

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

Remerciements	2
Résumé	6
Abstract	8
Chapitre 1 : Introduction	9
1.1 L'ovogenèse et l'ovocyte	10
1.2 La spermatogenèse	11
1.2.1 Rappels sur le testicule	13
1.2.2 La phase de multiplication	14
1.2.3 La méiose	16
1.2.4 La spermio-génèse	20
1.3 Structure et fonction du spermatozoïde	22
1.3.1 La tête	23
1.3.2 Le flagelle	25
1.4 L'infertilité masculine	27
1.4.1 Les différents phénotypes d'infertilité masculine	27
Anomalies liées à la quantité spermatique	28
Anomalies liées à la morphologie	28
Anomalies liées à la mobilité	30
1.4.2 La génétique de l'infertilité	30
Les causes fréquentes	30
Les nouveaux gènes	33
1.5 Les techniques d'analyses génétiques	35
1.5.1 Approche “gènes candidats”	35
1.5.2 Les puces	36
Les puces à expression	37
Les puces à SNP, plateforme génotypage	38
Les puces à indels	39
Limitation	39
1.5.3 Le séquençage NGS	40
La capture des parties à séquencer, avantages et inconvénients	41
L'amplification	42
La réaction de séquence	45
1.6 L'analyse bioinformatique des données de NGS	48
1.6.1 Les données fournies par le NGS	48
Un <i>read</i> , c'est quoi ?	48
Le format FASTQ	49
1.6.2 L'alignement	50
1.6.3 L'appel des variants	52
1.6.4 L'annotation des variants	54
1.6.5 Le filtrage des variants	57
1.6.6 Conclusion NGS	58

1.7 Problématique : Un patient, 50.000 variants, à la fin il ne peut en rester qu'un. Et après ?	60
Problématique : Un patient, 50.000 variants, à la fin il ne peut en rester qu'un. Et après ?	62
Chapitre 2 : Mise en place d'une stratégie pour l'analyse des données exomiques – application en recherche clinique	63
2.1 Méthode : Description du pipeline	65
2.1.1 L'alignement des <i>reads</i>	65
2.1.2 L'appel des variants	66
2.1.3 L'annotation	69
2.1.4 Le filtrage des variants	71
2.2 Résultats 1 : Analyse de 3 phénotypes par des cas familiaux	73
2.2.1 Résultats des différentes étapes de l'analyse	74
Résultat de l'alignement	74
L'appel des variants	76
L'annotation des variants	82
Le filtrage des variants	84
2.2.2 Article n°1	89
Contexte et objectifs	90
Principaux résultats	109
2.2.3 Article n°2	111
Contexte et objectifs	112
Principaux résultats	127
2.2.4 Article n°3	129
Contexte et objectifs	130
Principaux résultats	144
2.3 Résultats 2 : Étude d'une cohorte de femmes infertiles	147
2.3.1 Article n°4	147
Contexte et objectifs	148
Principaux résultats	172
2.4 Résultats 3 : Étude d'une large cohorte de patients MMAF	173
2.4.1 Article n°5	173
Contexte et objectifs	174
Principaux résultats	236
Chapitre 3 : Investigation génétique et physiologique de la globozoospermie	239
3.1 Introduction sur la globozoospermie	240
3.2 Résultats 1 : Les mécanismes mutationnels entraînant la délétion au locus de <i>DPY19L2</i> chez l'humain	242
3.2.1 Article n°6 :	242
Contexte et objectifs	243
Principaux résultats	258

3.3 Résultat 2 : La transcriptomique	260
3.3.1 Article n°7 :	260
Contexte et objectifs	261
Principaux résultats :	273
Conclusion et discussion	278
Annexe A : Article annexe	285
Annexe B : Tables des variants restant après application des filtres pour les cas familiaux	297
Annexe C : Table des variants retrouvés sur le gène <i>PATL2</i>	301
Annexe D : Variants retrouvés au sein de notre cohorte de patients MMAF	303
References	308

Liste des tableaux

1.1	Durée de vie moyenne des cellules germinales humaines	11
2.1	Liste simplifiée des conséquences prédictes par VEP avec leur description et impact associée	70
2.2	Tableau récapitulatif des familles séquencées et de leur phénotype . .	73
2.3	Comparaison des génotypes des variants appelés par les deux procédures	80
2.4	Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs de la famille AZ	110
2.5	Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs de la famille FF	127
2.6	Liste des différents individus présentant un phénotype MMAF séquencé en WES	174
B.1	Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P1 et P2 de la famille MMAF1	298
B.2	Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P3 et P4 de la famille MMAF2	298
B.3	Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P5 et P6 de la famille MMAF3	298
B.4	Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P8 et P9 de la famille MMAF4	298
B.5	Liste des variants ayant passé l'ensemble des filtres pour le patient P10 de la famille MMAF5	300
C.1	Table des variants retrouvés sur le gène <i>PATL2</i>	302
D.1	Variants homozygotes retrouvés sur le gène <i>CFAP43</i>	304
D.2	Variants homozygotes retrouvés sur le gène <i>CFAP44</i>	305

Table des figures

1.1	La fécondation, liaison spermatozoïde-ovocyte et sortie de la méiose	11
1.2	Schéma anatomique du testicule humain	13
1.3	Les différentes phases de la spermatogenèse	15
1.4	Les différentes étapes de la méiose gamétique masculine	16
1.5	Schéma simplifié d'un enjambement chromosomique (crossing-over) . .	18
1.6	Les différentes étapes de la première division méiotique masculine . .	19
1.7	Les différentes étapes de la deuxième division méiotique masculine . .	19
1.8	Principales étapes et modifications structurales lors de la spermiogenèse	21
1.9	Anatomie simplifiée du spermatozoïde	22
1.10	Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes	24
1.11	Structure simplifiée de l'axonème	25
1.12	Structure du flagelle d'un spermatozoïde	26
1.13	Classification morphologique de spermatozoïdes humains normaux et anormaux	29
1.14	Représentation schématique du chromosome Y	31
1.15	Les différents types de translocation	32
1.16	Représentation schématique des méthodes d'analyse d'expression génique par puce à ADN	37
1.17	Méthode de génotypage par discrimination allélique par hybridation .	38
1.18	Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée	41
1.19	Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS	44
1.20	Séquençage CRT tel qu'il est effectué par Illumina	45
1.21	Séquençage SNA tel qu'il est effectué par Ion Torrent	46
1.22	Séquençage SBL tel qu'il est effectué par SOLiD	47
1.23	Présentation d'un fichier FASTQ	49
1.24	Représentation schématique de l'alignement de reads paired-end . . .	51
1.25	Illustration schématique du processus d'appel des variants	52
1.26	Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel	53
1.27	Diagramme de Venn des prédictions de pathogénicité de variants de six logiciels	55
1.28	Représentation simplifié du processus d'annotation	56
1.29	Représentation simplifié du processus de filtrage des variants	57
1.30	Récapitulatif des différentes étapes du séquençage NGS dans le cadre d'une étude phénotype-génotype	59
2.1	Détail de l'appel général effectué pour les appels DS	68
2.2	Listes des différentes conséquences prédites par VEP et leur positionnement sur le transcript	69
2.3	Représentation schématique des phasages de deux variants avec les génotypes associés	72
2.4	Processus simplifié du contrôle qualité des <i>reads</i>	75

