-ci par séquençage exomique, a permis d'identifier le variant c.2130T>G, p.Y710*, causant un codon stop prématûr sur le gène *TEX15*, comme responsable du phénotype d'azoospermie de trois frères turcs issus d'une famille consanguine [71]. Ces résultats sont confortés par le fait que le phénotype de ces trois frères est très similaire à celui observé chez les souris *Tex15* KO, c'est-à-dire un arrêt prématûr de la méiose et une réduction significative de la taille des testicules [71].
 - c. **2016** : une étude démontre l'association de trois variants dans la séquence codante du gène *RAD21L* en se basant sur une étude statistique effectuée sur 38 japonais présentant un arrêt de la fertilité et 200 contrôles [72].
 - d. **2017** : plusieurs nouveaux gènes ont pu être liés au phénotype d'azoospermie. Par exemple, une analyse de trois familles par séquençage haut-débit a permis d'identifier trois gènes *MEIOB*, *TEX14* et *DNAH6* impliqués dans un phénotype d'azoospermie [73]. De même, une même délétion homozygote de quatre paires de bases menant à un décalage du cadre de lecture sur le gène *TDRD9* a pu être identifié chez cinq patients azoospérmes provenant

de la même famille [74].

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

- a. **La macrozoospermie** : Ce phénotype d'infertilité masculine rare est caractérisé par la présence de 100% des spermatozoïdes de l'éjaculat présentant une tête anormalement grosse ainsi que plusieurs flagelles. Il fut observé pour la première fois en 1978 [75], 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 [76]. Depuis, d'autres études ont permis d'associer d'autres variants sur ce même gène à ce phénotype [77]. Des anomalies du gène *AURKC* seraient ainsi responsables d'environ 83.7% des cas macrozoospermie chez des patients non apparentés [76]. Le gène *AURKC*, étant impliqué dans la méiose, conduit lorsqu'il est muté à un blocage de la première division méiotique, entraînant la production de spermatozoïdes tétraploïdes, c'est à dire, portant une quantité de matériel génétique quatre fois supérieure à la normale [78].
 - b. **La globozoospermie** : La globozoospermie est aussi un phénotype rare d'infertilité dont la prévalence est estimée à de 0,1%. Il fut identifié pour la première fois en 1971 et est caractérisé par la présence dans l'éjaculat d'une majorité de spermatozoïdes dépourvus d'acrosome, empêchant le spermatozoïde de franchir la zone pellucide de l'ovocyte et compromettant ainsi la fécondation [79–81]. 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*) [82] dont la protéine va, au cours de la spermatogenèse, fusionner avec les vésicules proacrosomales pour former l'acrosome [82, 83]. 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 [84]. cf globo
 - c. **Spermatozoïdes acéphaliques** : Ce phénotype rapporté plusieurs fois [85–87] 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* [88] qui avait précédemment été décrit comme localisant à la jonction noyau / flagelle du spermatozoïde [89].
3. **Le phénotype MMAF** : Le phénotype MMAF (*Multiple morphological abnormalities of the sperm flagella*) décrit les patients atteints d'asthenozoospermie dont les spermatozoïdes présentent de multiples anomalies morphologiques touchant en particulier les flagelles. Plus précisément, ce phénotype décrit les

asthenozoospermie résultant d'une mosaïque d'anomalies morphologiques au niveau du flagelle tel que l'absence totale de flagelle, des flagelles enroulés, courts, anguleux... [90, 91]. Récemment, le gène *DNAH1* (*Dynein Axonemal Heavy Chain 1*) codant pour une dynéine de la chaîne lourde de l'axonème a été retrouvé muté chez près d'un patient sur trois dans sa cohorte comportant 18 patients [91]. Deux autres études ont retrouvé des mutations dans le gène *DNAH1* chez des patients venant de Chine, d'Iran et d'Italie, laissant suggérer que ce gène est l'un des acteurs majeurs dans le syndrome MMAF [92, 93].

4. **Les échecs de fécondation du spermatozoïde :** Au moment de la fécondation, l'activation ovocytaire repose sur le relargage par le spermatozoïde de "facteurs spermatiques" qui déclenchent un signal de calcium, constitué d'oscillations Ca^{2+} . Ce processus est médié par une protéine spécifique du spermatozoïde, la *phospholipase C Zeta 1* (PLC ζ 1) codée par le gène *PLCZ1* [94, 95]. Plusieurs cas d'échec d'activation ovocytaire ont été liés à l'absence ou à la mauvaise localisation de la protéine PLC ζ 1. Malgré cela, aucune preuve génétique directe n'avait été reportée jusqu'à récemment où deux mutations au sein du gène *PLCZ1* furent retrouvées chez un patient [96] et un peu plus tard une mutation homozygote chez deux frères consanguins [97].

2.4 Généralités sur l'ovogenèse et l'ovocyte

Chez l'humain, la production d'ovocyte est un processus long commençant dès le développement embryonnaire. A ce stade, les ovocytes sont immatures et rentrent en phase de quiescence après avoir débuté l'étape de prophase I. Cette production est ensuite suivie d'une diapause de plusieurs dizaines d'années pour ensuite produire un ovocyte mature à chaque cycle menstruel. Les ovocytes vont dès lors compléter la MI et entamer la MII jusqu'au stade de la métaphase [98]. Cette phase est, chez la plupart des mammifères, la seule à laquelle l'ovocyte peut-être fécondé avec succès [99]. La MII se poursuivra ensuite dans le cas d'une fécondation [100].

La fécondation implique la fusion entre le spermatozoïde et l'ovocyte. Elle a été observée pour la première fois en 1876 par Oscar Hertwig chez l'oursin [101]. La première étape de la fécondation consiste en la liaison du spermatozoïde à la zone pellucide de l'ovocyte [102]. Cette liaison est permise chez l'humain grâce à quatre glycoprotéines : *zona pellucida sperm binding protein 1-4* (ZP1-4). Cette phase est ensuite suivie de l'activation ovocytaire. Cette étape est produite chez l'ensemble des animaux par la fertilisation d'un spermatozoïde. Elle entraîne une augmentation de la concentration cytosolique en Ca^{2+} [103]. Chez les mammifères, la fusion spermatozoïde-ovocyte est le déclencheur de séries distinctes d'oscillations du Ca^{2+} cytosolique nécessaires au développement normal de l'embryon [103, 104]. Ces observations ont rapidement permis l'émergence de l'hypothèse d'un facteur spermatique qui, lors de la fécondation, serait relargué et généreraient ces oscillations Ca^{2+} [103, 105]. Cette hypothèse est supportée par deux principales observations. Tout d'abord, le fait que la fusion des cytoplasmes du spermatozoïde et de l'ovocyte est le prélude à oscillations [106, 107]. Ensuite, le fait que l'injection d'un spermatozoïde ou d'extrait soluble de spermatozoïdes entraînait des oscillations Ca^{2+} similaires à celles observées lors de la fécondation [105, 108–111]. C'est en 2002 que la protéine PLC zeta (ζ) codée par le gène *PLCZ1* fut pour la première fois reportée comme étant, chez la souris, responsable de ces oscillations déclenchant ensuite une cascade de réactions dont découlent éventuellement l'activation ovocytaire et le développement embryonnaire [112] (**Figure : 2.15**).

Parmi les différents cas d'infertilité féminine, plusieurs études décrivent le cas de femmes produisant principalement des ovocytes non matures ou dégénérés. Ce phénotype peu décrit est connu sous le nom de *bad eggs syndrome* [113–116]. En 2016, une équipe a identifié des mutations hétérozygotes sur le gène *TUBB8*, une tubuline spécifique à l'ovocyte nécessaire à la création des fuseaux méiotiques, comme responsable du phénotype de déficience méiotique ovocytaire (*oocyte meiotic defect (OMD)*) dans une cohorte de femmes chinoises [117], faisant de *TUBB8* le premier gène formellement identifié dans le cadre d'OMD.

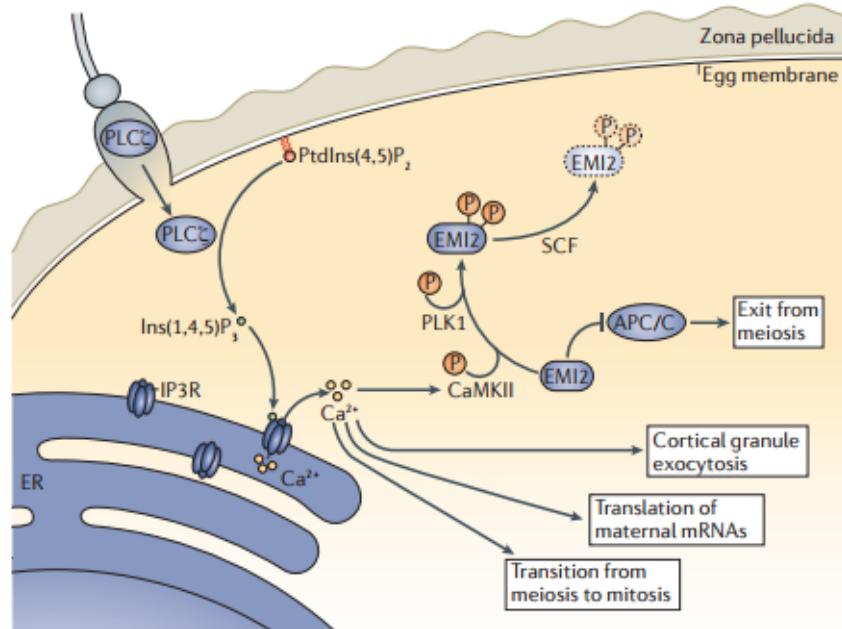


Figure 2.15 – La fécondation, liaison spermatozoïde-ovocyte et sortie de la méiose adapté d'après [118] : Suite à la fusion spermatozoïde-ovocyte, la protéine PLCZ1 induit la production de nosition 1,4,5-trisphosphate (Ins(1,4,5)P₃) qui se lie à son récepteur causant ainsi le relargage de Ca²⁺ qui, entre autre, fera sortir l'ovocyte de son stade de méiose pour rentrer en mitose

2.5 Les techniques d'analyses génétiques

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

2.5.1 Approche “gènes candidats”

L'approche “gènes candidats” consiste à rechercher des mutations chez un patient dans un ou plusieurs gènes cibles. Le choix des gènes cibles se fera en fonction de plusieurs critères. Le premier d'entre eux est l'étude de gènes reliés à des phénotypes proches du phénotype étudié dans différents modèles animaux et notamment murins. Dans ce cas, les mutations seront recherchées sur le gène orthologue humain [119]. 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, ceci étant dû, en grande partie, à l'hétérogénéité génétique des phénotypes étudiés, au nombre limité de patients testés [120] et aux connaissances souvent incomplètes sur le phénotype. De fait, cette approche a quasiment disparu au profit des méthodes à haut débit que sont les puces et le séquençage nouvelle génération (NGS). Néanmoins, cette méthode compte à son actif plusieurs succès retentissants avec dans le domaine de l'infertilité masculine, les gènes *SOHLH1* et *NR5A1* entraînant tous les trois un phénotype d'azoospermie [121, 122]. De même, le variant hétérozygote p.Ile215LeufsTer2 entraînant un décalage du cadre de lecture sur le gène *SYCP3*, fut décrit en 2003 comme entraînant la production d'une protéine *SYCP3* tronquée chez 2 des 19 hommes de l'étude [123]. Cette observation a mené l'équipe à conclure que ce variant hétérozygote entraînait un blocage méiotique menant à une azoospermie non-obstructive transmise sur le mode dominant. Néanmoins, ce variant ainsi que p.Ile215AsnfsTer27 sont retrouvés mutés à l'état hétérozygote chez respectivement 259 et 8 individus dans la base de données ExAC [124]. Puisqu'il est estimé qu'environ 1% des hommes souffrent d'azoospermie, cela signifierait que ces variants seraient responsables de 50% des cas. Ce qui n'est pas le cas dans l'étude de 2003 puisque seul 10,5% des patients sont porteurs du variant. Ces informations récentes permettent

ainsi de mettre en doute le liens entre le variant p.Ile215LeufsTer2 et le phénotype d'azoospermie de ces patients.

2.5.2 Les puces

Les puces à ADN furent initialement conçues dans le but de mesurer le niveau de transcription des transcrits provenant de plusieurs milliers de gènes lors d'une seule et unique expérience. Cette technologie a ainsi permis de déterminer des patterns d'expression de gènes à un état physiologique donné. L'analyse des "signatures" d'expression a ainsi permis de caractériser plusieurs cancers [125–128], mais aussi la réponse physiologique à plusieurs types de stimuli tel que la prise de certains médicaments [129].

Suite à cela, l'usage des puces à ADN dans le domaine biomédical s'est étendu pour ne plus être limité à la simple quantification de l'expression génique. Ainsi, cette technologie a également été utilisée afin de détecter des *single nucleotide polymorphisms* (SNPs) au sein de notre génome permettant notamment l'émergence du HapMap Project qui recense les SNPs de plusieurs milliers d'individus [130]. De même, l'utilisation des puces à ADN a permis la détection de *copy number variation* (CNVs).

Pendant plus de 10 ans, la grande qualité des puces, l'existence de protocoles d'hybridation standardisés ainsi que des algorithmes d'analyses robustes ont fait des puces à ADN l'outil d'analyse génomique le plus puissant avant l'arrivée du séquençage haut débit

Les puces à expression

L'utilisation principale des puces à ADN a été de mesurer l'expression des gènes dans un tissus donné. Dans cette application, l'ARN est extrait des cellules d'intérêt puis est généralement converti en ADNc. Dans un second temps, l'ADNc est hybridé à la puce qui subira par la suite une étape de lavage. Pour finir, l'intensité de fluorescence est mesurée à chaque spot de la puce et déterminera le niveau d'expression d'un gène.

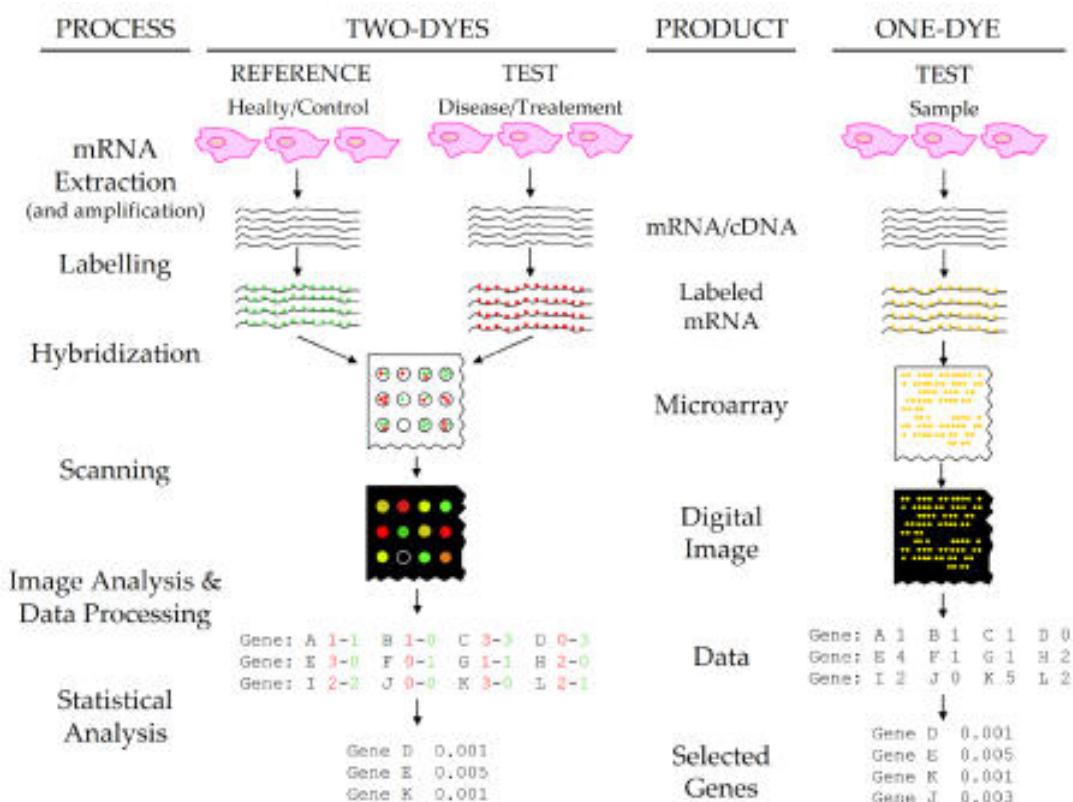


Figure 2.16 – Représentation schématique des méthodes d'analyse d'expression génique par puce à ADN d'après [131] : Présentation des méthodes à double et à simple colorant, respectivement à gauche et à droite. Pour les analyses à double colorant, une seule puce et nécessaire, les échantillons de la référence et du test sont mis en compétition sur la même puce, un signal de sortie vert indiquera une surexpression chez le test tandis qu'un signal rouge indiquera une sous-expression. Pour celles à simple colorant, deux puces sont nécessaires, une première pour la référence et une seconde pour le test. Les données des deux puces sont ensuite comparées pour déterminer quels sont les gènes différentiellement exprimés. Dans le cas de la CGH array, le principe est similaire, en remplaçant simplement l'ARNm par de l'ADNg.

Les puces à SNP, plateforme génotypage

Bien que leur utilité principale ait été d'analyser l'expression des gènes, les puces à ADN ont également été extrêmement utilisées comme moyen de génotyper les SNP (*single-nucleotide-polymorphism*). De nombreuses méthodes ont été mises en place pour cela ; cependant la plus employée est la méthode de discrimination allélique par hybridation telle qu'elle est utilisée par Affymetrix [132] malgré le “bruit de fond” causé par l’hybridation non spécifique dont elle souffre (**Figure : 2.17**).