2.5	Détermination de la valeur de couverture minimale	77
2.6	Densité de répartition du pourcentage de <i>reads</i> variants pour chaque position couverte	78
2.7	Contrôle qualité des variants appelés	79
2.8	Comparaison des variants obtenus par MAGIC + notre algorithme d'appel et BWA + GATK-HC	81
2.9	Annotation des variants	83
2.10	Filtrage des transcrits jugés "non pertinents" et des variants les chevauchant	85
2.11	Nombre d'individus et leur phénotypes composant la cohorte contrôle de chaque famille	86
2.12	Comparaison de l'efficacité de chacun des six filtres utilisés	88
2.13	Expression tissulaire des gènes <i>SPINK2</i> et <i>GUF1</i>	110
2.14	Expression tissulaire du gène <i>PLCZ1</i>	128
2.15	Couverture des six exons de <i>WBP2NL</i> pour les deux frères de la famille FF	128
2.16	Nombre de gènes passant l'ensemble des filtres par famille	146
3.1	Observation au microscope confocal de spermatozoïdes	241
3.2	Comparaison entre les spermatozoïdes des souris sauvages et globozoospermiques	241
3.3	Représentation schématique du mécanisme de NAHR	244
3.4	Quantification des sexes des souris observés lors de chaque naissance issues d'un croisement de deux souris hétérozygotes <i>Dpy19l2^{+/-}</i>	262
3.5	Principales fonctions moléculaires affectées chez les souris <i>Dpy19l2</i> KO	275
3.6	Présentation de cinq approches permettant la détection de CNVs à partir de données NGS	280
3.7	Analyse des variants restant sur chacun des 78 patients après filtrage	283

rm

Remerciements

Je remercie Messieurs le Professeurs Jacques VAN HELDEN et le Docteur Michaël MITCHELL d'avoir accepté la charge d'être rapporteurs de ma thèse.

Je remercie Madame le Professeur Christel THAUVIN et Monsieur le Docteur Julien THÉVENON d'avoir accepté de participer à mon jury et juger ce travail.

À Pierre pour m'avoir fait confiance toutes ces années et pour m'avoir permis de réaliser cette thèse.

À Nicolas, qui as dû s'arracher les yeux sur certains de mes codes, pour m'avoir inculqué, avec le temps, sa méthode et sa rigueur.

À l'ensemble des l'équipes GETI, en particulier Zine pour toutes nos discussions, les tacos et les bons moments.

À l'ensemble de l'équipe BCM, en particulier, Florient le *BCM R-Hero*, Keurcien (mon fils adoptif), Kévin et Thomas, pour toutes les discussions (du coup déterminisme ou pas ?), les barbecues, le *Fight Club*... À quand la *start-up* de la *Data Team* ?

À Carlos et Johnny Cage pour tous les fous rires que nous avons eu ensemble pendant ces dernières années.

À Agnès pour avoir pris le temps de relire mon manuscrit et corriger mes nombreuses fautes.

À mes amis, toujours présents malgré la distance.

À l'ensemble du Budo club de Malesherbes et plus particulièrement à Maître Hiram et Fanfan, qui m'ont vu grandir et qui, à travers leurs enseignements du Karaté, ont réussi à me transmettre des valeurs et un code de vie.

À ma famille pour son unité et sa solidarité.

À Dadette et Marco pour m'avoir accueilli pendant ces cinq années, et tous les débats parfois endiablés que nous avons pu avoir ensemble. “*Hakuna Matata*”.

À Aurélien pour tous ces moments de “luronage”, les grecs, les parties de tarots, nos années *Bell & Barksdale*. N'oublie jamais, “*le lion ne pactise pas avec les Hommes*”.

À mes parents, pour avoir toujours été à mes côtés quelques soit mes choix. Pour votre amour et votre dévouement indéfectible. Votre mission est désormais accomplie, profitez bien de votre vie de grands-parents.

À ma sœur pour toutes ces années de complicité. Bonne chance dans ta nouvelle vie qui commence.

À Estelle, mon garde-fou. Tu as su pendant ces années me conseiller et me soutenir sans jamais faillir. Tu as toujours été là pour moi. Ta capacité à me supporter est la preuve de ton courage et ta force. Tu m'as donné le plus beau des cadeaux en me permettant de construire, avec toi, une famille. Tu es la clef de voute sur laquelle je me repose (trop souvent). J'espère sincèrement que ces dix années ne sont que les

prémisses de notre Histoire et sache que tu pourras toujours compter sur moi comme je compte sur toi.

À Noham, mon fils, ma plus grande fierté et source de bonheur. Tu as bouleversé ma vie . Ton seul rire à le pouvoir de me rendre heureux et de me faire oublier tous mes tracas. Puisses-tu mener ta vie comme tu l'entends, faire tes propres choix, tes propres erreurs en sachant que je te soutiendrai toujours. Je t'aime.

Résumé

Ces dix dernières années, l'investigation des maladies génétiques a été bouleversée par l'émergence des techniques de séquençage haut-débit. Celles-ci permettent désormais de ne plus séquencer les gènes un par un, mais d'avoir accès à l'intégralité de la séquence génomique ou transcriptomique d'un individu. La difficulté devient alors d'identifier les variants causaux parmi une multitude d'artefacts techniques et de variants bénins, pour ensuite comprendre la physiopathologie des gènes identifiés.

L'application du séquençage haut débit est particulièrement prometteuse dans le champ de la génétique de l'infertilité masculine car il s'agit d'une pathologie dont l'étiologie est souvent génétique, qui est génétiquement très hétérogène et pour laquelle peu de gènes ont été identifiés. Mon travail de thèse est donc centré sur la l'infertilité et comporte deux parties majeures : l'analyse des données issues du séquençage haut débit d'homme infertiles et de modèles animaux et la caractérisation moléculaire d'un phénotype spécifique d'infertilité, la globozoospermie.

Le nombre de variants identifiés dans le cadre d'un séquençage exomique pouvant s'elever à plusieurs dizaines de milliers, l'utilisation d'un outil informatique performant est indispensable. Pour arriver à une liste de variants suffisamment restreinte pour pouvoir être interprétée, plusieurs traitements sont nécessaires. Ainsi, j'ai développé un pipeline d'analyse de données issues de séquençage haut-débit effectuant de manière successive l'intégralité des étapes de l'analyse bio-informatique, c'est-à-dire l'alignement des *reads* sur un génome de référence, l'appel des génotypes, l'annotation des variants obtenus ainsi que le filtrage de ceux considérés comme non pertinents dans le contexte de l'analyse. L'ensemble de ces étapes étant interdépendantes, les réaliser au sein du même pipeline permet de mieux les calibrer pour ainsi réduire le nombre d'erreurs générées. Ce pipeline a été utilisé dans cinq études au sein du laboratoire, et a permis l'identification de variants impactant des gènes candidats prometteurs pouvant expliquer le phénotype d'infertilité des patients. L'ensemble des variants retenus ont ensuite pu être validés expérimentalement.

J'ai également pris part aux investigations génétiques et moléculaires permettant la caractérisation du gène *DPY19L2*, identifié au laboratoire et dont la délétion homozygote entraîne une globozoospermie, caractérisée par la présence dans l'éjaculat de spermatozoïdes à tête ronde dépourvus d'acrosome. Pour cela, j'ai contribué à caractériser les mécanismes responsables de cette délétion récurrente, puis, en utilisant le modèle murin *Dpy19l2 knock out* (KO) mimant le phénotype humain, j'ai réalisé une étude comparative des transcriptomes testiculaires de souris sauvages et de souris KO *Dpy19l2^{-/-}*. Cette étude a ainsi permis de mettre en évidence la dérégulation de 76 gènes chez la souris KO. Parmi ceux-ci, 23 sont impliqués dans la liaison d'acides nucléiques et de protéines, pouvant ainsi expliquer les défauts d'ancre de l'acrosome au noyau chez les spermatozoïdes globozoocéphales.

Mon travail a donc permis de mieux comprendre la globozoospermie et de développer un pipeline d'analyse bioinformatique qui a déjà permis l'identification de plus de 15 gènes de la gamétogénèse humaine impliqués dans différents phénotypes d'infertilité.

Abstract

In the last decade, the investigations of genetic diseases have been revolutionized by the rise of high throughput sequencing (HTS). Thanks to these new techniques it is now possible to analyze the totality of the coding sequences of an individual (exome sequencing) or even the sequences of his entire genome or transcriptome. The understanding of a pathology and of the genes associated with it now depends on our ability to identify causal variants within a plethora of technical artifact and benign variants.

HTS is expected to be particularly useful in the field infertility as this pathology is expected to be highly genetically heterogeneous and only a few genes have so far been associated with it. My thesis focuses on male infertility and is divided into two main parts : HTS data analysis of infertile men and the molecular characterization of a specific phenotype, globozoospermia.

Several thousands of distinct variants can be identified in a single exome, thereby using effective informatics is essential in order to obtain a short and actionable list of variants. It is for this purpose that I developed a HTS data analysis pipeline performing successively all bioinformatics analysis steps : 1) reads mapping along a reference genome, 2) genotype calling, 3) variant annotation and 4) the filtering of the variants considered as non-relevant for the analysis. Performing all these independent steps within a single pipeline is a good way to calibrate them and therefore to reduce the number of erroneous calls. This pipeline has been used in five studies and allowed the identification of variants impacting candidate genes that may explain the patients' infertility phenotype. All these variants have been experimentally validated using Sanger sequencing.

I also took part in the genetic and molecular investigations which permitted to demonstrate that the absence of the *DPY19L2* gene induces male infertility due to globozoospermia, the presence in the ejaculate of only round-headed and acosomeless spermatozoa. Most patients with globozoospermia have a homozygous deletion of the whole gene. I contributed to the characterization of the mechanisms responsible for this recurrent deletion, then, using *Dpy19l2* knockout (KO) mice, I realized the comparative study of testicular transcriptome of wild type and *Dpy19l2*^{-/-} KO mice. This study highlighted a dysregulation of 76 genes in KO mice. Among them, 23 are involved in nucleic acid and protein binding, which may explain acosome anchoring defaults observed in the sperm of globozoospermic patients.

My work allowed a better understanding of globozoospermia and the development of a HTS data analysis pipeline. The latter allowed the identification of more than 15 human gametogenesis genes involved in different infertility phenotypes.

CHAPITRE 1

Introduction

1.1 L'ovogenèse et l'ovocyte

Chez l'humain, la production d'ovocyte est un processus long commençant dès le développement embryonnaire à ce stade, les ovocytes sont immatures et rentrent en phase de quiescence après avoir débuté la MI. 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 entammer la MII jusqu'au stade de la métaphase. La MII se poursuivra ensuite dans le cas d'une fécondation [1].

The zona pellucida (ZP) is a specialized extracellular coat that surrounds the plasma membrane of mammalian eggs. Its presence is essential for successful completion of oogenesis, fertilization and preimplantation development. The ZP is composed of only a few glycoproteins which are organized into long crosslinked fibrils that constitute the extracellular coat. résumé de l'intrò PATL2 Several reports describe that some infertile women repetitively produce mostly immature oocytes, a poorly-defined syndrome known as “oocyte factor” or “bad eggs syndrome” [2–5]. Heterozygous mutations of TUBB8, an oocyte specific tubulin necessary for the meiotic spindle were recently identified in a cohort of Chinese patients with OMD [6], establishing TUBB8 as the first human gene identified in the context of OMD.

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 [7]. La première étape de la fécondation consiste en la liaison du spermatozoïde à la zone pellucide de l'ovocyte [8]. 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+} [9]. 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 [9, 10]. Ces observations ont rapidement permis l'émergence de l'hypothèse d'un facteur spermatique qui, lors de la fécondation, était relargé et généreraient ces oscillations Ca^{2+} [9, 11]. Cette hypothèse étant supportée par deux principales observations. Tout d'abord, le fait que la fusion des cytoplasmes du spermatozoïdes et de l'ovocytes était le prélude de ces oscillations Des expérimentations [12, 13]. 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 [11, 14–17]. C'est en 2002 que la protéine PLC zeta (ζ) fut pour la première fois reporté comme étant, chez la souris, responsable de ces oscillations déclanchant ensuite une cascade de réactions dont découle éventuellement l'activation ovocytaire et le développement embryonnaire [18] (Figure : 1.1).