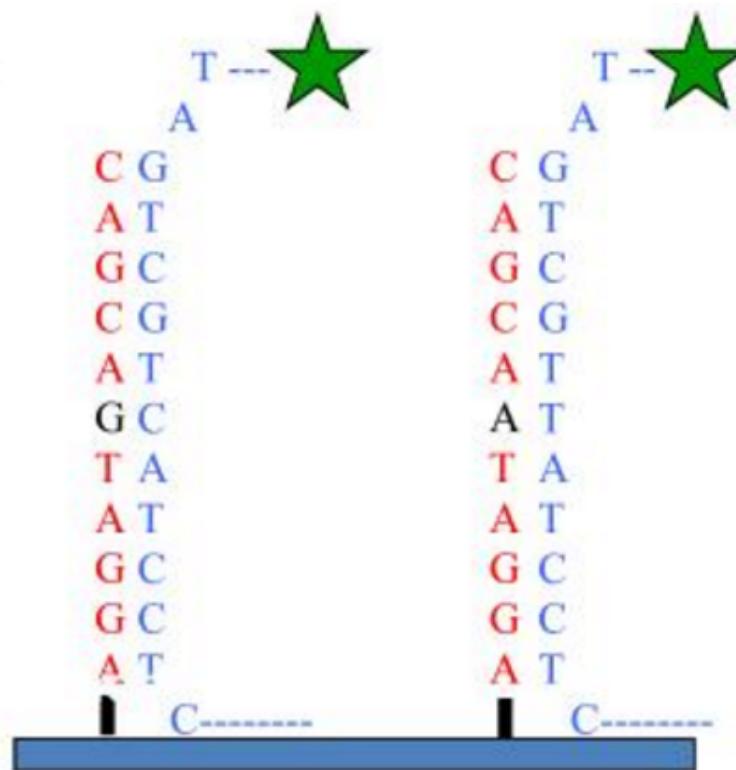


Figure 2.17 – Méthode de génotypage par discrimination allélique par hybridation d'après [133] : Des sondes complémentaires à chacun des allèles sont positionnées sur la puce. L'ADN génomique fragmenté et labélisé est mis en contact de la puce. Après nettoyage de la puce, l'analyse du signal émis par l'ADN génomique permettra de déterminer si l'individu est homozygote pour cette allèle (hybridation à une seule des deux sondes) ou bien hétérozygote pour cette allèle (hybridation aux deux sondes).

Les puces à indels

L'implication de réarrangements génomiques tel que des duplications, translocations ou délétions dans divers pathologies est bien connue. C'est afin de détecter ces réarrangements que la *Comparative Genomic Hybridization array* (CGH array) a été développée dès 1999 [134]. Son principe est très similaire à celui utilisé dans les puces à expression (**Figure : 2.16**) en remplaçant simplement l'ARN messager (ARNm) par de l'ADN génomique (ADNg). Ainsi, la présence d'un CNV sera facilement détectée en comparant le signal émis par un individu test avec celui émis par un contrôle.

Limitation

Bien que cette technologie ait été largement utilisée dans divers champs d'applications, elle présente deux limitations principales.

1. **Limitation n°1 :** Pour les génomes complexes (tel que les mammifères), il est difficile, si ce n'est impossible de *designer* une puce ne permettant pas de l'hybridation non spécifique. En effet, la séquence d'une puce prévue pour détecter le gène "A" pourra également détecter les gènes "B", "C" et "D" si ceux-ci présentent une forte homologie avec "A". Ce qui est particulièrement problématique dans le cas d'analyse de gènes d'une même famille.
2. **Limitation n°2 :** les puces détectent uniquement ce pour quoi elles ont été *designer*. Ainsi, si la solution que l'on hybride sur la puce contient des séquences d'ADN ou d'ARN pour lesquelles il n'y a aucune sonde complémentaire sur la puce, celles-ci ne seront pas détectées. Cela peut avoir de grandes répercussions puisque par exemple dans le cas des puces à expression, les gènes qui n'ont pas encore été annotés risquent de ne pas être représentés sur la puce.

2.5.3 Le séquençage NGS

Le terme séquençage de l'ADN fait référence à l'ensemble des techniques permettant de déterminer l'ordre des nucléotides A, T, C et G de l'intégralité ou d'une partie d'une molécule d'ADN. Avant de parler des nouvelles technologies de séquençage (NGS) faisons un bref historique du séquençage de l'ADN. En 1977 Frederick Sanger développe une technologie de séquençage d'ADN basée sur la méthode *chain-termination*. Ce procédé est désormais connu sous le nom de séquençage Sanger. D'autres méthodes furent développées à la même période, notamment celle de Walter Gilbert basée sur la modification chimique de l'ADN, cependant sa grande efficience et sa faible utilisation de la radioactivité permirent au séquençage Sanger de s'imposer comme référence dans la "première génération" de séquenceur à application commerciale et de recherche. Apparus 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 [135].

Contrairement à la méthode Sanger, le NGS est capable de "lire" des fragments d'ADN provenant d'un génome **entier**. On parle alors de séquençage de génomes entiers ou *whole genome sequencing* (WGS). Pour cela, la molécule d'ADN est "coupée" en plusieurs fragments d'une taille donnée. Ce sont ensuite ces fragments qui seront, après une étape d'amplification spécifique aux différentes plateformes, séquencés simultanément. C'est pourquoi on parle souvent de séquençage parallèle massif pour décrire le NGS. Le produit de ce séquençage est appelé *read*. Cette technologie est avantageuse de par la masse de *reads* qu'elle produit et par son faible coût par base séquencée [136]. 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, hormis dans le cas de la technologie PacBio, 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 plateforme contre une taille d'environ 800 nucléotides obtenus par Sanger (**Figure : 2.18**) ; c'est pour cela que les résultats du séquençage NGS sont appelés des *reads courts* ou *short reads*.

Étant donné que le NGS produit à l'heure actuelle des *reads* courts la notion de couverture est importante et représente l'un des critères majeur à considérer dans l'analyse des données [137]. La couverture est définie comme le nombre de *reads* qui, après l'étape d'alignement, se chevauchent les uns les autres au sein d'une région génomique spécifique. Par exemple, une couverture de 30x pour le gène XXXX signifie que chaque nucléotide de ce gène est chevauché par au moins 30 *reads* distincts.

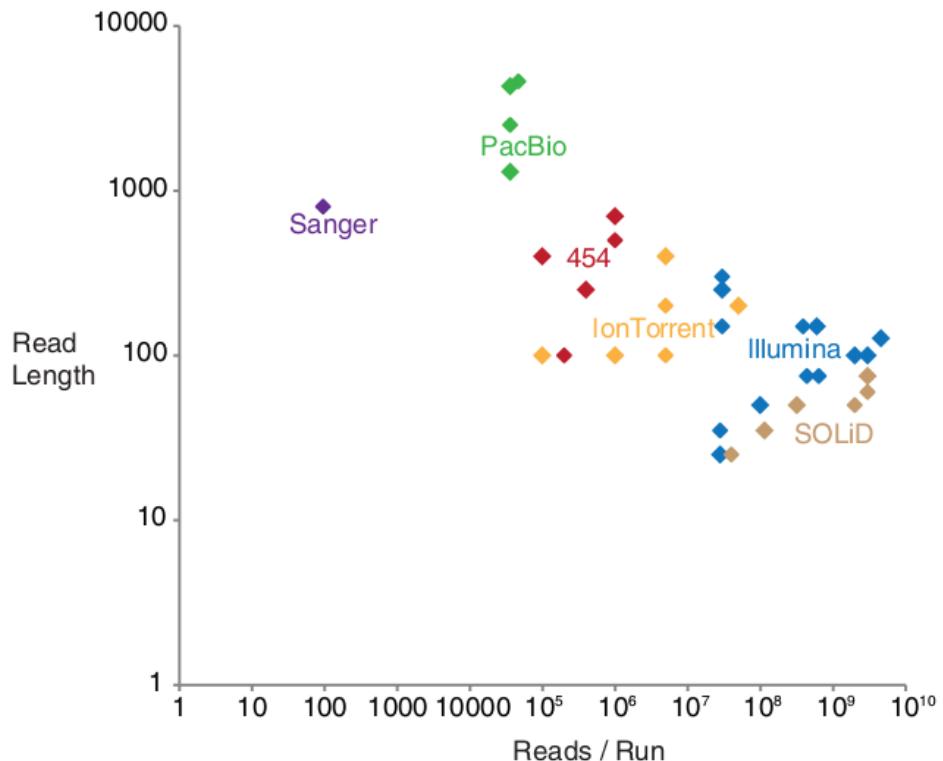


Figure 2.18 – Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée d'après [138] : Chaque point représente une plateforme de séquençage, la couleur détermine la marque du séquenceur.

La capture des parties à séquencer, avantages et inconvénients

Pour de nombreuses applications, il peut être intéressant de ne séquencer qu'une partie du génome et non pas son intégralité. Dans cette sous partie de génome ciblé on peut trouver par exemple : une région génomique spécifique à laquelle une pathologie a déjà été associée, l'ensemble des exons de certains gènes candidats, ou encore l'intégralité des exons de l'ensemble des gènes codant pour une protéine. Dans ce dernier cas, on parle alors de séquençage exomique ou *whole exome sequencing* (WES). Les principaux avantages du WES par rapport au WGS sont son coût réduit ainsi qu'une masse de données moins importante à stocker et à analyser. En effet, l'ensemble de l'exome ne représente qu'environ 1% du génome entier. On considère cependant que ces parties codantes contiennent plus de 90% des anomalies responsables de pathologies génétiques chez l'homme. Pour ces raisons, le WES est 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 [139]. Le procédé de séquençage

est identique au WGS, il est simplement précédé d'une étape d'enrichissement au cours de laquelle les exons sont capturés par hybridation à des sondes. De fait les exons capturés sont donc dépendants du kit de capture utilisé, cette technique permet donc de séquencer uniquement les exons connus et ciblés par les sondes. Il faut également noter que depuis quelques années, plusieurs études ont remis en cause l'intérêt du WES au profit du WGS, notamment car dans des conditions de séquençage standards, la proportion des régions codantes, définies à la fois par RefSeq et Ensembl, séquencée est plus importante dans le cas du WGS que dans le WES [140, 141]. 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 [141]. Ainsi, bien que le WES soit encore à l'heure actuelle le choix privilégié dans la majorité des études, la réduction des coûts de séquençage et du stockage des données, pourrait permettre prochainement au WGS de remplacer totalement le WES ainsi que l'ensemble des techniques impliquant la capture de séquences ciblées [141].

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

1. **La PCR en émulsion ou emPCR (Figure : 2.19 - a)** : Le patron d'ADN fragmenté simple brin est lié à une séquence adaptatrice complémentaire. Il est capturé par une gouttelette aqueuse appelée micelle contenant une bille recouverte d'adaptateur complémentaire à celui fixé sur le fragment d'ADN ainsi que tous les composants nécessaires à la réaction de PCR. En respectant un ratio nombre de molécules d'ADN / nombre de billes, on va fixer un seul fragment d'ADN sur chaque bille. Chacune de ces billes sera 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 : 2.19 - b)** : Les fragments d'ADN sont liés à des séquences adaptatrices et liés par une de leurs extrémités

à une amorce fixée sur un support solide. Du fait de la dilution, les molécules d'ADN se trouvent éloignées les unes des autres. L'extrémité libre du fragment interagit avec les amorces situées à proximité formant une structure en pont, d'où le nom de PCR en pont ou *bridge-PCR*. La PCR va alors synthétiser un deuxième brin complémentaire aux fragments immobilisés sur le support. En procédant à des cycles de température comme pour une réaction PCR classique, on obtient à l'emplacement de chaque molécule d'ADN un massif de molécules fixé sur la plaque, toutes identiques à la molécule initiale.

3. **Amplification par modèle mobile ou *walking-template* (Figure : 2.19 - c)** : L'ADN fragmenté est lié à un adaptateur et à une amorce complémentaire fixée sur un support solide. Le brin complémentaire du fragment sera synthétisé par PCR à partir de l'amorce fixée. La molécule double brin nouvellement formée sera ensuite partiellement dénaturée permettant à l'extrémité libre de se fixer à une séquence amorce voisine. Des amorces *reverse* sont ensuite utilisées pour resynthétiser un fragment d'ADN libre à partir des fragments fixés sur le support.

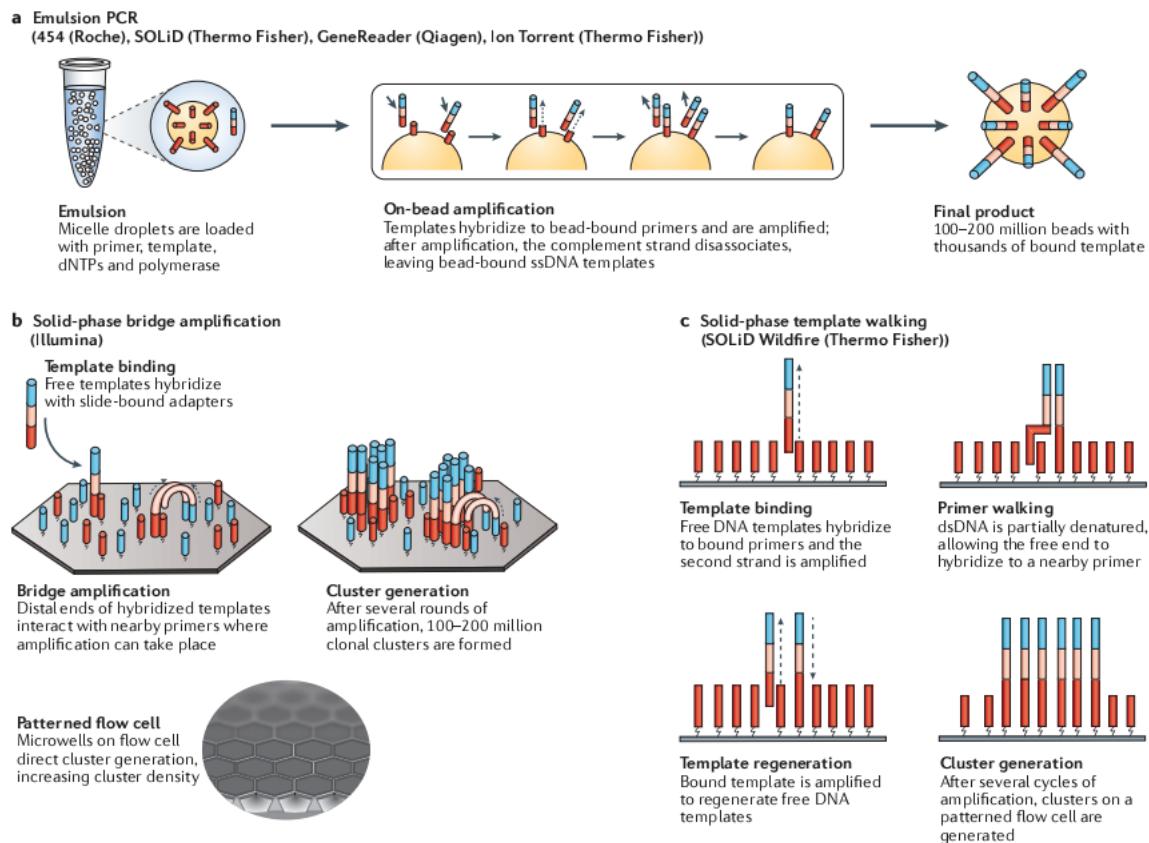


Figure 2.19 – Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS d'après [142] : a : PCR en émulsion. b : amplification par pont. c : amplification par modèle mobile.

La réaction de séquence

La réaction de séquence est l'étape suivant l'amplification. Elle 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 différenciés deux catégories de séquençage par synthèse [142] :
 - a. **Terminaison par cycle réversible, cyclic reversible termination (CRT)** (**Figure : 2.20**) : Cette méthode est caractérisée par l'utilisation de molécules terminatrices auxquelles le groupement 3' – OH est modifié de sorte à éviter l'elongation [143], on parlera de groupement 3' – bloqué. Une amorce liée au fragment d'ADN permettra l'initialisation du processus de polymérisation. À chaque cycle, un mix comprenant l'ensemble des quatre désoxynucléotides (dNTPs), préalablement labélisés par un fluorophore 3' – bloqué, est mis en contact du fragment. Après l'incorporation d'un unique dNTP au fragment, les dNTPs non liés sont éliminés et la nature du dNTP ajouté est identifiée grâce à son fluorophore. Le fluorophore et le groupement 3' – bloqué sont retirés permettant ainsi à un nouveau cycle de commencer.

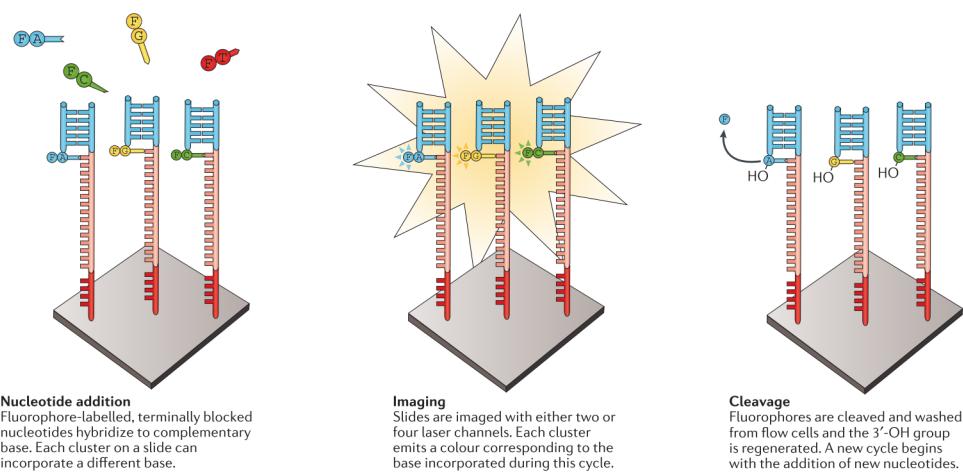


Figure 2.20 – Séquençage CRT tel qu'il est effectué par Illumina d'après [142] : a : ajout d'un dNTP labellisé par un fluorophore 3'-bloqué. b : identification du dNTP ajouté grâce au fluorophore. c : le fluorophore est clivé du dNTP et le groupement 3'-OH est reformé à partir du groupement 3'-bloqué permettant ainsi l'elongation.

b. **Addition de nucléotides uniques (SNA)** (**Figure : 2.21**) : l'initialisation de la méthode SNA est identique à celle de la méthode CRT. La différence se fait donc au moment de la phase d'elongation. Contrairement à la méthode CRT, le mix contenant les dNTPs ne contient qu'un seul type de dNTP. Quatre mixs différents sont donc présentés successivement au fragment d'ADN à séquencer, ceux-ci se fixeront uniquement s'ils sont complémentaires à la séquence. Ces dNTPs n'ont donc pas besoin d'être 3' – bloqué puisqu'un seul dNTP est ajouté à chaque itération. Après avoir présenté un mix, on vérifie si un dNTP s'est lié au fragment. Lors des séquences homopolymériques (plusieurs nucléotides identiques successifs dans la séquence), plusieurs dNTPs sont donc liés simultanément, cela sera détecté car le signal émis est proportionnel au nombre de nucléotides ajoutés.