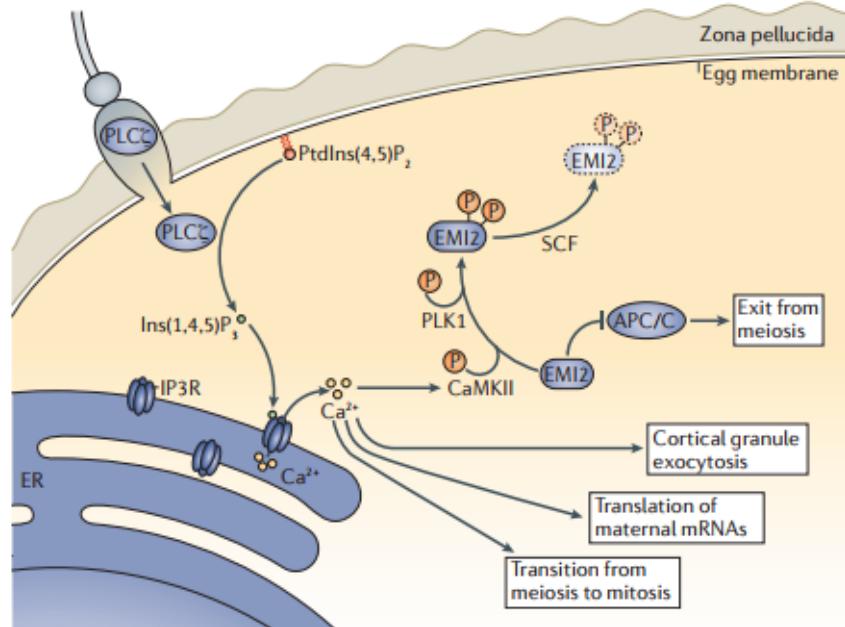


Figure 1.1 – La fécondation, liaison spermatozoïde-ovocyte et sortie de la méiose adapté d'après [19] : Suite à la fusion spermatozoïde-ovocyte, la protéine PLC ζ induit la production de nositol 1,4,5-trisphosphate (Ins(1,4,5)P₃) qui se lie à son récepteur causant ainsi le relargage de Ca²⁺ qui, entre autre, fera sortir l'ovocyte de son stade de méiose pour rentrer en mitose

1.2 La spermatogenèse

La spermatogenèse des mammifères est un processus long et complexe contrôlé par plusieurs mécanismes étroitement liés [20–22]. 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 spermatozoïdes. 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 [23]. Le cycle spermatogénétique est défini comme la succession chronologique des différents stades de différenciation d'une génération de cellules germinales (depuis la spermatogonie jusqu'au spermatozoïde). Chacune des étapes du cycle spermatogénétique a une durée fixe et constante selon les espèces (**Table : 1.1**).

Table 1.1 – Durée de vie moyenne des cellules germinales humaines

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

1.2.1 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 : 1.2**). Cette externalisation des testicules permet leur maintien à une température plus basse que celle du reste du corps, nécessaire à la spermatogenèse.

L'intérieur du testicule contient des tubes séminifères enroulés ainsi que du tissu entre les tubules appelé espace interstitiel. Les tubes séminifères sont de longs tubes compactés sous forme de boucles et dont les deux extrémités débouchent sur le *rete testis* (**Figure : 1.2**). C'est le long des parois du tube séminifère que se déroulera l'ensemble des étapes de la spermatogenèse.

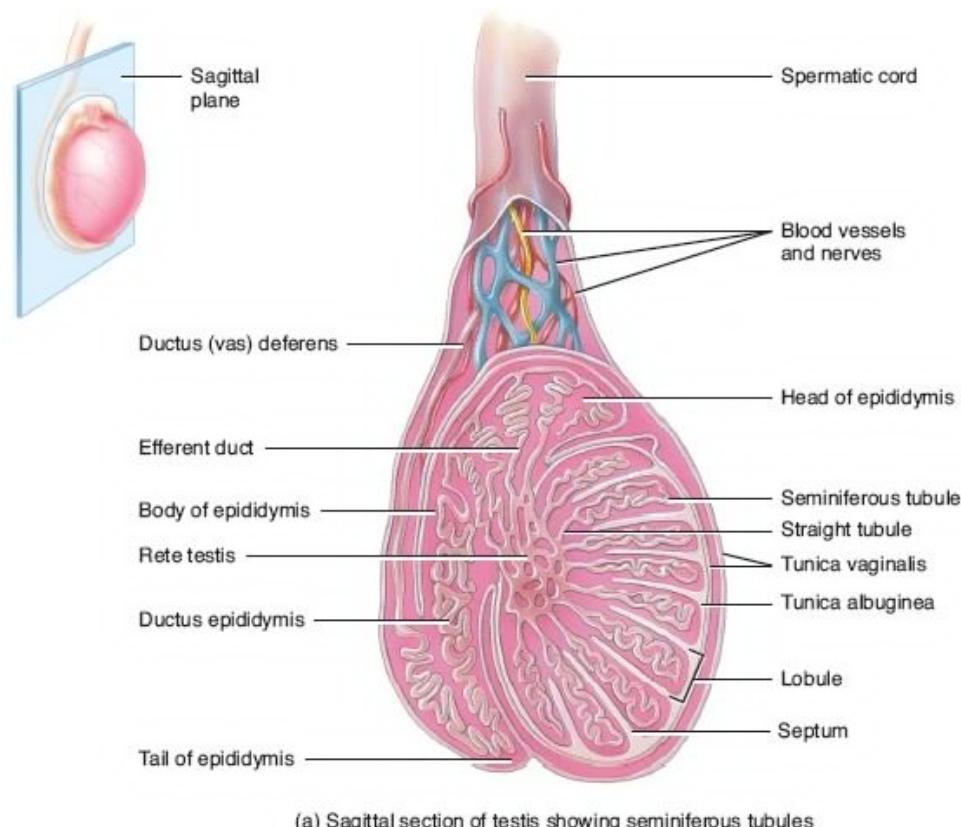


Figure 1.2 – Schéma anatomique du testicule humain.

1.2.2 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 [24–26] : 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 : 1.3**). 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 : 1.3**).