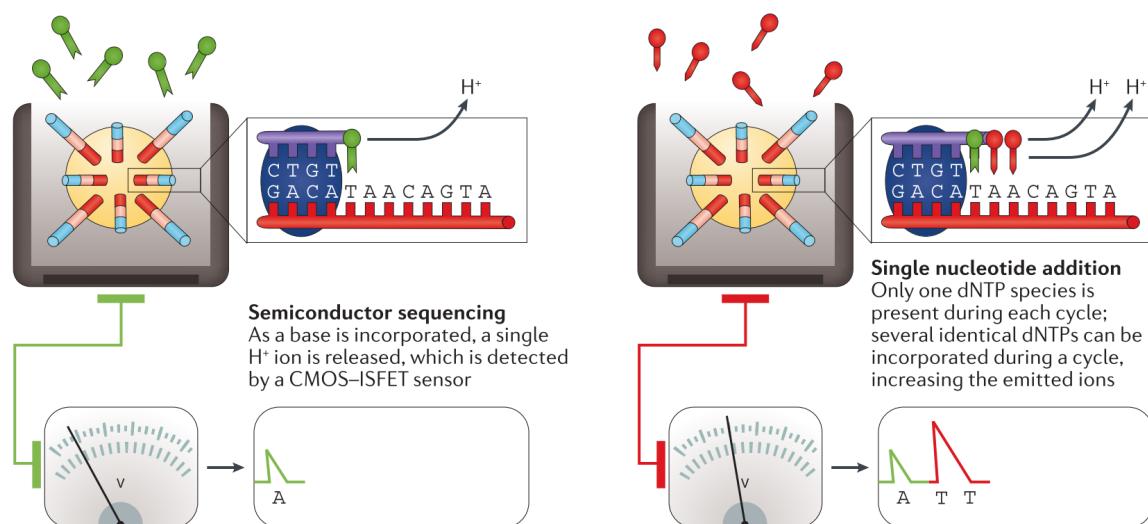


Figure 2.21 – Séquençage SNA tel qu'il est effectué par Ion Torrent d'après [142] : a : mise en présence du patron d'ADN à séquencer avec un mix contenant un seul type de dNTP, si le dNTP est complémentaire au patron, il se fixe et libère un proton permettant d'identifier la liaison. b : dans le cas d'homopolymère, autant de protons sont relâchés que de bases constituant l'homopolymère, le signal émis est donc plus fort permettant d'identifier le nombre des dNTPs liés.

2. **Séquençage par ligation (SBL)** (**Figure : 2.22**) : Par définition, cette méthode est basée sur l'hybridation et la ligation de l'ADN à une sonde liée à un fluorophore [144]. 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 liaisons phosphodiester. La sonde est constituée d'une ou deux bases connues, on parle alors de *one-base-encoded probes* ou de *two-bases-encoded probes* suivies d'une succession de bases "dégénérées" ou universelles, c'est à dire, de bases capables de s'apparier avec n'importe laquelle des quatre bases de l'ADN.

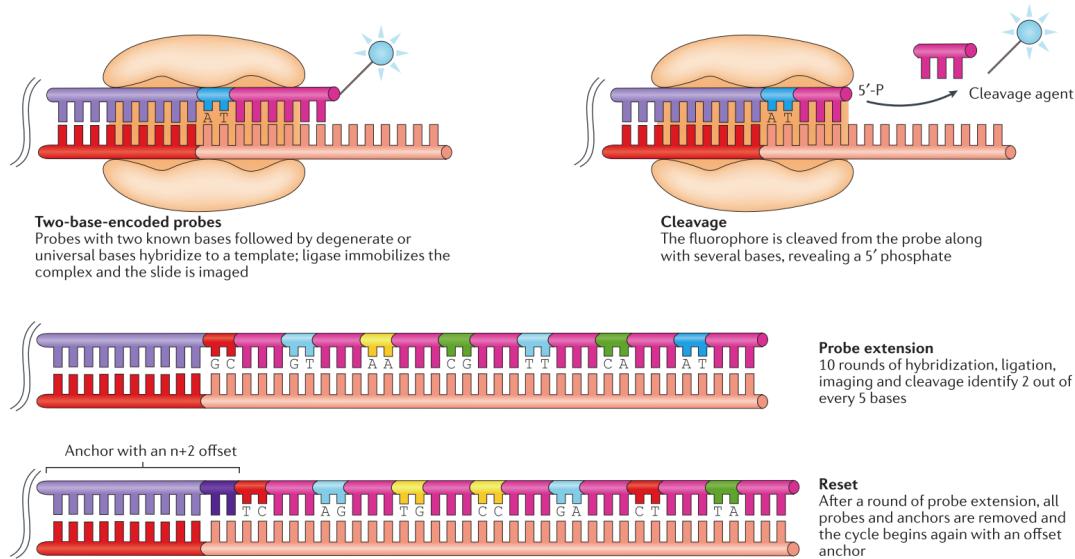


Figure 2.22 – Séquençage SBL tel qu'il est effectué par SOLiD d'après [142] : **a** : dans cette phase d'initialisation, un ensemble de sondes composées de deux nucléotides connus suivis de bases dégénérées et de bases universelles est reliée en 5' à un fluorophore. De fait qu'il n'y ait que quatre fluorophores différents, un même fluorophore est utilisé pour quatre combinaisons de dinucléotides parmi les seize possibles. **b** : après avoir été hybridé grâce à une ligase, aux brin d'ADN à séquencé, le fluorophore est identifié permettant ainsi de déduire l'une des quatre combinaisons possibles à ce locus. **c** : le fluorophore ainsi qu'une partie des bases non spécifiques sont clivés. **d** : les étapes **b** et **c** sont ensuite répétées 10 fois permettant à chaque fois d'identifier de réduire la liste des dinucléotides possibles au locus en question. **e** : les étapes **b**, **c**, et **d** sont répétées sur le même brin avec un décalage d'un nucléotide jusqu'à ce que chaque position ait été séquencée deux fois. **f** : en recoupant les informations obtenues à chaque itération du cycle, la séquence nucléotidique est reconstituée.

2.6 L'analyse bioinformatique des données de NGS

La stratégie consistant à séquencer en parallèle plusieurs millions de *reads* courts a engendré de nombreux et nouveaux défis bioinformatiques dans l'analyse et l'interprétation des données de séquençage et la recherche de variants dans le génome humain [145, 146]. Ces techniques ont été appliquées dans différents contextes, notamment la métagénomique [147], la détection de SNPs [148] et de variants structuraux [149, 150] mais également dans des études portant sur la méthylation de l'ADN [151], l'analyse de l'expression des ARNs messagers [152], dans la génétique du cancer [153] et la médecine personnalisée [154]. Cependant, pour l'ensemble de ces applications, la grande quantité de données générées par chaque analyse pose plusieurs défis informatiques [155]. 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 coût raisonnable. Ainsi, l'émergence du séquençage haut débit et notamment du WGS et du WES a permis de réunir une quantité jusqu'à présent inégalée d'informations sur les variations génétiques, et sur les gènes et leurs fonctions [156, 157]. Cependant, de par leur nature et leur quantité, l'acquisition de ces nouvelles données a engendré de nouvelles problématiques qui freinent les biologistes dans leurs recherches.

2.6.1 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 300 pb pour le NGS (**Figure : 2.18**). Depuis quelques années, un nouveau type de *read* est apparu, le *read paired-end*. Contrairement aux *reads* classiques (*single-end*), les deux extrémités (les *ends*) du fragment d'ADN sont désormais séquencées. La distance approximative 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 [158]. En plus de SNP, ce format permet de mettre en évidence des variants structuraux [159].

Le format FASTQ

Le format FASTQ (**Figure : 2.23**) est actuellement le format de données le plus couramment utilisé dans le cadre du séquençage haut-débit. Sa création est cependant antérieure à l'émergence du NGS puisqu'il fut inventé à la fin du XX^{ème} par Jim Mullikin au Wellcome Trust Sanger Institute alors que le séquençage commençait à prendre de l'ampleur grâce à des projets tels que le Projet Génome Humain. La quantité de données générées par ces programmes a nécessité une analyse automatisée. C'est ainsi que chaque base séquencée s'est vu associer un score de qualité appelé *Phred-score*. Chaque séquence générait ainsi deux fichiers, un fichier FASTA contenant les séquences et un fichier QUAL contenant les scores *Phred* associés à chaque base du fichier FASTA [160]. Plus tard, afin de n'avoir à manipuler qu'un seul fichier, les fichiers FASTA et QUAL furent fusionnés en ce que l'on appelle désormais le fichier FASTQ. Ce format est aujourd'hui le plus utilisé par les différents séquenceurs. On peut cependant noter certaines différences dans les formats FASTQ provenant des différentes plateformes, puisqu'à l'époque, aucune spécification officielle n'avait été donnée [160].

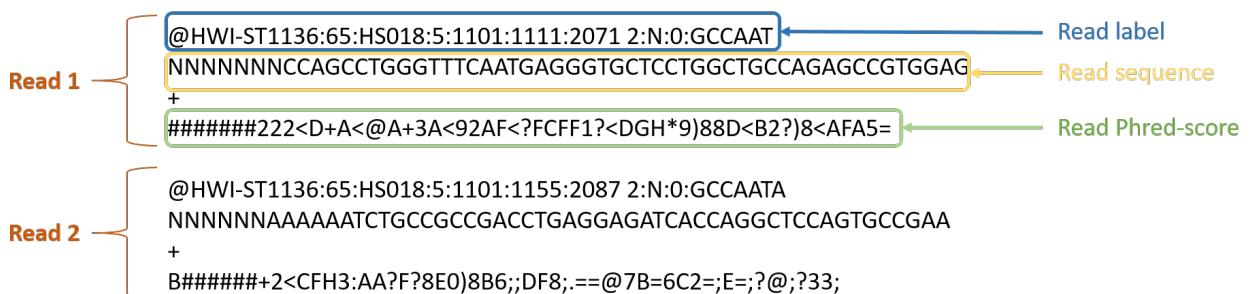


Figure 2.23 – Présentation d'un fichier FASTQ : Chaque *read* présent au sein d'un fichier FASTQ est composé d'un label, d'une séquence et d'un score de qualité associé à chaque nucléotide de la séquence.

2.6.2 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 censée représenter la séquence consensus d'une espèce donnée. Par exemple, la séquence de référence humaine GRCh37 (*Genome Reference Consortium human build 37*) a été créée à partir de treize 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 lequel les *reads* seraient les pièces et le génome de référence le modèle (**Figure : 2.24**). Elle constitue probablement l'étape la plus importante de l'analyse des données issues du séquençage haut débit [161] car elle est la base sur laquelle repose l'ensemble des étapes effectuées en aval, notamment l'appel des variants [162]. Cependant, l'étape d'alignement est sujette à de nombreuses erreurs dont certaines proviennent directement des erreurs survenues lors de l'étape de séquençage. D'autres, sont dues aux caractéristiques des régions séquencées comme par exemple les séquences répétées [163] qui pourront entraîner l'alignement d'un même *read* à plusieurs régions du génome [164]. De nombreux aligneurs ont émergé afin de répondre au mieux à cette problématique tel que Bowtie [165], Bowtie2 [163], BWA [166], NovoAlign, MAGIC [167]. De nombreuses études ont cependant montré de grandes différences entre ces aligneurs, au niveau du temps de calcul, de leur coût en mémoire et de leur taux d'erreur [168–170].

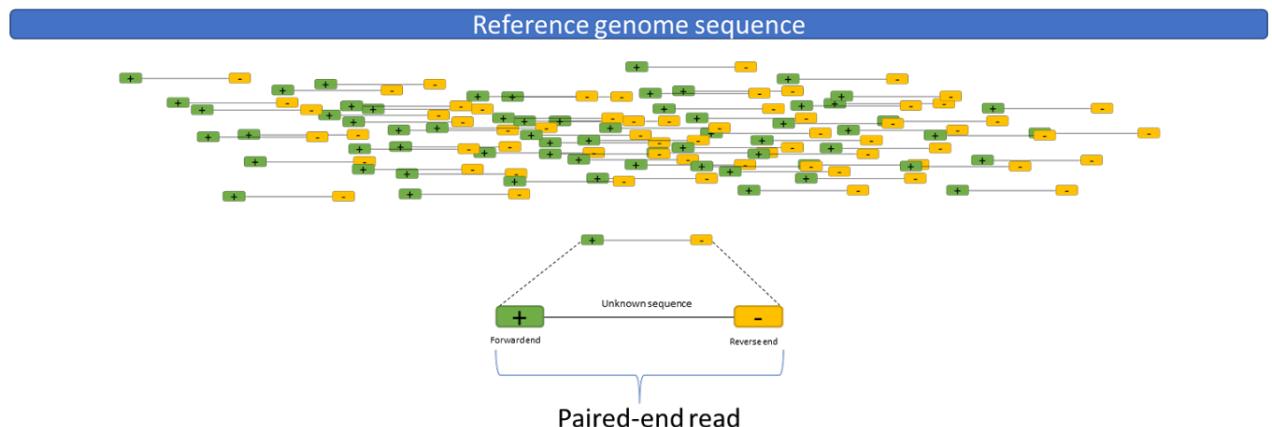
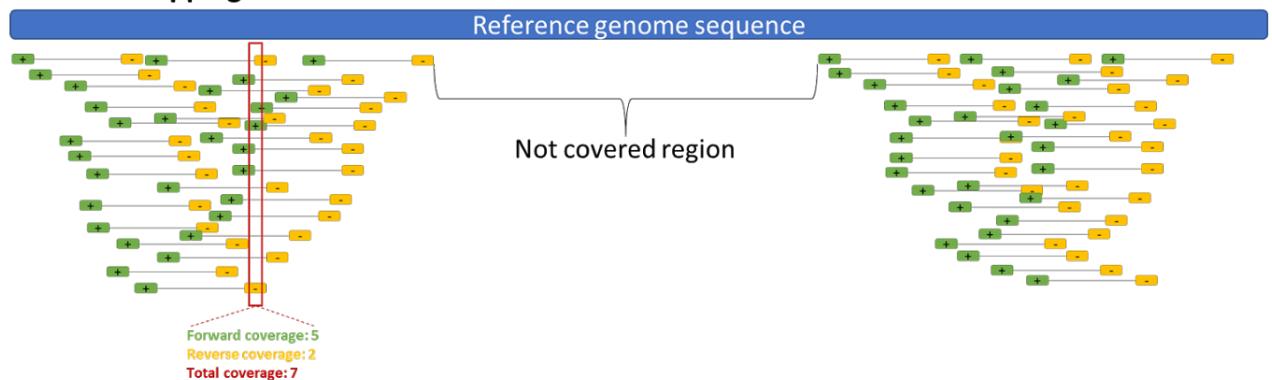
A: Before mapping**B: After mapping**

Figure 2.24 – Représentation schématique de l’alignement de reads paired-end : **A** : représentation du génome de référence ainsi que de *reads paired-end* avant l’étape d’alignement. Les *reads paired-end* sont composés d’une extrémité *forward* (en vert) complémentaire du brin sens du génome de référence et d’une extrémité réverse (en jaune), complémentaire du brin anti-sens du génome de référence. Chacune de ces extrémités est séparée par un insert de taille connue mais de séquence inconnue. **B** : après l’étape d’alignement, chaque *read* est positionné sur la région du génome avec laquelle il présente la plus grande homologie de séquence. Le nombre de *reads* différents recouvrant une même position du génome de référence est appelé couverture.

2.6.3 L'appel des variants

L'appel des variants, ou *variant calling*, fait référence à l'ensemble des méthodes permettant d'identifier des SNVs ou des indels à partir des résultats de l'alignement. Cette étape est souvent différenciée de l'alignement, cependant, les résultats de l'appel étant extrêmement dépendants de l'alignement, il est conseillé d'effectuer son appel en tenant compte de l'aligneur choisi [162, 171, 172]. On appellera variant toute différence de séquence observée entre un individu et la séquence de référence utilisée. Pour reprendre la comparaison avec la construction d'un puzzle, cette étape consiste à détecter quelles sont les pièces qui présentent des différences avec le modèle.



Figure 2.25 – Illustration schématique du processus d'appel des variants : Pour chaque position couverte, le pourcentage de *read* portant un allèle variant est analysé. Lorsque l'on est proche des 100% l'appel est homozygote pour le variant, lorsque l'on est proche des 50% l'appel est hétérozygote. Lorsqu'à une position donnée, peu de *reads* portent un variant, la cause est souvent une erreur de séquençage.

De nombreux logiciels d'appel des variants, ou *callers*, basés sur des algorithmes différents ont émergé ces dernières années pour répondre à cette problématique. Parmi les plus connus on note SAMtools [173], Genome Analysis Tool Kit - HaplotypeCaller (GATK-HC) [174], Freebayes [175], SOAPsnp [176] et Torrent Variant Caller (TVC). Les quatre premiers, peuvent être utilisés pour analyser des données provenant de tout type de plateforme de séquençage tandis que TVC a été développé spécifiquement pour les données provenant de Ion Proton. La pluparts de ces *callers* se basent soit sur des méthodes heuristiques, soit sur des méthodes probabilistes :

- 1. Les méthodes heuristiques :** dans ces méthodes, le nombre d'allèle variant et le nombre d'allèle référence observés sont comptés pour chaque individu sur l'ensemble des positions séquencées. Le génotype de l'individu à cette position donnée est ensuite déterminé en utilisant des valeurs seuils. Ces valeurs seuils peuvent être arbitraires, néanmoins ces méthodes montrent de meilleurs résultats lorsque celles-ci sont déterminées empiriquement [177].
- 2. Les méthodes probabilistes :** ces méthodes utilisent le théorème de Bayes afin de calculer pour chaque position la vraisemblance ($p(D|G)$) de l'ensemble

des génotypes (G) possibles (homozygote référence, hétérozygote et homozygote variant). Avec, D représentant l'ensemble des informations fournies par les données (couverture, qualité, nombre de *reads* variants / références...) pour un individu donné à une position donnée. Ainsi, en utilisant une probabilité *a priori* $p(G)$, le théorème de Bayes permet de calculer la probabilité *a posteriori* $p(D|G)$. Le génotype ayant la probabilité *a posteriori* la plus élevée est choisi, ainsi, $\hat{G} = \text{argmax}(p(G|D))$. Le ratio entre la plus forte probabilité et la seconde plus forte probabilité peut être ensuite utilisé pour mesurer la confiance dans le génotype appelé. De nombreux logiciels comme SAMtools, SOAPsnp et Genome Analysis Tool Kit - HaplotypeCaller (GATK-HC) utilisent cette approche. Ceux-ci utilisent la même méthode bayésienne pour calculer leurs probabilités *a posteriori* et ensuite déterminer le génotype. En revanche, ces logiciels diffèrent dans les probabilités *a priori* qu'ils utilisent. Par exemple, SAMtools et GATK fixent leur probabilités *a priori* pour un variant hétérozygote inconnu à 0.001 et à 0.2 pour un variant hétérozygote connu. Bien que, GATK utilise des étapes de pré-traitement et de post-traitement plus avancées, comme le ré-alignement local autours des éventuels indels. SOAPsnp utilise un model plus complexe pour déterminer les probabilités *a priori* en traitant, par exemple, les transitions ($A \leftrightarrow G$, $C \leftrightarrow T$) et les transversions ($A/G \leftrightarrow C/T$) différemment.

Les données issues de NGS peuvent présenter un taux d'erreur important. Ce taux d'erreur est multifactoriel et inclut 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 [137]. 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 [178]. 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 trois catégories et certains aligneurs auront tendance à être plus sujets à l'un de ces types d'erreur qu'à l'autre (**Figure : 2.26**) :

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.
3. Autres : incluent l'ensemble des autres types d'appel erronés indépendamment de l'allèle de référence.

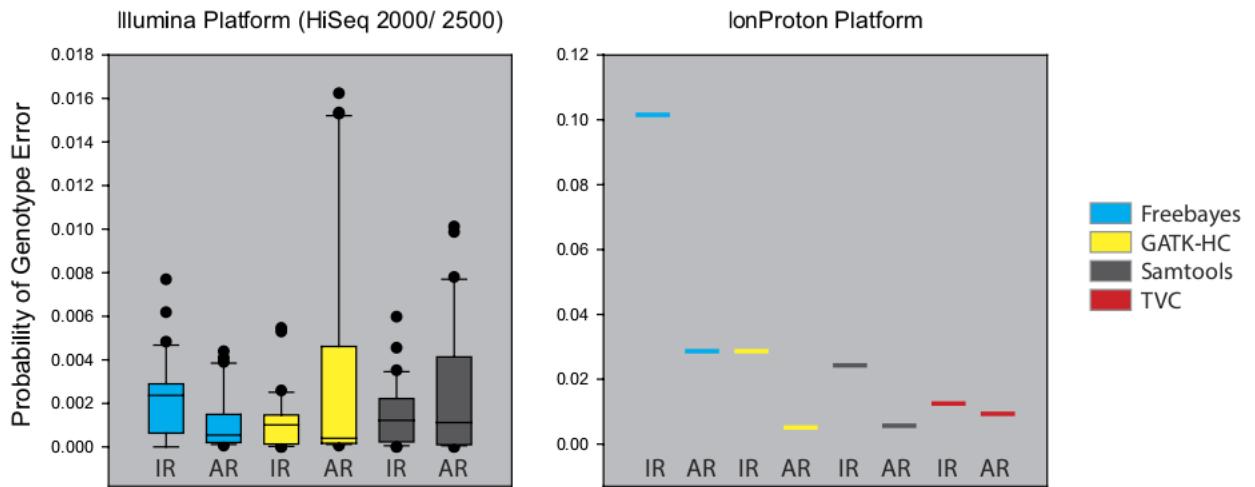


Figure 2.26 – Représentation des erreurs d'appel de type *IR* et *AR* en fonction de la plateforme de séquençage et du logiciel d'appel d'après [178] : Pour la plateforme Illumina, on peut voir que Freebayes favorise les appels variant-homozygote tandis que GATK-HC et Samtools favorisent les appels hétérozygotes. Pour la plateforme Ion Proton, les quatre logiciels entraînent des erreurs de type IR

De même que pour l'aligneur, le choix du logiciel d'appel est crucial car il existe de nombreuses différences dans les variants appelés par différents logiciels se basant sur les mêmes données brutes [179–181]. En effet, en 2013, une étude comparant les résultats de 5 *callers* montraient que seulement 57,4% des variants étaient appelés par les 5 *callers* et que 80,7% des variants étaient appelés par au moins 3 d'entre eux. Ce taux chutait drastiquement pour les indels puisque la concordance était cette fois seulement de 26,8% pour les indels non retrouvés par les 3 *callers* [180]. Ces résultats sont cependant à pondérer avec une étude de 2015 comparant 4 *callers* et montrant que 91,7% des SNVs séquencés sur une plateforme Illumina étaient appelés par 3 *callers*, cependant, pour les variants séquencés sur Ion Proton, seulement 27,3% des variants étaient appelés par au moins 3 *callers* et 57,4% des variants n'étaient appelés que par un seul des *callers* [178].

2.6.4 L'annotation des variants

Traditionnellement, les scientifiques et les laboratoires dans lesquels ils travaillaient développaient leur expertise dans un nombre de pathologies et de gènes associés limité. L'émergence du NGS est en train de remettre en cause cette pratique car la totalité de l'exome ou du génome peut permettre de couvrir tous les gènes en une seule et même analyse. De nombreux praticiens maintiennent cependant une spécialisation pour certains groupes de pathologies qui est précieuse pour l'analyse des données et l'obtention d'un diagnostic. En effet il est courant de retrouver entre 20.000 et 25.000 variants différents par exome [182]. Afin de pouvoir lier un variant à une pathologie, il est indispensable d'annoter cet ensemble de variants, c'est à dire d'associer à ces variants l'ensemble des informations qui les caractérisent afin de pouvoir les replacer dans leur contexte biologique. Ces informations serviront ensuite d'indicateur afin de filtrer ou de prioriser un variant. Cette dernière étape de l'analyse est, elle aussi, cruciale puisqu'elle permet de réduire le nombre de variants à considérer. On peut généralement distinguer deux niveaux d'annotations d'un variant (**Figure : 2.28**) :

1. **Au niveau du variant** : Ce niveau d'annotation regroupe l'ensemble des informations **spécifiques** à un variant
 - a. **Informations issues des résultats du séquençage** : la couverture du variant ainsi que la qualité qui lui est associée peuvent permettre de considérer un variant comme étant fiable ou non. Le génotype associé à ce variant est également une information importante.
 - b. **La fréquence du variant dans la population générale** : l'émergence du séquençage haut-débit a permis de gros consortiums tels que ESP6500 (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA), 1KG [183]. Ces consortiums ont pu mettre à disposition du public des 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) [124] qui n'a effectué aucun séquençage mais qui a regroupé les données de plusieurs gros jeux (notamment 1000Genome et ESP) afin de leur appliquer la même analyse bioinformatique harmonisant ainsi les données provenant de 60.706 individus non apparentés. Cette masse d'information permet de se faire une idée de la fréquence d'un variant dans la population générale et même au sein de sous-populations humaines. On considère qu'un variant fréquent ne peut pas être délétère, sinon il aurait été contre-sélectionné au cours de l'évolution.
 - c. **Son impact sur le transcrit** : Dans la plupart des analyses phénotype-génotype, les chercheurs se limitent aux variants chevauchant des transcrits codants pour une protéine. Il est donc important de savoir l'impact d'un variant sur ce transcrit, c'est à dire si le variant va causer une mutation synonyme, un faux-sens ou une mutation tronquante. Des logiciels tels que *Variant Effect Predictor* (VEP) [184], SnpEff [185] ou encore ANNOVAR

[186] vont prédire l'impact qu'aura un variant sur les différents transcrits qu'il chevauche. D'autres logiciels tel que SIFT [187], PROVEAN [188], Polyphen2, ou encore CADD vont, eux, chercher à prédire la pathogénicité de ce variant, c'est à dire la probabilité que ce variant soit délétère pour la fonction de la protéine. Bien que cette information soit importante, elle est à pondérer, étant donné le peu de concordance qu'il existe entre les prédictions de ces différents logiciels (**Figure : 2.27**).

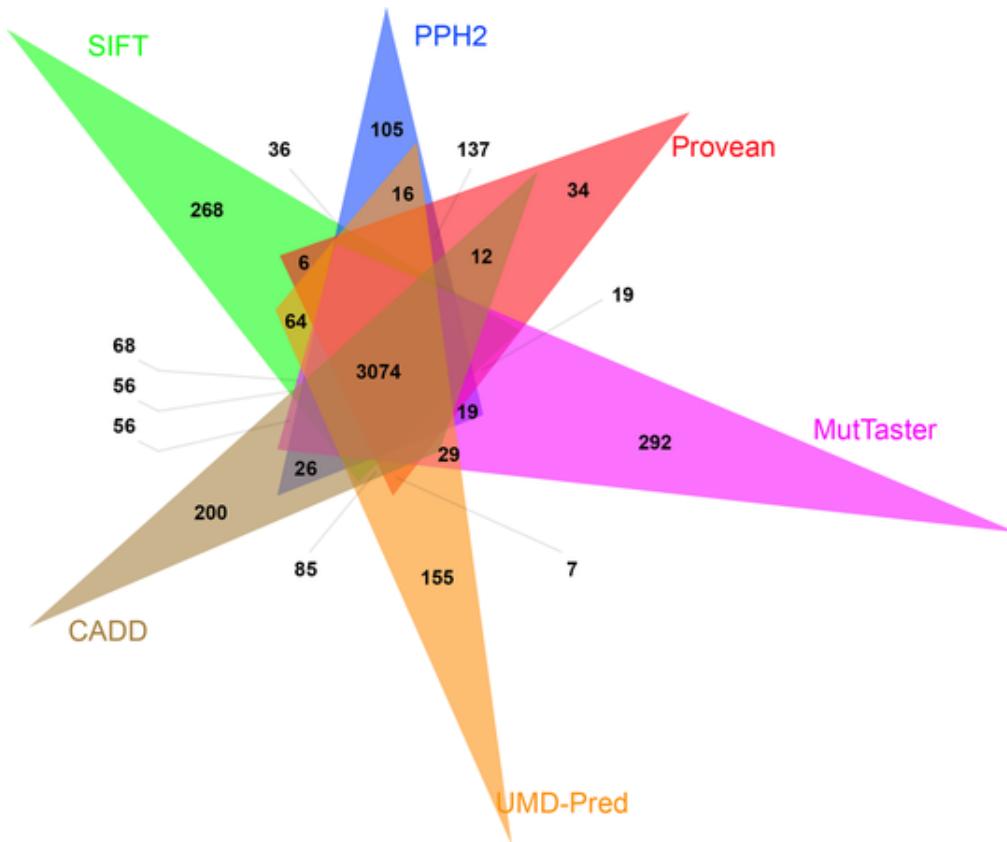


Figure 2.27 – Diagramme de Venn des prédictions de pathogénicité de variants de six logiciels d'après [189] : Les six logiciels utilisés sont : CADD [190] (marron), SIFT [187] (vert), PolyPhen2 [191] (bleu), Provean [188] (rouge) MutationTaster [192] (violet) et UMD-Predictor [193] (orange).

2. Au niveau du gène (ou transcrit) : L'annotation au niveau du gène consiste à récupérer l'ensemble des informations disponibles non plus sur le variant uniquement mais sur le ou les gènes qu'il impacte. Ce “dézoom” permet d'ajouter des informations complémentaires particulièrement utiles notamment lorsque peu d'informations sont disponibles sur le variant lui-même. En pratique, la plupart des variants connus pour impliquer une pathologie sont des variants privés, c'est à dire spécifiques à une famille ou à un individu, limitant ainsi la quantité d'information disponible sur ce variant. Élargir l'annotation au niveau des gènes 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 ANalysis THrough Evolutionary Relationships* (PANTHER) [194] qui permettent par exemple de classer une liste de gènes en fonction de leurs fonctions moléculaires, des processus biologiques et des voies de signalisation dans lesquelles ils sont impliqués. On peut également noter *the Human Phenotype Ontology project* (HPO) [195] qui fournit un vocabulaire standardisé pour les anomalies phénotypiques observées dans les pathologies humaines et une liste de gènes connus pour être associés à ces phénotypes. Plus récemment, on a pu voir émerger des “scores mutationnels” tel que RVIS [196] ou encore le pLI [124]. 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 mendélienne” devraient être moins tolérants aux variations que les autres.

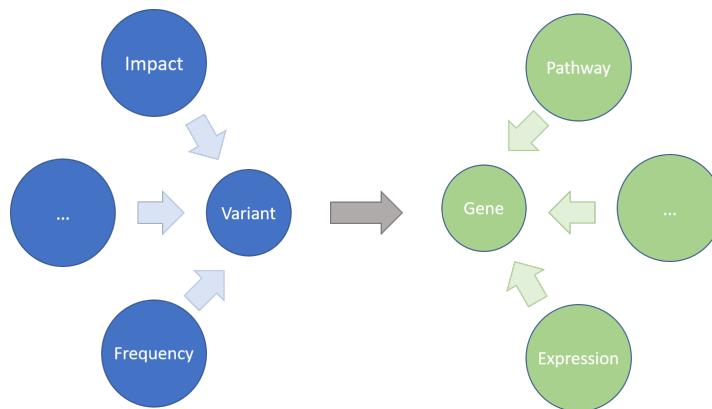


Figure 2.28 – Représentation simplifiée du processus d'annotation : On peut observer deux niveaux d'annotation, le premier est l'annotation des variants consistant à ajouter un maximum d'information sur un variant spécifique (sa fréquence, son impact...). La deuxième est au niveau du gène, consistant à récupérer pour les gènes impactés par les variants l'ensemble des informations disponibles tel que les processus biologiques dans lesquels il est impliqué, ou encore son expression tissulaire.

2.6.5 Le filtrage des variants

L'étape de filtrage a pour principal objectif de restreindre le nombre de variants obtenu à l'issus de l'appel afin que ceux-ci puissent être analysés par un être humain. Pour cela il utilise l'ensemble des informations obtenues lors de l'étape d'annotation afin de filtrer les variants ayant le moins de risque d'être responsables du phénotype. Communément, les variants ayant une forte fréquence dans les bases de données ExAC, ESP6500 ou encore 1KG sont filtrés avec l'idée sous-jacente que des variants observés fréquemment dans la population ne peuvent être responsables de phénotypes sévères.

Comme nous l'avons vu, le développement d'outils permettant l'analyse et le filtrage des données NGS est extrêmement important puisqu'il permet aux biologistes de faire face à la masse de données générée par le séquençage haut-débit l'aident ainsi dans ses prises de décisions. Il est à noter que la plupart de ces données filtrées sont extrêmement dépendantes du jeu de transcrits utilisés. Les prédictions seront donc différentes si l'on se base sur les transcrits RefSeq, Ensembl ou UCSC [197] bien que les transcrits du *Consensus Coding Sequence project* (CCDS) soient bien représentés par ces trois listes [199]. De même, pour une même liste de gène, de nombreuses différences seront observées en fonction du ou des logiciels de prédition utilisés [189, 197].

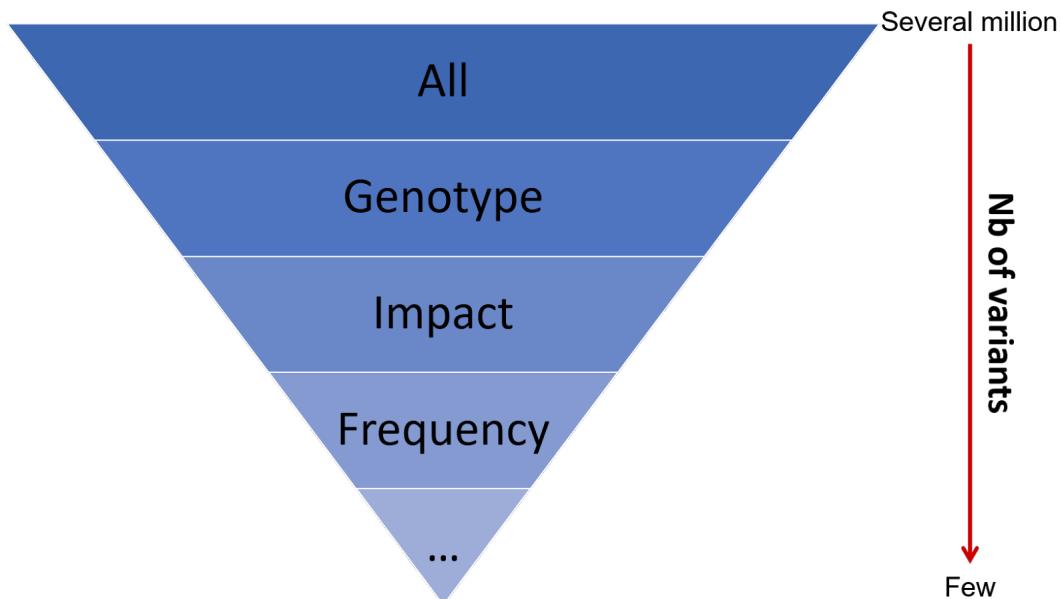


Figure 2.29 – Représentation simplifiée du processus de filtrage des variants : L'ensemble des annotations ajoutées lors de l'étape précédente servent alors de support pour filtrer (ou non) les variants. Il est par exemple commun de filtrer les variants ayant une forte fréquence dans la population générale ou encore ceux ayant un impact faible sur la protéine. De même dans le cas d'étude sur des pathologies ayant un mode de transmission récessif, les variants hétérozygotes pourront également être également filtrés.