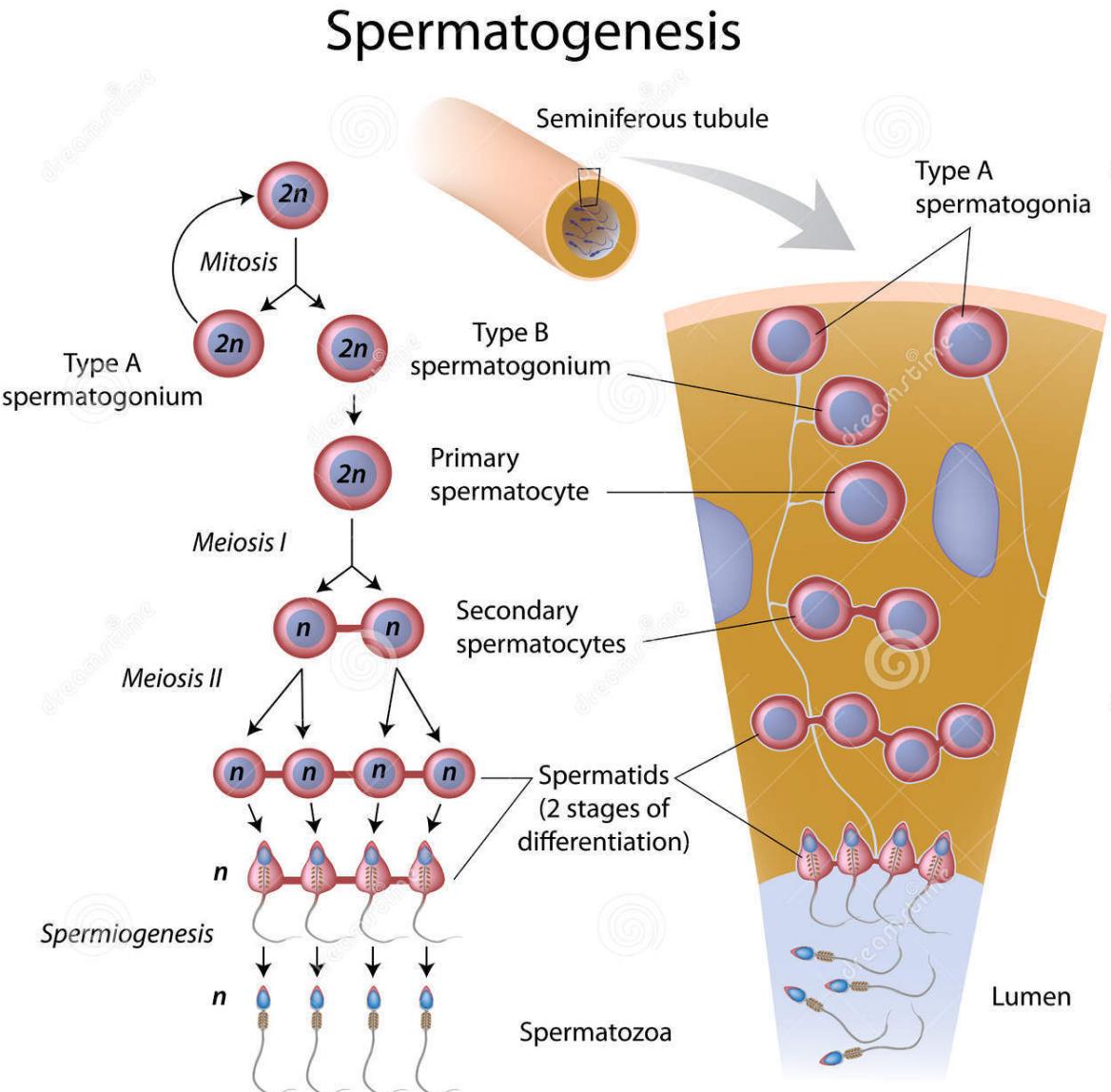


Figure 1.3 – Les différentes phases de la spermatogenèse
d'après medizin-kompakt : L'évolution de la spermatogenèse est strictement coordonnée, tant dans le sens transversal des tubules, que dans le sens longitudinal. Le sens transversal correspond au sens unique de la différenciation germinale, les 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.

1.2.3 La méiose

La méiose, ou phase de maturation, est l'étape au cours de laquelle, à partir de cellules diploïdes (les spermatogonies B) vont se former des cellules haploïdes, les spermatocytes secondaires (spermatocytes II). Ce résultat est le fruit de deux divisions successives (**Figure : 1.4**) 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 : 1.5**).

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 : 1.4**) :

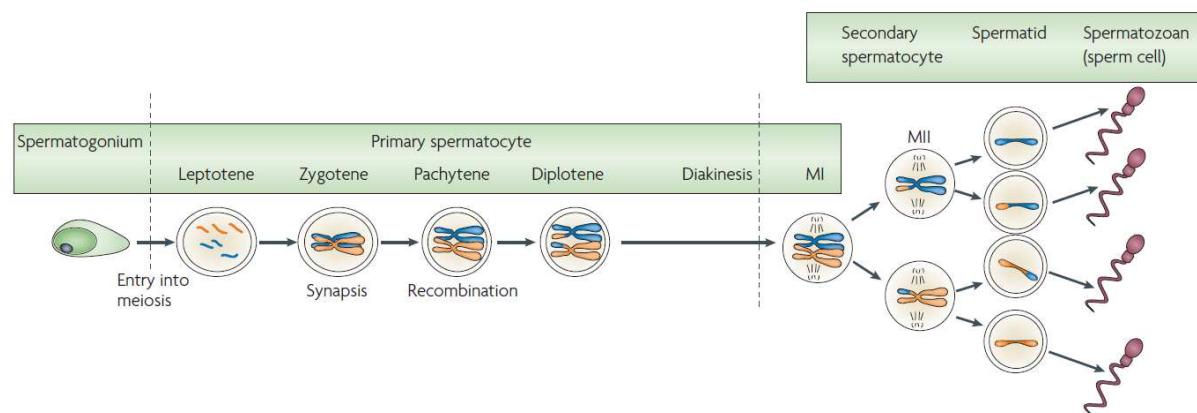


Figure 1.4 – Les différentes étapes de la méiose gamétique masculine d'après [27] : Les spermatogonies B servent de point d'initiation et se différencient en spermatocyte primaire. Les cinq étapes de la prophase I sont suivies d'une première division cellulaire, donnant deux cellules haploïdes au sein 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 : 1.6)

- 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 : 1.5).
 - 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) [28].
 - 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 : 1.7) la MII est similaire à une division mitotique et peut se décomposer en quatre parties distinctes :

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

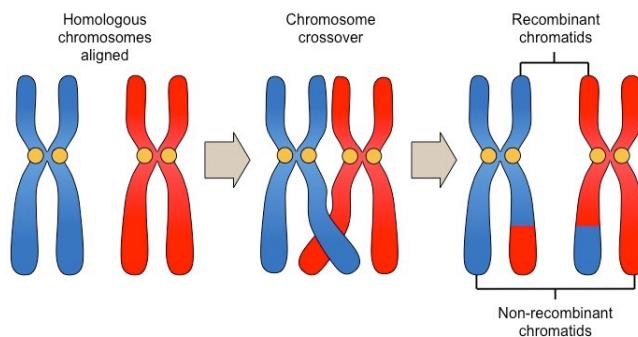


Figure 1.5 – Schéma simplifié d'un enjambement chromosomique (crossing-over).