2.6.6 Conclusion NGS

En moins de 10 ans, les technologies NGS sont passées du séquençage de panels de gènes (environ 100 Mb pour le Roche GS FLX system) au séquençage de génomes entiers (environs 1500 GB pour l'Illumina Hiseq 4000) et d'une utilisation exclusive à la recherche à l'analyse en routine dans un cadre de diagnostics cliniques. Le nombre croissant d'études utilisant le WGS ou le WES démontre le pouvoir de ces approches dans des analyses phénotype-génotype impliquant des pathologies à transmission mendélienne. De plus, la diminution constante des coûts par génomes / exomes séquencés laisse supposer que ces technologies deviendront d'ici peut le fer de lance de la génétique clinique moderne. Cependant, la quantité de données produites crée de nouvelles problématiques pour les généticiens qui se retrouvent désormais face au "déluge de données génétiques" [200]. 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 à chaque étape (**Figure : 2.30**). Bien que de nombreux efforts soient faits pour palier la contrainte instaurée par les *reads* courts dans le cadre d'analyse génomique, les solutions informatiques et bioinformatiques proposées jusqu'à présent restent en dessous des besoins créés par NGS [201]. Cette masse de données produite, à l'origine du succès du séquençage haut-débit dans le domaine de la génomique et de la post-génomique, se trouve désormais être un frein à la compréhension et à l'interprétation des réseaux de gènes et leurs implications dans des pathologies. La limitation de cette technologie n'est donc plus le séquençage d'un, de plusieurs, ou de l'ensemble des gènes, mais plutôt l'analyse et l'interprétation des données générées. Le processus allant de l'extraction de l'ADN à l'identification d'un variant responsable d'une pathologie comprend de nombreuses étapes apportant avec elles leurs lots d'erreurs. Bien que dans chacune de ces phases, de nombreux acteurs soient en concurrence et cherchent à atteindre une solution idéale, celle-ci n'a toujours pas été trouvée et la prolifération des logiciels et algorithmes d'analyses, bien que nécessaire, peut également parfois augmenter la confusion.

Malgré les dizaines de milliers d'exomes et de génomes ayant été jusqu'à présent étudiés, notre compréhension des mécanismes moléculaires qui sous-tendent la variété génomique humaine reste limitée, et ce particulièrement dans le contexte de l'analyse de pathologies génétiques. En effet, à l'heure actuelle, plus de 3700 pathologies à transmission mendélienne ont été caractérisées mais un nombre similaire a toujours une cause inconnue [202]. L'élucidation de ces mystères passera probablement par une harmonisation des méthodes de production des données ainsi que par l'amélioration des techniques d'analyses.

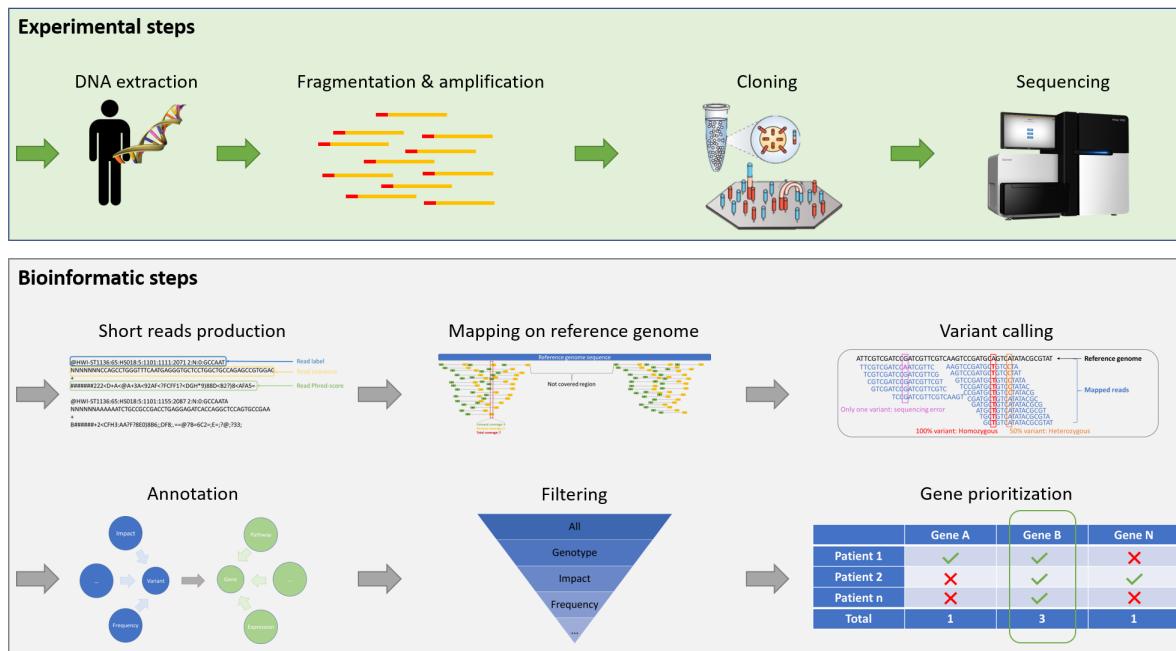


Figure 2.30 – Récapitulatif des différentes étapes du séquençage NGS dans le cadre d'une étude phénotype-génotype : L'ADN est d'abord extrait, puis fragmenté. Les fragments sont liés à des adaptateurs puis amplifiés. Ces amplifiats sont alors isolés et soumis à une amplification clonale (schéma de l'amplification clonale est adapté d'après [142]). Chacun des clones est ensuite séquencé. Les *reads* générés à l'issu du séquençage sont stockés dans des fichiers FASTQ qui serviront de base pour l'étape d'alignement à la suite de laquelle, les variantes et leur génotype seront appelés puis annotés. Ces annotations serviront ensuite pour filtrer les variantes jugées non pertinentes dans le cadre de l'étude, les variantes / gènes restants seront ensuite priorisés de sorte à identifier le/les variant(s) responsable(s) du phénotype.

2.7 Problématique : Un patient, 50.000 variants. “*There can be only one*”. Et après ?

Malgré l’identification régulière de nouveaux gènes depuis plusieurs années, la très grande majorité des cas d’infertilité reste encore inexpliquée et sans cause génétique connue. Ce faible taux de succès peut s’expliquer en partie par une grande hétérogénéité génétique pour de nombreux phénotypes. Par exemple, il est estimé qu’entre 1500 et 2000 gènes sont impliqués dans le contrôle de la spermatogénèse parmi lesquels 300 à 600 sont spécifiquement exprimés dans les cellules germinales masculines, on s’attend logiquement à ce que des anomalies génétiques portant sur ces gènes perturbent la fertilité masculine [203].

Dans ce contexte, l’objectif de ma thèse a été double. D’une part j’ai cherché à montrer que l’utilisation des nouvelles technologies de séquençage ainsi que l’utilisation d’une stratégie d’analyse des données adaptée pouvaient permettre d’identifier de nouvelles causes d’infertilité, améliorant ainsi l’efficacité des diagnostics génétiques pour ces pathologies. D’autre part, ma thèse a également eu pour but d’améliorer nos connaissances sur les mécanismes moléculaires impliqués dans la spermatogénèse.

L’ensemble de ce manuscrit a été rédigé sous l’environnement de développement Rstudio [204] en langage Rmarkdown grâce au package thesisdown <https://github.com/ismayc/thesisdown>. L’ensemble du manuscrit, les codes des différents graphiques qui le composent ainsi que ceux des principaux scripts développés pendant ma thèse sont consultables sur ma page GitHub : <https://github.com/tkaraouzene/>.

CHAPITRE 3

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

3.1 Méthode : Description d'ExSQLibur

3.1.1 L'alignement des *reads*

3.1.2 L'appel des variants

3.1.3 L'annotation

3.1.4 Le filtrage des variants

3.2 Résultats 1 : Analyse de 3 phénotypes par des cas familiaux

3.2.1 Résultats des différentes étapes de l'analyse

Résultat de l'alignement

L'appel des variants

L'annotation des variants

Le filtrage des variants

3.2.2 Article n°1

Contexte et objectifs

Principaux résultats

3.2.3 Article n°2

Contexte et objectifs

Principaux résultats

3.2.4 Article n°3

Contexte et objectifs

Principaux résultats

3.3 Résultats 2 : Étude d'une cohorte de femmes infertiles

3.3.1 Article n°4

Contexte et objectifs

Principaux résultats

3.4 Résultats 3 : Étude d'une large cohorte de patients MMAF

3.4.1 Article n°5

Contexte et objectifs

Principaux résultats

CHAPITRE 4

Investigation génétique et physiologique de la globozoospermie

4.1 Introduction sur la globozoospermie

4.2 Résultats 1 : Les mécanismes mutationnels entraînant la délétion au locus de *DPY19L2* chez l'humain

4.2.1 Article n°6 :

Contexte et objectifs

Principaux résultats

4.3 Résultat 2 : La transcriptomique

4.3.1 Article n°7 :

Contexte et objectifs

Principaux résultats :

CHAPITRE 5

This chunk ensures that the thesisdown package is

CHAPITRE 6

Article annexe 1

References

1. L. Gnessi, A. Fabbri, and G. Spera : “Gonadal peptides as mediators of development and functional control of the testis : An integrated system with hormones and local environment.” *Endocrine Reviews*. vol. 18, no. 4, pp. 541–609, 1997.
2. R.M. Sharpe, C. McKinnell, T. McLaren, M. Millar, T.P. West, S. Maguire, J. Gaughan, V. Syed, B. J?gou, J.B. Kerr, and P.T.K. Saunders : “Interactions Between Androgens, Sertoli Cells and Germ Cells in the Control of Spermatogenesis.” Molecular and cellular endocrinology of the testis. pp. 115–142. *Springer Berlin Heidelberg*, Berlin, Heidelberg (1994).
3. A.L. KIERSZENBAUM : “Mammalian Spermatogenesis *< i>in Vivo</i>* and *< i>in Vitro</i>* : A Partnership of Spermatogenic and Somatic Cell Lineages*.” *Endocrine Reviews*. vol. 15, no. 1, pp. 116–134, 1994.
4. L. JOHNSON, C.S. PETTY, and W.B. NEAVES : “A Comparative Study of Daily Sperm Production and Testicular Composition in Humans and Rats.” *Biol Reprod.* vol. 22, no. 5, pp. 1233–1243, 1980.
5. Y. Clermont : “The cycle of the seminiferous epithelium in man.” *American Journal of Anatomy*. vol. 112, no. 1, pp. 35–51, 1963.
6. Y. Clermont : “Renewal of spermatogonia in man.” *American Journal of Anatomy*. vol. 118, no. 2, pp. 509–524, 1966.
7. E. Goossens and H. Tournaye : “Adult stem cells in the human testis.” *Seminars in Reproductive Medicine*. vol. 31, no. 1, pp. 39–48, 2013.
8. H. Sasaki and Y. Matsui : “Epigenetic events in mammalian germ-cell development : reprogramming and beyond.” *Nat Rev Genet.* vol. 9, no. 2, pp. 129–140, 2008.
9. A.H. Handyside : “Molecular origin of female meiotic aneuploidies.” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. vol. 1822, no. 12, pp. 1913–1920, 2012.
10. J.B. Reece, L.A. Urry, M.L.(L. Cain, S.A. Wasserman, P.V. Minorsky, R.B. Jackson, and N.A. Campbell : “Campbell biology.”, 2014.
11. L.H. Yves Clermont, Richard Oko : “Cell and molecular biology of the testis.” *Oxford University Press*, 1993.
12. D. Escalier, J.M. Gallo, M. Albert, G. Meduri, D. Bermudez, G. David, and J. Schrevel : “Human acrosome biogenesis : immunodetection of proacrosin in primary spermatocytes and of its partitioning pattern during meiosis.” *Development (Cambridge, England)*. vol. 113, no. 3, pp. 779–788, 1991.
13. G.M.H. Hamilton, D. W., Waites : “Cellular and Molecular Events in Spermiogenesis.” *Cambridge University Press*, 1990.
14. Z. Papic, G. Katona, and Z. Skrabalo : “The cytologic identification and quantification of testicular cell subtypes. Reproducibility and relation to histologic findings in the

- diagnosis of male infertility." *Acta cytologica.* vol. 32, no. 5, pp. 697–706, 1988.
15. U. Schenck and W.B. Schill : "Cytology of the human seminiferous epithelium." *Acta cytologica.* vol. 32, no. 5, pp. 689–96,
16. M.M. Adelman and E.M. Cahill : "Atlas of sperm morphology." *ASCP Press*, 1989.
17. World Health Organization : "WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction." *Cambridge University Press*, 1992.
18. A Ogura, J. Matsuda, and R. Yanagimachi : "Birth of normal young after electro-fusion of mouse oocytes with round spermatids." *Proceedings of the National Academy of Sciences of the United States of America.* vol. 91, no. 16, pp. 7460–7462, 1994.
19. A. Ogura, J. Matsuda, T. Asano, O. Suzuki, and R. Yanagimachi : "Mouse oocytes injected with cryopreserved round spermatids can develop into normal offspring." *Journal of Assisted Reproduction and Genetics.* vol. 13, no. 5, pp. 431–434, 1996.
20. I. Sasagawa and R. Yanagimachi : "Spermatids from mice after cryptorchid and reversal operations can initiate normal embryo development." *Journal of andrology.* vol. 18, no. 2, pp. 203–209, 1997.
21. A. Tanaka, M. Nagayoshi, Y. Takemoto, I. Tanaka, H. Kusunoki, S. Watanabe, K. Kuroda, S. Takeda, M. Ito, and R. Yanagimachi : "Fourteen babies born after round spermatid injection into human oocytes." *Proceedings of the National Academy of Sciences.* vol. 112, no. March 2014, pp. 201517466, 2015.
22. B. Asimakopoulos : "Is There a Place for Round and Elongated Spermatids Injection in?" vol. 1, no. 1, pp. 1–6, 2003.
23. R.D. Moreno, J. Palomino, and G. Schatten : "Assembly of spermatid acrosome depends on microtubule organization during mammalian spermiogenesis." *Developmental Biology.* vol. 293, no. 1, pp. 218–227, 2006.
24. L. Hermo, R.M. Pelletier, D.G. Cyr, and C.E. Smith : "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.* vol. 73, no. 4, pp. 320–363, 2010.
25. A. Toure, B. Rode, G.R. Hunnicutt, D. Escalier, and G. Gacon : "Septins at the annulus of mammalian sperm." *Biological Chemistry.* vol. 392, no. 8-9, pp. 799–803, 2011.
26. A.L. Kierszenbaum and L.L. Tres : "The acrosome-acroplaxome-manchette complex and the shaping of the spermatid head." *Archives of histology and cytology.* vol. 67, no. 4, pp. 271–84, 2004.
27. C. Cho, W.D. Willis, E.H. Goulding, H. Jung-Ha, Y.C. Choi, N.B. Hecht, and

- E.M. Eddy : "Haploinsufficiency of protamine-1 or -2 causes infertility in mice." *Nature genetics.* vol. 28, no. 1, pp. 82–6, 2001.
28. A.L. Kierszenbaum and L.L. Tres : "RNA transcription and chromatin structure during meiotic and postmeiotic stages of spermatogenesis." *Federation proceedings.* vol. 37, no. 11, pp. 2512–6, 1978.
29. W.S. Ward : "The structure of the sleeping genome : implications of sperm DNA organization for somatic cells." *Journal of cellular biochemistry.* vol. 55, no. 1, pp. 77–82, 1994.
30. R.E. Braun : "Packaging paternal chromosomes with protamine." *Nature genetics.* vol. 28, no. 1, pp. 10–12, 2001.
31. K. Inaba : "Molecular Architecture of the Sperm Flagella : Molecules for Motility and Signaling." *Zoological Science.* vol. 20, no. 9, pp. 1043–1056, 2003.
32. E.M. Eddy : "The scaffold role of the fibrous sheath." *Society of Reproduction and Fertility supplement.* vol. 65, pp. 45–62, 2007.
33. C.L. Borg, K.M. Wolski, G.M. Gibbs, and M.K. O'Bryan : "Phenotyping male infertility in the mouse : how to get the most out of a 'non-performer'." *Human reproduction update.* vol. 16, no. 2, pp. 205–24, 2010.
34. J. Boivin, L. Bunting, J.A. Collins, and K.G. Nygren : "International estimates of infertility prevalence and treatment-seeking : potential need and demand for infertility medical care." *Human Reproduction.* vol. 22, no. 6, pp. 1506–1512, 2007.
35. J.G.(.G. Grudzinskas and J. Yovich : "Gametes : the spermatozoon." *Cambridge University Press,* 1995.
36. M. Michael and K. Joel : "Zellformen in normalen und pathologischen Ejakulaten und ihre klinische Bedeutung." *Schweiz. Med. Wsch.* 1937.
37. M. Tomlinson, C. Barrati, A. Bolton, E. Lenton, H. Roberts, and I. Cooke : "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.* vol. 42, no. 2, pp. 223–224, 1993.
38. J. MacLeod : "The Significance of Deviations in Human Sperm Morphology." Presented at the (1970).
39. M.J. Tomlinson, C.L.R. Barratt, and I.D. Cooke : "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.* vol. 60, no. 6, pp. 1069–1075, 1993.
40. L.F. Kurilo, I.A. Liubashevskaya, V.P. Dubinskaya, and T.N. Gaeva : "[Karyological analysis of the count of immature germ cells in the ejaculate]." *Urologiiia i nefrologiiia.*