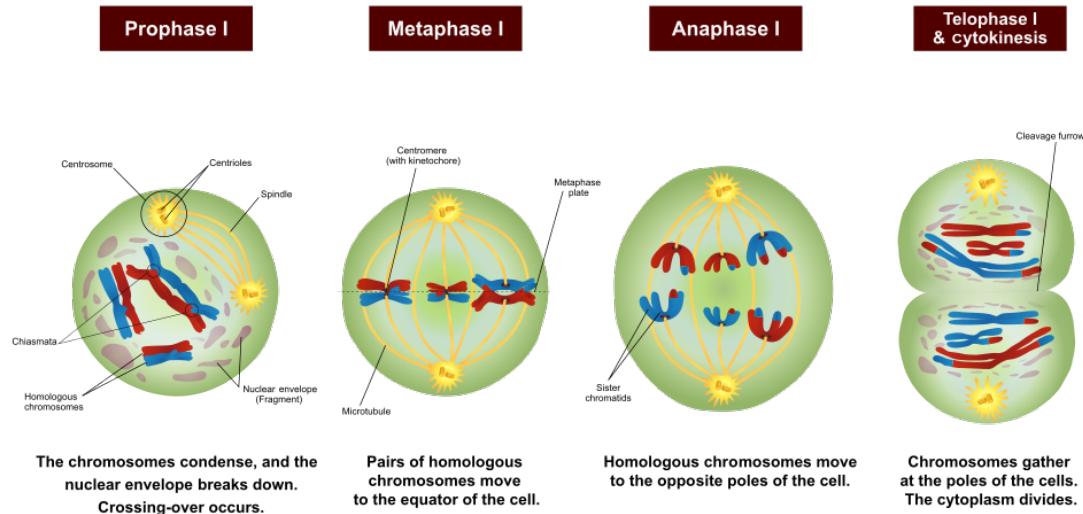


Figure 1.6 – *Les différentes étapes de la première division méiotique masculine* adapté d'après [29].

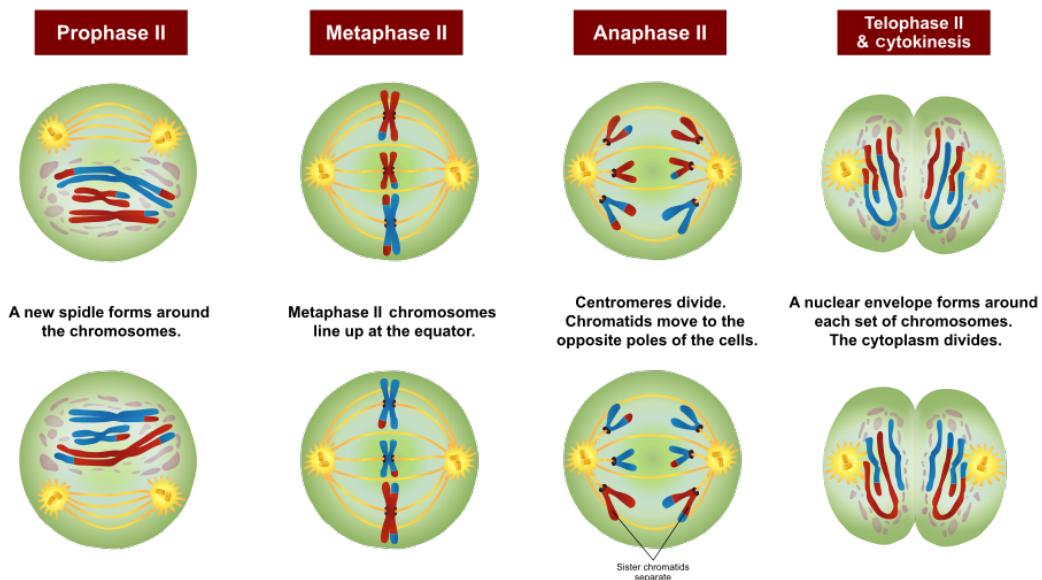


Figure 1.7 – *Les différentes étapes de la deuxième division méiotique masculine* adapté d'après [29].

1.2.4 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 : 1.8**). 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 [30]. 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 [31]. Pendant l'élongation de la spermatide, le noyau se condense et devient hautement polarisé [32]. Les spermatides sont situées dans le compartiment adluminal, à proximité de la lumière du tube séminifère. Ce sont de petites cellules (8 à 10 µm) que l'on peut schématiquement diviser en trois classes :

1. **Les spermatides rondes** (**Figure : 1.8 - 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 [24, 33–36]. Plusieurs études animales ont pu démontrer le potentiel des spermatides rondes à donner la vie à des individus sains et fertiles, [37–39], la même chose ayant été également observée plus récemment chez l'homme [40] bien que le taux de fécondation et d'implantation soit extrêmement faible [41]. Ils possèdent un noyau rond avec une chromatine pâle et homogène. C'est à partir de ces étapes que démarre la biogenèse de l'acrosome avec la production par l'appareil de Golgi des vésicules pro-acrosomales (phase de Golgi). Les deux centrioles contenus dans le cytoplasme vont se déplacer au futur pôle caudal. Le centriole proximal est inactif alors que le centriole distal donne naissance à un ensemble de microtubules à l'origine de l'axonème du futur flagelle.
2. **Les spermatides en élongation** (**Figure : 1.8 - 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 [42]. 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 [41].

3. Les spermatides en condensation (Figure : 1.8 - 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 physiologique 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 [43].

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

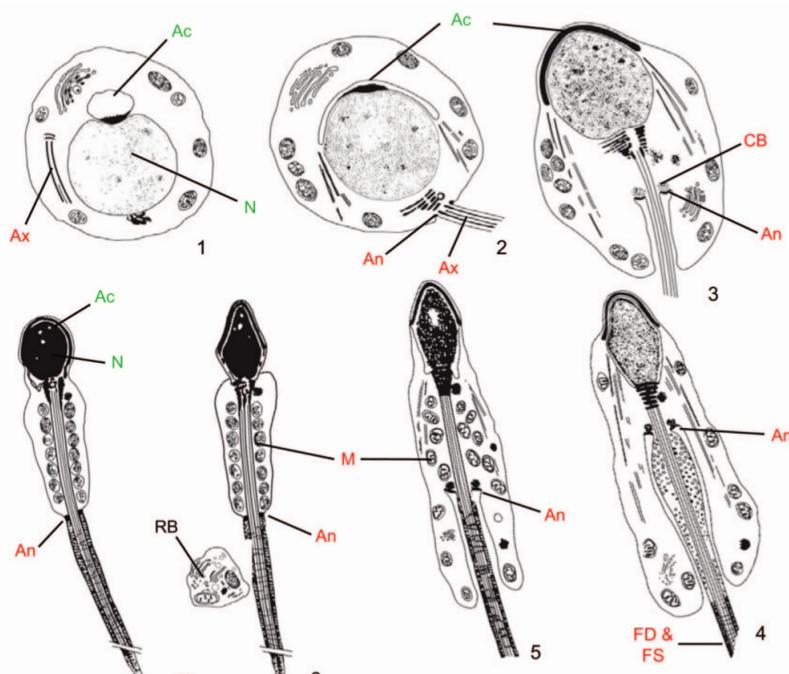


Figure 1.8 – Principales étapes et modifications structurales lors de la spermiogenèse d'après [44] : 1. La spermatide immature avec un gros noyau arrondi. La vésicule acrosomale est attachée au noyau, l'ébauche du flagelle n'atteint pas le noyau. 2. La vésicule acrosomale a augmenté de taille et apparaît 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.

1.3 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 : 1.9**). 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.

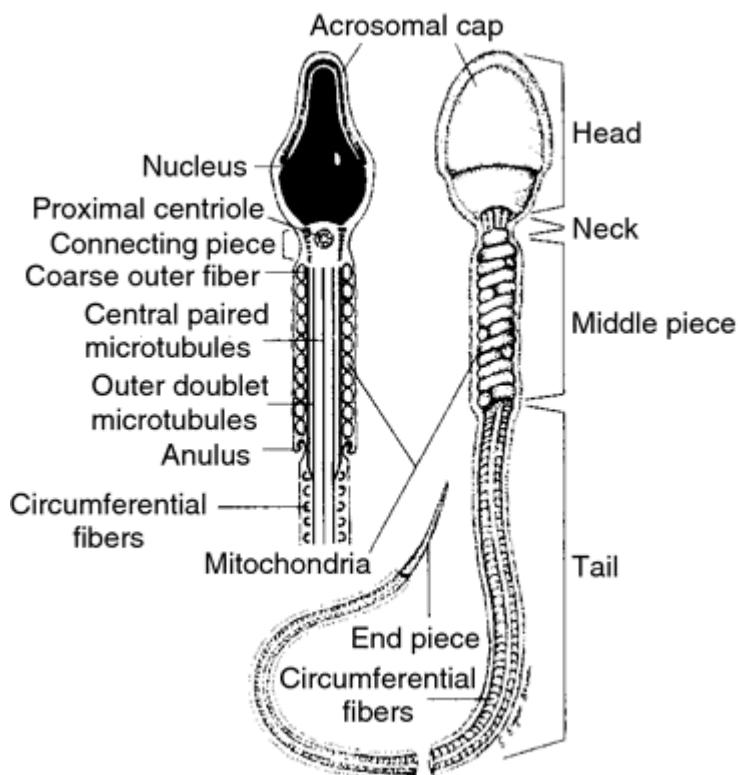


Figure 1.9 – *Anatomie simplifiée du spermatozoïde* d'après medical-dictionary.

1.3.1 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 [45]. 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** : 1.10). 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** : 1.10). L'intégrité des deux protéines composant ce dimère est nécessaire pour la procréation [46]. 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 [47]. Le noyau du spermatozoïde est donc un noyau au repos, transcriptionnellement inactif [48].