- no. 2, pp. 45–7, 1993.
41. K. SPERLING and R. KADEN : “Meiotic Studies of the Ejaculated Seminal Fluid of Humans with Normal Sperm Count and Oligospermia.” *Nature*. vol. 232, no. 5311, pp. 481–481, 1971.
42. S.M. Girgis, A.N. Etriby, A.A. Ibrahim, and S.A. Kahil : “Testicular biopsy in azoospermia. A review of the last ten years’ experiences of over 800 cases.” *Fertility and sterility*. vol. 20, no. 3, pp. 467–77, 1969.
43. T.G. Cooper, E. Noonan, S. von Eckardstein, J. Auger, H.W.G. Baker, H.M. Behre, T.B. Haugen, T. Kruger, C. Wang, M.T. Mbizvo, and K.M. Vogelsong : “World Health Organization reference values for human semen characteristics.” *Human Reproduction Update*. vol. 16, no. 3, pp. 231–245, 2010.
44. T.J. Colgan, Y.C. Bedard, H.T. Strawbridge, M.B. Buckspan, and P.G. Klotz : “Reappraisal of the Value of Testicular Biopsy in the Investigation of Infertility.” *Fertility and Sterility*. vol. 33, no. 1, pp. 56–60, 1980.
45. H.S. Levin : “Testicular biopsy in the study of male infertility.” *Human Pathology*. vol. 10, no. 5, pp. 569–584, 1979.
46. K.O. Soderström and J. Suominen : “Histopathology and ultrastructure of meiotic arrest in human spermatogenesis.” *Archives of pathology & laboratory medicine*. vol. 104, no. 9, pp. 476–82, 1980.
47. T.-W. WONG, F.H.I. STRAUS, and N.E. WARNER : “TESTICULAR BIOPSY IN THE STUDY OF MALE INFERTILITY : II. POST... : Obstetrical & Gynecological Survey.” *Obstetrical & Gynecological Survey*. vol. 28, no. 9, pp. 660–661, 1973.
48. G. Palermo, H. Joris, P. Devroey, and A.C. Van Steirteghem : “Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte.” *Lancet (London, England)*. vol. 340, no. 8810, pp. 17–8, 1992.
49. J. Auger, F. Eustache, A.G. Andersen, D.S. Irvine, N. Jørgensen, N.E. Skakkebæk, J. Suominen, J. Toppari, M. Vierula, P. Jouannet, N.E. Skakkebaek, J. Suominen, J. Toppari, M. Vierula, and P. Jouannet : “Sperm morphological defects related to environment, lifestyle and medical history of 1001 male partners of pregnant women from four European cities.” *Human reproduction (Oxford, England)*. vol. 16, no. 12, pp. 2710–7, 2001.
50. C. Lindholmer : “The importance of seminal plasma for human sperm motility.” *Biology of reproduction*. vol. 10, no. 5, pp. 533–42, 1974.
51. L. Björndahl : “The usefulness and significance of assessing rapidly progressive spermatozoa.” *Asian journal of andrology*. vol. 12, no. 1, pp. 33–5, 2010.
52. R.J. Aitken, M. Sutton, P. Warner, and D.W. Richardson : “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*.

- vol. 73, no. 2, pp. 441–9, 1985.
53. F. Tüttelmann, M. Simoni, S. Kliesch, S. Ledig, B. Dworniczak, P. Wieacker, and A. Röpke : “Copy number variants in patients with severe oligozoospermia and Sertoli-cell-only syndrome.” *PloS one*. vol. 6, no. 4, pp. e19426, 2011.
54. H. Skaletsky, T. Kuroda-Kawaguchi, P.J. Minx, H.S. Cordum, L. Hillier, L.G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.-F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.-P. Yang, R.H. Waterston, R.K. Wilson, S. Rozen, and D.C. Page : “The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.” *Nature*. vol. 423, no. 6942, pp. 825–837, 2003.
55. J. Hotaling and D.T. Carrell : “Clinical genetic testing for male factor infertility : current applications and future directions.” *Andrology*. vol. 2, no. 3, pp. 339–350, 2014.
56. K.L. O’Flynn O’Brien, A.C. Varghese, and A. Agarwal : “The genetic causes of male factor infertility : A review.” *Fertility and Sterility*. vol. 93, no. 1, pp. 1–12, 2010.
57. C. Ravel, I. Berthaut, J.L. Bresson, J.P. Siffroi, and Genetics Commission of the French Federation of CECOS : “Prevalence of chromosomal abnormalities in phenotypically normal and fertile adult males : large-scale survey of over 10 000 sperm donor karyotypes.” *Human Reproduction*. vol. 21, no. 6, pp. 1484–1489, 2006.
58. A. Bojesen and C.H. Gravholt : “Morbidity and mortality in Klinefelter syndrome (47,XXY).” *Acta Paediatrica*. vol. 100, no. 6, pp. 807–813, 2011.
59. J. Gekas, F. Thepot, C. Turleau, J.P. Siffroi, J.P. Dadoune, S. Briault, M. Rio, G. Bourouillou, F. Carré-Pigeon, R. Wasels, B. Benzacken, and Association des Cytogeneticiens de Langue Francaise : “Chromosomal factors of infertility in candidate couples for ICSI : an equal risk of constitutional aberrations in women and men.” *Human reproduction (Oxford, England)*. vol. 16, no. 1, pp. 82–90, 2001.
60. D.J. Elliott and H.J. Cooke : “The molecular genetics of male infertility.” *BioEssays*. vol. 19, no. 9, pp. 801–809, 1997.
61. C. Krausz and G. Forti : “Clinical aspects of male infertility.” *Results and problems in cell differentiation*. vol. 28, pp. 1–21, 2000.
62. E. Vorona, M. Zitzmann, J. Gromoll, A.N. Schüring, and E. Nieschlag : “Clinical, Endocrinological, and Epigenetic Features of the 46,XX Male Syndrome, Compared with 47,XXY Klinefelter Patients.” *The Journal of Clinical Endocrinology & Metabolism*. vol. 92, no. 9, pp. 3458–3465, 2007.
63. J. Yu, Z. Chen, Y. Ni, and Z. Li : “CFTR mutations in men with congenital bilateral absence of the vas deferens (CBAVD) : a systemic review and meta-analysis.”

- Human Reproduction.* vol. 27, no. 1, pp. 25–35, 2012.
64. Ö. Ayhan, M. Balkan, A. Guven, R. Hazan, M. Atar, A. Tok, and A. Tolun : “Truncating mutations in TAF4B and ZMYND15 causing recessive azoospermia.” *Journal of medical genetics.* vol. 51, no. 4, pp. 239–44, 2014.
65. C.J. Jorgez, N. Wilken, J.B. Addai, J. Newberg, H.V. Vangapandu, A.W. Pastuszak, S. Mukherjee, J.A. Rosenfeld, L.I. Lipshultz, and D.J. Lamb : “Genomic and genetic variation in E2F transcription factor-1 in men with nonobstructive azoospermia.” *Fertility and Sterility.* vol. 103, no. 1, pp. 44–52.e1, 2015.
66. K. Agger, E. Santoni-Rugiu, C. Holmberg, O. Karlström, and K. Helin : “Conditional E2F1 activation in transgenic mice causes testicular atrophy and dysplasia mimicking human CIS.” *Oncogene.* vol. 24, no. 5, pp. 780–789, 2005.
67. A.N. Yatsenko, A.P. Georgiadis, A. Röpke, A.J. Berman, T. Jaffe, M. Olszewska, B. Westernströer, J. Sanfilippo, M. Kurpisz, A. Rajkovic, S.A. Yatsenko, S. Kliesch, S. Schlatt, and F. Tüttelmann : “X-linked TEX11 mutations, meiotic arrest, and azoospermia in infertile men.” *The New England journal of medicine.* vol. 372, no. 22, pp. 2097–107, 2015.
68. F. Yang, S. Silber, N.A. Leu, R.D. Oates, J.D. Marszalek, H. Skaletsky, L.G. Brown, S. Rozen, D.C. Page, and P.J. Wang : “TEX11 is mutated in infertile men with azoospermia and regulates genome-wide recombination rates in mouse.” *EMBO molecular medicine.* vol. 7, no. 9, pp. 1198–210, 2015.
69. E. Maor-Sagie, Y. Cinnamon, B. Yaacov, A. Shaag, H. Goldsmidt, S. Zenvirt, N. Laufer, C. Richler, and A. Frumkin : “Deleterious mutation in SYCE1 is associated with non-obstructive azoospermia.” *Journal of assisted reproduction and genetics.* vol. 32, no. 6, pp. 887–91, 2015.
70. Y. Tenenbaum-Rakover, A. Weinberg-Shukron, P. Renbaum, O. Lobel, H. Eideh, S. Gulsuner, D. Dahary, A. Abu-Rayyan, M. Kanaan, E. Levy-Lahad, D. Bercovich, and D. Zangen : “Minichromosome maintenance complex component 8 (MCM8) gene mutations result in primary gonadal failure.” *Journal of Medical Genetics.* vol. 52, no. 6, pp. 391–399, 2015.
71. F. Yang, S. Eckardt, N.A. Leu, K.J. McLaughlin, and P.J. Wang : “Mouse TEX15 is essential for DNA double-strand break repair and chromosomal synapsis during male meiosis.” *The Journal of Cell Biology.* vol. 180, no. 4, pp. 673–679, 2008.
72. G. Minase, T. Miyamoto, Y. Miyagawa, M. Iijima, H. Ueda, Y. Saijo, M. Namiki, and K. Sengoku : “Single-nucleotide polymorphisms in the human <i>RAD21L</i> gene may be a genetic risk factor for Japanese patients with azoospermia caused by meiotic arrest and Sertoli cell-only syndrome.” *Human Fertility.* pp. 1–4, 2017.
73. M. Gershoni, R. Hauser, L. Yogev, O. Lehavi, F. Azem, H. Yavetz, S. Pietrokovski, and S.E. Kleiman : “A familial study of azoospermic men identifies three novel causative mutations in three new human azoospermia genes.” *Genetics in Medicine.*

vol. 19, no. 9, pp. 998–1006, 2017.

74. M. Arafat, I. Har-Vardi, A. Harlev, E. Levitas, A. Zeadna, M. Abofoul-Azab, V. Dyomin, V.C. Sheffield, E. Lunenfeld, M. Huleihel, and R. Parvari : “Mutation in TDRD9 causes non-obstructive azoospermia in infertile men.” *Journal of Medical Genetics*. vol. 54, no. 9, pp. 633–639, 2017.
75. M. Nistal, R. Paniagua, and A. Herruzo : “Multi-tailed spermatozoa in a case with asthenospermia and teratospermia.” *Virchows Archiv B*. vol. 26, no. 1, pp. 111–118, 1978.
76. K. Dieterich, R. Soto Rifo, A.K. Faure, S. Hennebicq, B. Ben Amar, M. Zahi, J. Perrin, D. Martinez, B. Sèle, P.-S. Jouk, T. Ohlmann, S. Rousseaux, J. Lunardi, and P.F. Ray : “Homozygous mutation of AURKC yields large-headed polyploid spermatozoa and causes male infertility.” *Nature genetics*. vol. 39, no. 5, pp. 661–5, 2007.
77. M. Ben Khelifa, R. Zouari, R. Harbuz, L. Halouani, C. Arnoult, J. Lunardi, and P.F. Ray : “A new AURKC mutation causing macrozoospermia : implications for human spermatogenesis and clinical diagnosis.” *Molecular Human Reproduction*. vol. 17, no. 12, pp. 762–768, 2011.
78. K. Dieterich, R. Zouari, R. Harbuz, F. Vialard, D. Martinez, H. Bellayou, N. Prisant, A. Zoghmar, M.R. Guichaoua, I. Koscienski, M. Kharouf, M. Noruzinia, S. Nadifi, A. Sefiani, J. Lornage, M. Zahi, S. Viville, B. Sele, P.-S. Jouk, M.-C. Jacob, D. Escalier, Y. Nikas, S. Hennebicq, J. Lunardi, and P.F. Ray : “The Aurora Kinase C c.144delC mutation causes meiosis I arrest in men and is frequent in the North African population.” *Human Molecular Genetics*. vol. 18, no. 7, pp. 1301–1309, 2009.
79. A. Dam, I. Feenstra, J. Westphal, L. Ramos, R. van Golde, and J. Kremer : “Globozoospermia revisited.” *Human Reproduction Update*. vol. 13, no. 1, pp. 63–75, 2006.
80. C.G.S. Sen, A.F. Holstein, and C. Schirren : “über die Morphogenese rundköpfiger Spermatozoen des Menschen.” *Andrologia*. vol. 3, no. 3, pp. 117–125, 1971.
81. A.F. Holstein, C. Schirren, and C.G. Schirren : “Human spermatids and spermatozoa lacking acrosomes.” *Journal of reproduction and fertility*. vol. 35, no. 3, pp. 489–91, 1973.
82. A.H. Dam, I. Koscienski, J.A. Kremer, C. Moutou, A.-S. Jaeger, A.R. Oudakker, H. Tournaye, N. Charlet, C. Lagier-Tourenne, H. van Bokhoven, and S. Viville : “Homozygous Mutation in SPATA16 Is Associated with Male Infertility in Human Globozoospermia.” *The American Journal of Human Genetics*. vol. 81, no. 4, pp. 813–820, 2007.
83. L. Lu, M. Lin, M. Xu, Z.-M. Zhou, and J.-H. Sha : “Gene functional research using polyethylenimine-mediated in vivo gene transfection into mouse spermatogenic cells.”

- Asian Journal of Andrology.* vol. 8, no. 1, pp. 53–59, 2006.
84. R. Harbuz, R. Zouari, V. Pierre, M. Ben Khelifa, M. Kharouf, C. Coutton, G. Merdassi, F. Abada, J. Escoffier, Y. Nikas, F. Vialard, I. Koscinski, C. Triki, N. Sermondade, T. Schweitzer, A. Zhioua, F. Zhioua, H. Latrous, L. Halouani, M. Ouafi, M. Makni, P.-S. Jouk, B. Sèle, S. Hennebicq, V. Satre, S. Viville, C. Arnoult, J. Lunardi, and P.F. Ray : “A recurrent deletion of DPY19L2 causes infertility in man by blocking sperm head elongation and acrosome formation.” *American journal of human genetics.* vol. 88, no. 3, pp. 351–61, 2011.
85. H.E. Chemes and V.Y. Rawe : “The making of abnormal spermatozoa : cellular and molecular mechanisms underlying pathological spermiogenesis.” *Cell and Tissue Research.* vol. 341, no. 3, pp. 349–357, 2010.
86. D. Panidis, D. Rousso, A. Kourtis, C. Gianoulis, K. Papathanasiou, and J. Kalachanis : “Headless spermatozoa in semen specimens from fertile and subfertile men.” *The Journal of reproductive medicine.* vol. 46, no. 11, pp. 947–50, 2001.
87. H.E. Chemes, C. Carizza, F. Scarinci, S. Brugo, N. Neuspiller, and L. Schwarsztein : “Lack of a head in human spermatozoa from sterile patients : a syndrome associated with impaired fertilization.” *Fertility and sterility.* vol. 47, no. 2, pp. 310–6, 1987.
88. F. Zhu, F. Wang, X. Yang, J. Zhang, H. Wu, Z. Zhang, Z. Zhang, X. He, P. Zhou, Z. Wei, J. Gecz, and Y. Cao : “Biallelic SUN5 Mutations Cause Autosomal-Recessive Acephalic Spermatozoa Syndrome.” *The American Journal of Human Genetics.* vol. 99, no. 4, pp. 942–949, 2016.
89. S. Yassine, J. Escoffier, R. Abi Nahed, R.A. Nahed, V. Pierre, T. Karaouzene, P.F. Ray, and C. Arnoult : “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.* vol. 10, no. 3, pp. e0118698, 2015.
90. C. Coutton, J. Escoffier, G. Martinez, C. Arnoult, and P.F. Ray : “Teratozoospermia : spotlight on the main genetic actors in the human.” *Human Reproduction Update.* vol. 21, no. 4, pp. 455–485, 2015.
91. M. Ben Khelifa, C. Coutton, R. Zouari, T. Karaouzène, J. Rendu, M. Bidart, S. Yassine, V. Pierre, J. Delaroche, S. Hennebicq, D. Grunwald, D. Escalier, K. Pernet-Gallay, P.S. Jouk, N. Thierry-Mieg, A. Touré, C. Arnoult, and P.F. Ray : “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.* vol. 94, no. 1, pp. 95–104, 2014.
92. X. Wang, H. Jin, F. Han, Y. Cui, J. Chen, C. Yang, P. Zhu, W. Wang, G. Jiao, W. Wang, C. Hao, and Z. Gao : “Homozygous *< i>DNAH1</i>* frameshift mutation causes multiple morphological anomalies of the sperm flagella in Chinese.” *Clinical Genetics.* vol. 91, no. 2, pp. 313–321, 2017.
93. A. Amiri-Yekta, C. Coutton, Z.-E. Kherraf, T. Karaouzène, P. Le Tanno, M.H. Sanati, M. Sabbaghian, N. Almadani, M.A. Sadighi Gilani, S.H. Hosseini, S. Bahrami,

- A. Daneshpour, M. Bini, C. Arnoult, R. Colombo, H. Gourabi, and P.F. Ray : “Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new *< i>DNAH1</i>* mutations.” *Human Reproduction*. vol. 31, no. 12, pp. 2872–2880, 2016.
94. M. Nomikos, J. Kashir, K. Swann, and F.A. Lai : “Sperm PLC ζ : From structure to Ca $^{2+}$ oscillations, egg activation and therapeutic potential.” *FEBS Letters*. vol. 587, no. 22, pp. 3609–3616, 2013.
95. S.N. Amdani, C. Jones, and K. Coward : “Phospholipase C zeta (PLC ζ) : Oocyte activation and clinical links to male factor infertility.” *Advances in Biological Regulation*. vol. 53, no. 3, pp. 292–308, 2013.
96. E. Heytens, J. Parrington, K. Coward, C. Young, S. Lambrecht, S.-Y. Yoon, R.A. Fissore, R. Hamer, C.M. Deane, M. Ruas, P. Grasa, R. Soleimani, C.A. Cuvelier, J. Gerris, M. Dhont, D. Deforce, L. Leybaert, and P. De Sutter : “Reduced amounts and abnormal forms of phospholipase C zeta (PLC ζ) in spermatozoa from infertile men.” *Human reproduction (Oxford, England)*. vol. 24, no. 10, pp. 2417–28, 2009.
97. J. Escoffier, H.C. Lee, S. Yassine, R. Zouari, G. Martinez, T. Karaouzène, C. Coutton, Z.-E. Kherraf, L. Halouani, C. Triki, S. Nef, N. Thierry-Mieg, S.N. Savinov, R. Fissore, P.F. Ray, and C. Arnoult : “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*. vol. 25, no. 5, pp. 878–91, 2016.
98. J.J. Eppig : “Coordination of nuclear and cytoplasmic oocyte maturation in eutherian mammals.” *Reproduction, fertility, and development*. vol. 8, no. 4, pp. 485–9, 1996.
99. J.J. EPPIG, M.M. VIVEIROS, C.M. BIVENS, and R. DE LA FUENTE : “Regulation of Mammalian Oocyte Maturation.” *The ovary*. pp. 113–129. Elsevier (2004).
100. E. Tosti and Y. Ménézo : “Gamete activation : basic knowledge and clinical applications.” *Human Reproduction Update*. vol. 22, no. 4, pp. 420–439, 2016.
101. O. Hertwig : “Beitr{ä}ge zur Kenntniss der Bildung, Befruchtung und Theilung des thierischen Eies.” *W. Engelmann*, 1875.
102. P.M. Wassarman and E.S. Litscher : “Mammalian fertilization : the egg’s multi-functional zona pellucida.” *The International journal of developmental biology*. vol. 52, no. 5-6, pp. 665–76, 2008.
103. S.A. Stricker : “Comparative Biology of Calcium Signaling during Fertilization and Egg Activation in Animals.” *Developmental Biology*. vol. 211, no. 2, pp. 157–176, 1999.
104. S. Miyazaki, H. Shirakawa, K. Nakada, and Y. Honda : “Essential role of the inositol 1,4,5-trisphosphate receptor/Ca $^{2+}$ release channel in Ca $^{2+}$ waves and Ca $^{2+}$ oscillations at fertilization of mammalian eggs.” <http://www.sciencedirect.com/>

- science/article/pii/S0012160683711681, (1993).
105. K. Swann : “A cytosolic sperm factor stimulates repetitive calcium increases and mimics fertilization in hamster eggs.” *Development*. vol. 110, no. 4, 1990.
106. F.Z. Sun, J. Hoyland, X. Huang, W. Mason, R.M. Moor, and P. Rossi : “A comparison of intracellular changes in porcine eggs after fertilization and electroactivation.” *Development (Cambridge, England)*. vol. 115, no. 4, pp. 947–56, 1992.
107. Y. Lawrence, M. Whitaker, and K. Swann : “Sperm-egg fusion is the prelude to the initial Ca²⁺ increase at fertilization in the mouse.” *Development (Cambridge, England)*. vol. 124, no. 1, pp. 233–41, 1997.
108. H. Wu, C.L. He, and R.A. Fissore : “Injection of a porcine sperm factor triggers calcium oscillations in mouse oocytes and bovine eggs.” *Molecular Reproduction and Development*. vol. 46, no. 2, pp. 176–189, 1997.
109. H. Wu, C.-L. He, B. Jehn, S.J. Black, and R.A. Fissore : “Partial Characterization of the Calcium-Releasing Activity of Porcine Sperm Cytosolic Extracts.” *Developmental Biology*. vol. 203, no. 2, pp. 369–381, 1998.
110. S.A. Stricker : “Intracellular Injections of a Soluble Sperm Factor Trigger Calcium Oscillations and Meiotic Maturation in Unfertilized Oocytes of a Marine Worm.” *Developmental Biology*. vol. 186, no. 2, pp. 185–201, 1997.
111. T.S. Tang, J.B. Dong, X.Y. Huang, and F.Z. Sun : “Ca(2+) oscillations induced by a cytosolic sperm protein factor are mediated by a maternal machinery that functions only once in mammalian eggs.” *Development (Cambridge, England)*. vol. 127, no. 5, pp. 1141–50, 2000.
112. C.M. Saunders, M.G. Larman, J. Parrington, L.J. Cox, J. Royse, L.M. Blayney, K. Swann, and F.A. Lai : “PLC zeta : a sperm-specific trigger of Ca(2+) oscillations in eggs and embryo development.” *Development (Cambridge, England)*. vol. 129, no. 15, pp. 3533–44, 2002.
113. G. Hartshorne, S. Montgomery, and L. Klentzeris : “A case of failed oocyte maturation in vivo and in vitro.” *Fertility and sterility*. vol. 71, no. 3, pp. 567–70, 1999.
114. D. Levran, J. Farhi, H. Nahum, M. Glezerman, and A. Weissman : “Maturation arrest of human oocytes as a cause of infertility : case report.” *Human reproduction (Oxford, England)*. vol. 17, no. 6, pp. 1604–9, 2002.
115. S. Beall, C. Brenner, and J. Segars : “Oocyte maturation failure : a syndrome of bad eggs.” *Fertility and sterility*. vol. 94, no. 7, pp. 2507–13, 2010.
116. A. Hourvitz, E. Maman, M. Brengauz, R. Machtlinger, and J. Dor : “In vitro maturation for patients with repeated in vitro fertilization failure due to ‘oocyte maturation abnormalities’.” *Fertility and Sterility*. vol. 94, no. 2, pp. 496–501, 2010.
117. R. Feng, Q. Sang, Y. Kuang, X. Sun, Z. Yan, S. Zhang, J. Shi, G. Tian, A.

- Luchniak, Y. Fukuda, B. Li, M. Yu, J. Chen, Y. Xu, L. Guo, R. Qu, X. Wang, Z. Sun, M. Liu, H. Shi, H. Wang, Y. Feng, R. Shao, R. Chai, Q. Li, Q. Xing, R. Zhang, E. Nogales, L. Jin, L. He, M.L. Gupta, N.J. Cowan, and L. Wang : "Mutations in *< i>TUBB8*" and Human Oocyte Meiotic Arrest." *New England Journal of Medicine*. vol. 374, no. 3, pp. 223–232, 2016.
118. D. Clift and M. Schuh : "Restarting life : fertilization and the transition from meiosis to mitosis." *Nature Reviews Molecular Cell Biology*. vol. 14, no. 9, pp. 549–562, 2013.
119. P. de Boer, M. de Vries, and L. Ramos : "A mutation study of sperm head shape and motility in the mouse : lessons for the clinic." *Andrology*. vol. 3, no. 2, pp. 174–202, 2015.
120. E. ElInati, P. Kuentz, C. Redin, S. Jaber, F. Vanden Meerschaut, J. Makarian, I. Koscinski, M.H. Nasr-Esfahani, A. Demirol, T. Gurgan, N. Louanjli, N. Iqbal, M. Bisharah, F.C. Pigeon, H. Gourabi, D. De Briel, F. Brugnon, S.A. Gitlin, J.-M. Grillo, K. Ghaedi, M.R. Deemeh, S. Tanhaei, P. Modarres, B. Heindryckx, M. Benkhalifa, D. Nikiforaki, S.C. Oehninger, P. De Sutter, J. Muller, and S. Viville : "Globozoospermia is mainly due to DPY19L2 deletion via non-allelic homologous recombination involving two recombination hotspots." *Human Molecular Genetics*. vol. 21, no. 16, pp. 3695–3702, 2012.
121. Y. Choi, S. Jeon, M. Choi, M.-h. Lee, M. Park, D.R. Lee, K.-Y. Jun, Y. Kwon, O.-H. Lee, S.-H. Song, J.-Y. Kim, K.-A. Lee, T.K. Yoon, A. Rajkovic, and S.H. Shim : "Mutations in SOHLH1 gene associate with nonobstructive Azoospermia." *Human Mutation*. vol. 31, no. 7, pp. 788–793, 2010.
122. A. Bashamboo, B. Ferraz-de-Souza, D. Lourenço, L. Lin, N.J. Sebire, D. Montjean, J. Bignon-Topalovic, J. Mandelbaum, J.-P. Siffroi, S. Christin-Maitre, U. Radhakrishna, H. Rouba, C. Ravel, J. Seeler, J.C. Achermann, and K. McElreavey : "Human male infertility associated with mutations in NR5A1 encoding steroidogenic factor 1." *American journal of human genetics*. vol. 87, no. 4, pp. 505–12, 2010.
123. T. Miyamoto, S. Hasuike, L. Yogev, M.R. Maduro, M. Ishikawa, H. Westphal, and D.J. Lamb : "Azoospermia in patients heterozygous for a mutation in SYCP3." *The Lancet*. vol. 362, no. 9397, pp. 1714–1719, 2003.
124. M. Lek, K.J. Karczewski, E.V. Minikel, K.E. Samocha, E. Banks, T. Fennell, A.H. O'Donnell-Luria, J.S. Ware, A.J. Hill, B.B. Cummings, T. Tukiainen, D.P. Birnbaum, J.A. Kosmicki, L.E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D.N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M.I. Kurki, A.L. Moonshine, P. Natarajan, L. Orozco, G.M. Peloso, R. Poplin, M.A. Rivas, V. Ruano-Rubio, S.A. Rose, D.M. Ruderfer, K. Shakir, P.D. Stenson, C. Stevens, B.P. Thomas, G. Tiao, M.T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D.M. Altshuler, D. Ardiissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J.C. Florez, S.B. Gabriel, G. Getz, S.J. Glatt, C.M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M.I. McCarthy, D.

- McGovern, R. McPherson, B.M. Neale, A. Palotie, S.M. Purcell, D. Saleheen, J.M. Scharf, P. Sklar, P.F. Sullivan, J. Tuomilehto, M.T. Tsuang, H.C. Watkins, J.G. Wilson, M.J. Daly, D.G. MacArthur, and D.G. Exome Aggregation Consortium : “Analysis of protein-coding genetic variation in 60,706 humans.” *Nature*. vol. 536, no. 7616, pp. 285–91, 2016.
125. U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine : “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.” *Proceedings of the National Academy of Sciences of the United States of America*. vol. 96, no. 12, pp. 6745–50, 1999.
126. T. Wang, D. Hopkins, C. Schmidt, S. Silva, R. Houghton, H. Takita, E. Repasky, and S.G. Reed : “Identification of genes differentially over-expressed in lung squamous cell carcinoma using combination of cDNA subtraction and microarray analysis.” *Oncogene*. vol. 19, no. 12, pp. 1519–1528, 2000.
127. D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D’Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, and W.R. Sellers : “Gene expression correlates of clinical prostate cancer behavior.” *Cancer cell*. vol. 1, no. 2, pp. 203–9, 2002.
128. L.J. van ’t Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend : “Gene expression profiling predicts clinical outcome of breast cancer.” *Nature*. vol. 415, no. 6871, pp. 530–536, 2002.
129. A. Brachat, B. Pierrat, A. Xynos, K. Brecht, M. Simonen, A. Brüngger, and J. Heim : “A microarray-based, integrated approach to identify novel regulators of cancer drug response and apoptosis.” *Oncogene*. vol. 21, no. 54, pp. 8361–8371, 2002.
130. D.J. Cutler, M.E. Zwick, M.M. Carrasquillo, C.T. Yohn, K.P. Tobin, C. Kashuk, D.J. Mathews, N.A. Shah, E.E. Eichler, J.A. Warrington, and A. Chakravarti : “High-throughput variation detection and genotyping using microarrays.” *Genome research*. vol. 11, no. 11, pp. 1913–25, 2001.
131. V. Trevino, F. Falciani, and H.A. Barrera-Saldaña : “DNA microarrays : a powerful genomic tool for biomedical and clinical research.” *Molecular medicine (Cambridge, Mass.)*. vol. 13, no. 9-10, pp. 527–41, 2007.
132. D.G. Wang, J.B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M.S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T.J. Hudson, R. Lipshutz, M. Chee, and E.S. Lander : “Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.” *Science (New York, N.Y.)*. vol. 280, no. 5366, pp. 1077–82,

1998.

133. R. Bumgarner : “Overview of DNA microarrays : types, applications, and their future.” *Current protocols in molecular biology*. vol. Chapter 22, pp. Unit 22.1., 2013.
134. P.O. Brown, J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamen- schikov, C.F. Williams, S.S. Jeffrey, and D. Botstein : “Genome-wide analysis of DNA copy-number changes using cDNA microarrays.” *Nature Genetics*. vol. 23, no. 1, pp. 41–46, 1999.
135. F.S. Collins, M. Morgan, and A. Patrinos : “The Human Genome Project : Lessons from Large-Scale Biology.” *Science*. vol. 300, no. 5617, pp. 286–290, 2003.
136. M.L. Metzker : “Sequencing technologies - the next generation.” *Nature reviews. Genetics*. vol. 11, no. 1, pp. 31–46, 2010.
137. D. Sims, I. Sudbery, N.E. Ilott, A. Heger, and C.P. Ponting : “Sequencing depth and coverage : key considerations in genomic analyses.” *Nature reviews. Genetics*. vol. 15, no. 2, pp. 121–32, 2014.
138. B.P. Hodkinson and E.A. Grice : “Next-Generation Sequencing : A Review of Technologies and Tools for Wound Microbiome Research.” *Advances in wound care*. vol. 4, no. 1, pp. 50–58, 2015.
139. S.B. Ng, E.H. Turner, P.D. Robertson, S.D. Flygare, W. Abigail, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. Evan, M. Bamshad, D. a Nickerson, and J. Shendure : “Targeted Capture and Massicely Parallel Sequencing of twelve human exomes.” *Nature*. vol. 461, no. 7261, pp. 272–276, 2010.
140. S.H. Lelieveld, M. Spielmann, S. Mundlos, J. a Veltman, and C. Gilissen : “Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions.” *Human mutation*. vol. 36, no. 8, pp. 815–22, 2015.
141. J. Meienberg, R. Bruggmann, K. Oexle, and G. Matyas : “Clinical sequencing : is WGS the better WES ?” *Human Genetics*. vol. 135, no. 3, pp. 359–362, 2016.
142. S. Goodwin, J.D. McPherson, and W.R. McCombie : “Coming of age : ten years of next-generation sequencing technologies.” *Nat Rev Genet*. vol. 17, no. 6, pp. 333–351, 2016.
143. J. Guo, N. Xu, Z. Li, S. Zhang, J. Wu, D.H. Kim, M. Sano Marma, Q. Meng, H. Cao, X. Li, S. Shi, L. Yu, S. Kalachikov, J.J. Russo, N.J. Turro, and J. Ju : “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*. vol. 105, no. 27, pp. 9145–9150, 2008.
144. A.E. Tomkinson, S. Vijayakumar, J.M. Pascal, and T. Ellenberger : “DNA Ligases : Structure, Reaction Mechanism, and Function.” *Chemical Reviews*. vol. 106, no. 2, pp. 687–699, 2006.
145. B. Wold and R.M. Myers : “Sequence census methods for functional genomics.”

Nature Methods. vol. 5, no. 1, pp. 19–21, 2007.

146. M.Q. Yang, B.D. Athey, H.R. Arabnia, A.H. Sung, Q. Liu, J.Y. Yang, J. Mao, and Y. Deng : “High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics.” *BMC genomics.* vol. 10 Suppl 1, pp. I1, 2009.
147. J. Qin, R. Li, J. Raes, M. Arumugam, S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, T. Yamada, D.R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-m. Batto, T. Hansen, D.L. Paslier, A. Linneberg, H.B. Nielsen, E. Pelletier, P. Renault, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, and H. Yang : “A human gut microbial gene catalog established by metagenomic sequencing.” *Nature.* vol. 464, no. 7285, pp. 59–65, 2010.
148. C.P. Van Tassell, T.P.L. Smith, L.K. Matukumalli, J.F. Taylor, R.D. Schnabel, C.T. Lawley, C.D. Haudenschild, S.S. Moore, W.C. Warren, and T.S. Sonstegard : “SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries.” *Nature Methods.* vol. 5, no. 3, pp. 247–252, 2008.
149. C. Alkan, J.M. Kidd, T. Marques-bonet, G. Aksay, F. Hormozdiari, J.O. Kitzman, C. Baker, M. Malig, S.C. Sahinalp, R.A. Gibbs, and E.E. Eichler : “Personalized Copy-Number and Segmental Duplication Maps using Next-Generation Sequencing.” *Nature Genetics.* vol. 41, no. 10, pp. 1061–1067, 2010.
150. P. Medvedev, M. Stanciu, and M. Brudno : “Computational methods for discovering structural variation with next-generation sequencing.” *Nature Methods.* vol. 6, no. 11s, pp. S13–S20, 2009.
151. K.H. Taylor, R.S. Kramer, J.W. Davis, J. Guo, D.J. Duff, D. Xu, C.W. Caldwell, and H. Shi : “Ultradeep Bisulfite Sequencing Analysis of DNA Methylation Patterns in Multiple Gene Promoters by 454 Sequencing.” *Cancer Research.* vol. 67, no. 18, pp. 8511–8518, 2007.
152. M. Sultan, M.H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O’Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo : “A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome.” *Science.* vol. 321, no. 5891, pp. 956–960, 2008.
153. A. Guffanti, M. Iacono, P. Pelucchi, N. Kim, G. Soldà, L.J. Croft, R.J. Taft, E. Rizzi, M. Askarian-Amiri, R.J. Bonnal, M. Callari, F. Mignone, G. Pesole, G. Bertalot, L. Bernardi, A. Albertini, C. Lee, J.S. Mattick, I. Zucchi, and G. De Bellis : “A transcriptional sketch of a primary human breast cancer by 454 deep sequencing.” *BMC Genomics.* vol. 10, no. 1, pp. 163, 2009.
154. C. Auffray, Z. Chen, and L. Hood : “Systems medicine : the future of medical genomics and healthcare.” *Genome medicine.* vol. 1, no. 1, pp. 2, 2009.
155. D.S. Horner, G. Pavesi, T. Castrignano’, P.D.O. de Meo, S. Liuni, M. Sammeth, E. Picardi, and G. Pesole : “Bioinformatics approaches for genomics and post genomics

- applications of next-generation sequencing.” *Briefings in Bioinformatics.* vol. 11, no. 2, pp. 181–197, 2009.
156. E.R. Mardis : “The impact of next-generation sequencing technology on genetics.” *Trends in Genetics.* vol. 24, no. 3, pp. 133–141, 2008.
157. D.R. Bentley : “Whole-genome re-sequencing.” *Current Opinion in Genetics and Development.* vol. 16, no. 6, pp. 545–552, 2006.
158. H. Li, J. Ruan, R. Durbin, H. Li, J. Ruan, and R. Durbin : “Mapping short DNA sequencing reads and calling variants using mapping quality scores Mapping short DNA sequencing reads and calling variants using mapping quality scores.” pp. 1851–1858, 2008.
159. J.O. Korbel, A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, J. Nicholas, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, a C. Eugenia, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, and M. Snyder : “Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome.” *October.* vol. 318, no. 5849, pp. 420–426, 2009.
160. P.J.A. Cock, C.J. Fields, N. Goto, M.L. Heuer, and P.M. Rice : “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.” *Nucleic Acids Research.* vol. 38, no. 6, pp. 1767–1771, 2009.
161. P. Flicek and E. Birney : “Sense from sequence reads : methods for alignment and assembly.” *Nature methods.* vol. 6, no. 11 Suppl, pp. S6–S12, 2009.
162. R. Nielsen, J.S. Paul, A. Albrechtsen, and Y.S. Song : “Genotype and SNP calling from next-generation sequencing data.” *Nature reviews. Genetics.* vol. 12, no. 6, pp. 443–51, 2011.
163. B. Langmead and S.L. Salzberg : “Fast gapped-read alignment with Bowtie 2.” *Nature Methods.* vol. 9, no. 4, pp. 357–359, 2012.
164. T.J. Treangen and S.L. Salzberg : “Repetitive DNA and next-generation sequencing : computational challenges and solutions.” *Nat Rev Genet.* vol. 13, no. 1, pp. 36–46, 2013.
165. B. Langmead, C. Trapnell, M. Pop, and S. Salzberg : “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.” *Genome biology.* vol. 10, no. 3, pp. R25, 2009.
166. H. Li and R. Durbin : “Fast and accurate short read alignment with Burrows-Wheeler transform.” *Bioinformatics.* vol. 25, no. 14, pp. 1754–1760, 2009.
167. Z. Su, P.P. Łabaj, S.S. Li, J. Thierry-Mieg, D. Thierry-Mieg, W. Shi, C. Wang, G.P. Schroth, R. a Setterquist, J.F. Thompson, W.D. Jones, W. Xiao, W. Xu, R.V. Jensen, R. Kelly, J. Xu, A. Conesa, C. Furlanello, H.H. Gao, H. Hong, N. Jafari, S. Letovsky, Y. Liao, F. Lu, E.J. Oakeley, Z. Peng, C.A. Praul, J. Santoyo-Lopez, A. Scherer, T. Shi, G.K. Smyth, F. Staedtler, P. Sykacek, X.-X. Tan, E.A. Thompson, J.

- Vandesompele, M.D. Wang, J.J.J. Wang, R.D. Wolfinger, J. Zavadil, S.S. Auerbach, W. Bao, H. Binder, T. Blomquist, M.H. Brilliant, P.R. Bushel, W. Cai, J.G. Catalano, C.-W. Chang, T. Chen, G. Chen, R. Chen, M. Chierici, T.-M. Chu, D.-A. Clevert, Y. Deng, A. Derti, V. Devanarayanan, Z. Dong, J. Dopazo, T. Du, H. Fang, Y. Fang, M. Fasold, A. Fernandez, M. Fischer, P. Furió-Tari, J.C. Fuscoe, F. Caimet, S. Gaj, J. Gandara, H.H. Gao, W. Ge, Y. Gondo, B. Gong, M. Gong, Z. Gong, B. Green, C. Guo, L.-W.L. Guo, L.-W.L. Guo, J. Hadfield, J. Hellmanns, S. Hochreiter, M. Jia, M. Jian, C.D. Johnson, S. Kay, J. Kleinjans, S. Lababidi, S. Levy, Q.-Z. Li, L. Li, P. Li, Y. Li, H. Li, J. Li, S.S. Li, S.M. Lin, F.J. López, X. Lu, H. Luo, X. Ma, J. Meehan, D.B. Megherbi, N. Mei, B. Mu, B. Ning, A. Pandey, J. Pérez-Florido, R.G. Perkins, R. Peters, J.H. Phan, M. Pirooznia, F. Qian, T. Qing, L. Rainbow, P. Rocca-Serra, L. Sambourg, S.-A. Sansone, S. Schwartz, R. Shah, J. Shen, T.M. Smith, O. Stegle, N. Stralis-Pavese, E. Stupka, Y. Suzuki, L.T. Szkołnicki, M. Tinning, B. Tu, J. van Delft, A. Vela-Boza, E. Venturini, S.J. Walker, L. Wan, W. Wang, J.J.J. Wang, J.J.J. Wang, E.D. Wieben, J.C. Willey, P.-Y. Wu, J. Xuan, Y. Yang, Z. Ye, Y. Yin, Y. Yu, Y.-C. Yuan, J. Zhang, K.K. Zhang, W.W. Zhang, W.W. Zhang, Y. Zhang, C. Zhao, Y. Zheng, Y. Zhou, P. Zumbo, W. Tong, D.P. Kreil, C.E. Mason, and L. Shi : “A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium.” *Nature Biotechnology*. vol. 32, no. 9, pp. 903–14, 2014.
168. M. Ruffalo, T. Laframboise, and M. Koyutürk : “Comparative analysis of algorithms for next-generation sequencing read alignment.” *Bioinformatics*. vol. 27, no. 20, pp. 2790–2796, 2011.
169. S. Thankaswamy-Kosalai, P. Sen, and I. Nookaew : “Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics.” *Genomics*. 2017.
170. S. Bao, R. Jiang, W. Kwan, B. Wang, X. Ma, and Y.-Q. Song : “Evaluation of next-generation sequencing software in mapping and assembly.” *Journal of Human Genetics*. vol. 56, no. May, pp. 406–414, 2011.
171. M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. del Angel, M.A. Rivas, M. Hanna, A. McKenna, T.J. Fennell, A.M. Kernytsky, A.Y. Sivachenko, K. Cibulskis, S.B. Gabriel, D. Altshuler, M.J. Daly, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, and E. Pritchard : “A framework for variation discovery and genotyping using next-generation DNA sequencing data.” *Nature Genetics*. vol. 43, no. 5, pp. 491–498, 2011.
172. G. Lunter and M. Goodson : “Stampy : A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.” *Genome Research*. vol. 21, no. 6, pp. 936–939, 2011.
173. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin : “The Sequence Alignment/Map format and SAMtools.”

Bioinformatics. vol. 25, no. 16, pp. 2078–2079, 2009.

174. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M.A. DePristo : “The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data.” *Genome research.* vol. 20, no. 9, pp. 1297–303, 2010.
175. E. Garrison and G. Marth : “Haplotype-based variant detection from short-read sequencing.” *arXiv.* 2012.
176. R. Li, Y. Li, X. Fang, H. Yang, J. Wang, K. Kristiansen, and J. Wang : “SNP detection for massively parallel whole-genome resequencing.” *Genome Research.* vol. 19, no. 6, pp. 1124–1132, 2009.
177. D.J. Hedges, D. Hedges, D. Burges, E. Powell, C. Almonte, J. Huang, S. Young, B. Boese, M. Schmidt, M.A. Pericak-Vance, E. Martin, X. Zhang, T.T. Harkins, and S. Züchner : “Exome sequencing of a multigenerational human pedigree.” *PloS one.* vol. 4, no. 12, pp. e8232, 2009.
178. S. Hwang, E. Kim, I. Lee, and E.M. Marcotte : “Systematic comparison of variant calling pipelines using gold standard personal exome variants.” *Scientific Reports.* vol. 5, no. December, pp. 17875, 2015.
179. C.F. Baes, M.A. Dolezal, J.E. Koltes, B. Bapst, E. Fritz-Waters, S. Jansen, C. Flury, H. Signer-Hasler, C. Stricker, R. Fernando, R. Fries, J. Moll, D.J. Garrick, J.M. Reecy, and B. Greddler : “Evaluation of variant identification methods for whole genome sequencing data in dairy cattle.” *BMC genomics.* vol. 15, no. 1, pp. 948, 2014.
180. J. O’Rawe, T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W.E. Johnson, Z. Wei, K. Wang, and G.J. Lyon : “Low concordance of multiple variant-calling pipelines : practical implications for exome and genome sequencing.” *Genome Medicine.* vol. 5, no. 3, pp. 28, 2013.
181. J.A. Rosenfeld, C.E. Mason, T.M. Smith, C. Wallin, and M. Diekhans : “Limitations of the Human Reference Genome for Personalized Genomics.” *PLoS ONE.* vol. 7, no. 7, pp. e40294, 2012.
182. C. Gonzaga-Jauregui, J.R. Lupski, and R.A. Gibbs : “Human genome sequencing in health and disease.” *Annual review of medicine.* vol. 63, pp. 35–61, 2012.
183. T.1.G.P. 1000 Genomes Project Consortium, A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, and G.R. Abecasis : “A global reference for human genetic variation.” *Nature.* vol. 526, no. 7571, pp. 68–74, 2015.
184. W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham : “The Ensembl Variant Effect Predictor.” *Genome biology.* vol. 17, no. 1, pp. 122, 2016.
185. P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, and D.M. Ruden : “A program for annotating and predicting the effects of single

- nucleotide polymorphisms, SnpEff.” *Fly.* vol. 6, no. 2, pp. 80–92, 2012.
186. K. Wang, M. Li, and H. Hakonarson : “ANNOVAR : functional annotation of genetic variants from high-throughput sequencing data.” *Nucleic Acids Research.* vol. 38, no. 16, pp. e164–e164, 2010.
187. P. Kumar, S. Henikoff, and P.C. Ng : “Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.” *Nature protocols.* vol. 4, no. 7, pp. 1073–1081, 2009.
188. Y. Choi, G.E. Sims, S. Murphy, J.R. Miller, and A.P. Chan : “Predicting the Functional Effect of Amino Acid Substitutions and Indels.” *PLoS ONE.* vol. 7, no. 10, 2012.
189. D. Salgado, M.I. Bellgard, J.P. Desvignes, and C. B ?rouard : “How to Identify Pathogenic Mutations among All Those Variations : Variant Annotation and Filtration in the Genome Sequencing Era.” *Human Mutation.* vol. 37, no. 12, pp. 1272–1282, 2016.
190. M. Kircher, D.M. Witten, P. Jain, B.J. O’Roak, G.M. Cooper, and J. Shendure : “A general framework for estimating the relative pathogenicity of human genetic variants.” *Nature Genetics.* vol. 46, no. 3, pp. 310–315, 2014.
191. I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, and S.R. Sunyaev : “A method and server for predicting damaging missense mutations.” *Nature methods.* vol. 7, no. 4, pp. 248–9, 2010.
192. J.M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow : “MutationTaster evaluates disease-causing potential of sequence alterations.” *Nature Methods.* vol. 7, no. 8, pp. 575–576, 2010.
193. D. Salgado, J.-P. Desvignes, G. Rai, A. Blanchard, M. Miltgen, A. Pinard, N. Lévy, G. Collod-Béroud, and C. Béroud : “UMD-Predictor : A High-Throughput Sequencing Compliant System for Pathogenicity Prediction of any Human cDNA Substitution.” *Human Mutation.* vol. 37, no. 5, pp. 439–446, 2016.
194. H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P.D. Thomas : “PANTHER version 11 : expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements.” *Nucleic Acids Research.* vol. 45, no. D1, pp. D183–D189, 2017.
195. S. Köhler, S.C. Doelken, C.J. Mungall, S. Bauer, H.V. Firth, I. Bailleul-Forestier, G.C.M. Black, D.L. Brown, M. Brudno, J. Campbell, D.R. FitzPatrick, J.T. Eppig, A.P. Jackson, K. Freson, M. Girdea, I. Helbig, J.A. Hurst, J. Jähn, L.G. Jackson, A.M. Kelly, D.H. Ledbetter, S. Mansour, C.L. Martin, C. Moss, A. Mumford, W.H. Ouwehand, S.-M. Park, E.R. Riggs, R.H. Scott, S. Sisodiya, S. Van Vooren, R.J. Wapner, A.O.M. Wilkie, C.F. Wright, A.T. Vulto-van Silfhout, N. de Leeuw, B.B.A. de Vries, N.L. Washington, C.L. Smith, M. Westerfield, P. Schofield, B.J. Ruef, G.V. Gkoutos, M. Haendel, D. Smedley, S.E. Lewis, and P.N. Robinson : “The Human

Phenotype Ontology project : linking molecular biology and disease through phenotype data.” *Nucleic acids research*. vol. 42, no. Database issue, pp. D966–74, 2014.

196. S. Petrovski, Q. Wang, E.L. Heinzen, A.S. Allen, D.B. Goldstein, E. Davydov, D. Goode, M. Sirota, G. Cooper, A. Sidow, I. Adzhubei, S. Schmidt, L. Peshkin, V. Ramensky, A. Gerasimova, W. Lee, P. Yue, Z. Zhang, N. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, S. Hicks, D. Wheeler, S. Plon, M. Kimmel, G. Cooper, J. Shendure, B. Neale, Y. Kou, L. Liu, A. Ma’ayan, K. Samocha, B. O’Roak, L. Vives, S. Girirajan, E. Karakoc, N. Krumm, S. Sanders, M. Murtha, A. Gupta, J. Murdoch, M. Raubeson, J. de Ligt, M. Willemsen, B. van Bon, T. Kleefstra, H. Yntema, A. Rauch, D. Wieczorek, E. Graf, T. Wieland, S. Ende, I. Iossifov, M. Ronemus, D. Levy, Z. Wang, I. Hakker, J. Tennessen, A. Bigham, T. O’Connor, W. Fu, E. Kenny, K. Pruitt, J. Harrow, R. Harte, C. Wallin, M. Diekhans, A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, E. Heinzen, K. Swoboda, Y. Hitomi, F. Gurrieri, S. Nicole, J. Eppig, J. Blake, C. Bult, J. Kadin, J. Richardson, B. Georgi, B. Voight, M. Bucan, N. Goldman, Z. Yang, W. Li, C. Wu, C. Luo, M. Nei, T. Gojobori, C. Zhang, J. Wang, M. Long, C. Fan, K. Goh, M. Cusick, D. Valle, B. Childs, M. Vidal, E. DeLong, D. DeLong, D. Clarke-Pearson, X. Robin, N. Turck, A. Hainard, N. Tiberti, and F. Lisacek : “Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes.” *PLoS Genetics*. vol. 9, no. 8, pp. e1003709, 2013.
197. D.J. McCarthy, P. Humburg, A. Kanapin, M. a Rivas, K. Gaulton, J.-B. Cazier, and P. Donnelly : “Choice of transcripts and software has a large effect on variant annotation.” *Genome medicine*. vol. 6, no. 3, pp. 26, 2014.
198. S. Zhao and B. Zhang : “A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification.” *BMC genomics*. vol. 16, no. 1, pp. 97, 2015.
199. K.D. Pruitt, J. Harrow, R.A. Harte, C. Wallin, M. Diekhans, D.R. Maglott, S. Searle, C.M. Farrell, J.E. Loveland, B.J. Ruef, E. Hart, M.M. Suner, M.J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J.L. Cherry, V. Curwen, M. DiCuccio, M. Kellis, J. Lee, M.F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B.L. Maidak, J. Mudge, M.R. Murphy, T. Murphy, J. Rajan, B. Rajput, L.D. Riddick, C. Snow, C. Steward, D. Webb, J.A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman : “The consensus coding sequence (CCDS) project : Identifying a common protein-coding gene set for the human and mouse genomes.” *Genome Research*. vol. 19, no. 7, pp. 1316–1323, 2009.
200. M.C. Schatz and B. Langmead : “The DNA Data Deluge : Fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze.” *IEEE spectrum*. vol. 50, no. 7, pp. 26–33, 2013.
201. J.D. McPherson : “Next-generation gap.” *Nature Methods*. vol. 6, no. 11s, pp. S2–S5, 2009.
202. J. Amberger, C. Bocchini, and A. Hamosh : “A new face and new challenges for

- Online Mendelian Inheritance in Man (OMIM)." *Human Mutation.* vol. 32, no. 5, pp. 564–567, 2011.
203. M.M. Matzuk and D.J. Lamb : "The biology of infertility : research advances and clinical challenges." *Nature Medicine.* vol. 14, no. 11, pp. 1197–1213, 2008.
204. RStudio Team : "RStudio : Integrated Development Environment for R." *RStudio, Inc.*, Boston, MA, 2015.