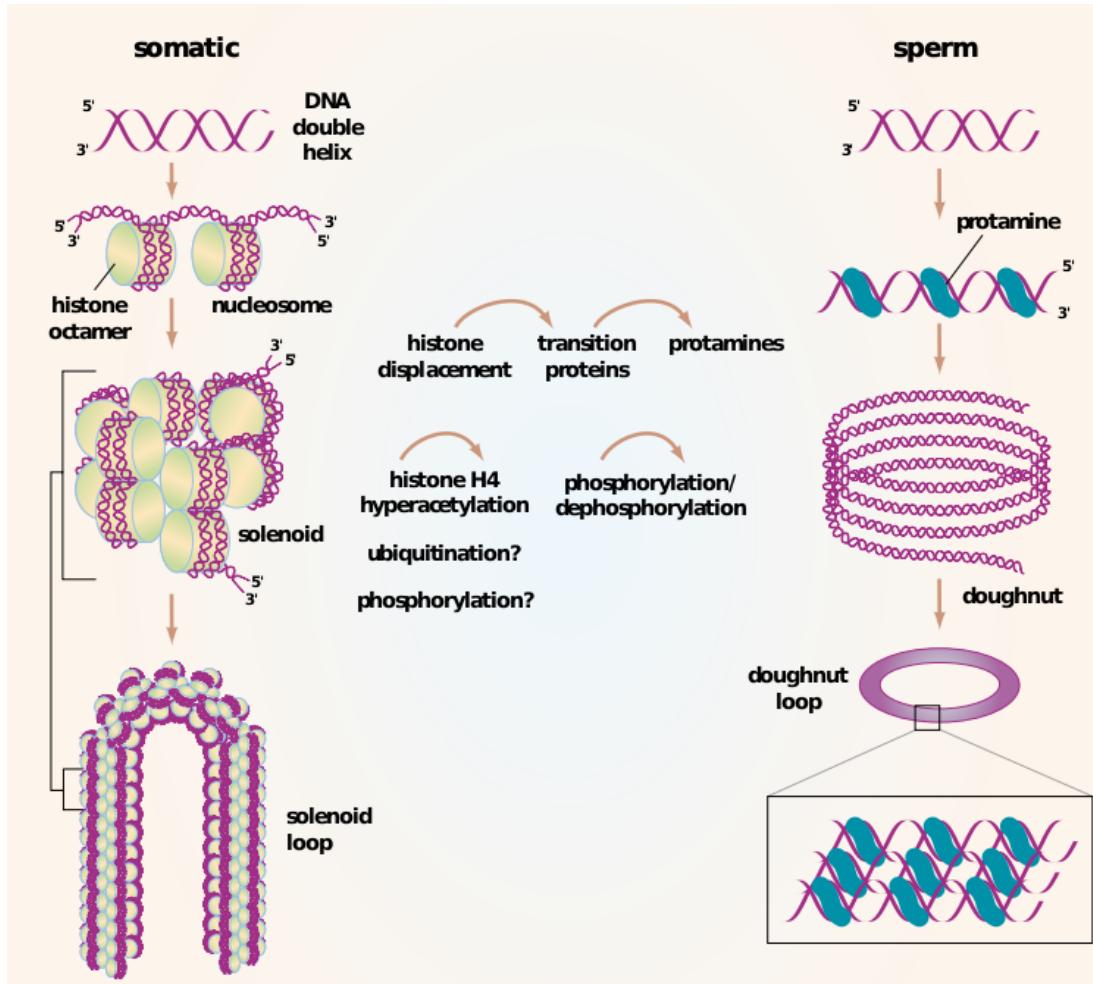


Figure 1.10 – Schéma de la compaction de l'ADN dans les cellules somatiques et dans les spermatozoïdes d'après [49] : Dans les cellules somatiques, l'ADN est enroulé sous forme de nucléosome. Les nucléosomes vont s'agencer entre eux pour former un solénoïde qui sera attaché à la matrice nucléaire par 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.

1.3.2 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 [50] (**Figure : 1.11**), 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.

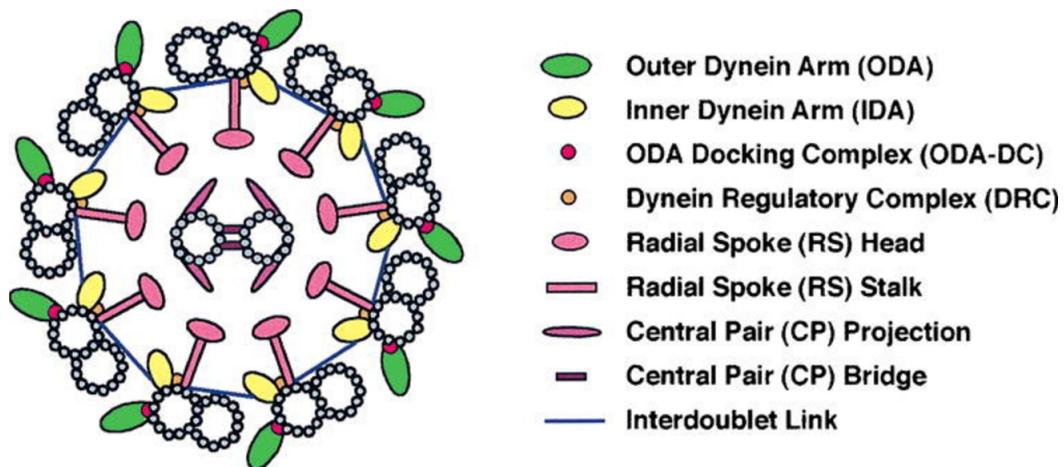


Figure 1.11 – Structure simplifiée de l'axonème d'après [50] : L'axonème est constitué de neuf doublets de microtubules périphériques reliés entre eux par des liens de nexine et d'un doublet central relié aux doublets périphériques par des ponts radiaux.

Le flagelle du spermatozoïde peut être divisé en trois parties distinctes (**Figure : 1.12**) :

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 [51].
3. **La pièce terminale** : elle est située au niveau de l'extrémité distale du flagelle et ne contient que l'axonème [50].

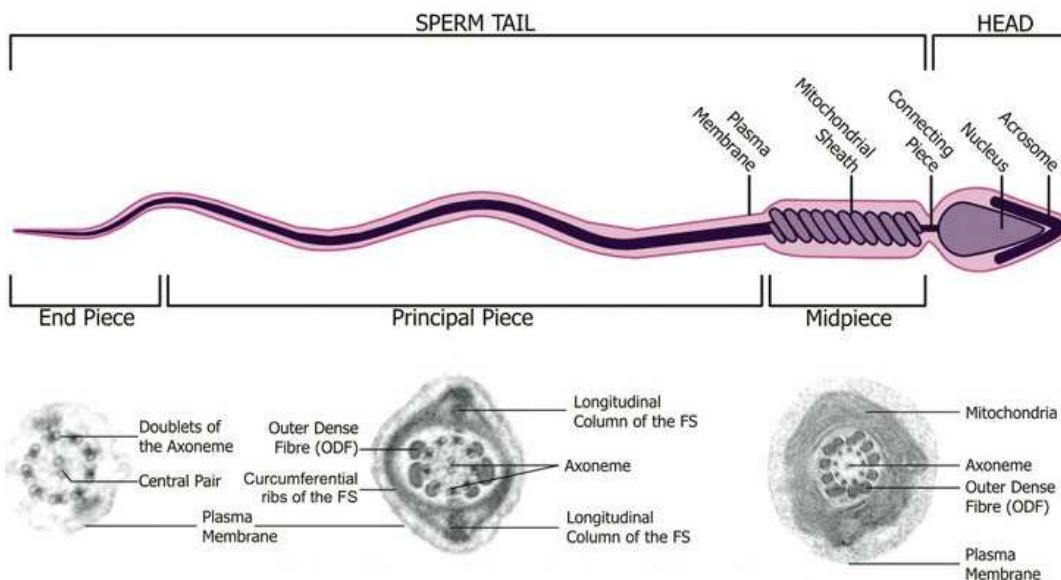


Figure 1.12 – Structure du flagelle d'un spermatozoïde d'après [52] : Coupes transversales en microscopie électronique. Le flagelle se compose de trois parties : la pièce intermédiaire, contenant les mitochondries, la pièce principale et la pièce terminale. L'axonème, en position centrale, parcourt tout le flagelle. Des structures périaxonémiales sont observables : les fibres denses dans la pièce intermédiaire et principale, et la gaine fibreuse dans la pièce principale seulement.

1.4 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 [53]. 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 [54].

1.4.1 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” [55], chez les individus présentant une quantité basse de spermatozoïdes [57, 58] ou en étant dépourvu [59]. Cependant, leur nombre augmente tandis que la quantité de spermatozoïde diminue [60].

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 [61].

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 [62]. Un arrêt de la spermatogenèse a été observé dans 4 à 30% des biopsies testiculaires des hommes présentant une oligospermie sévère [63–66]. 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) [67].
2. **L'azoospermie** : comme l'oligozoospermie, l'azoospermie est un phénotype d'infertilité masculine cette fois-ci caractérisé par l'absence totale de spermatozoïdes dans l'éjaculat. On distingue des causes excrétoires empêchant l'excrétion des spermatozoïdes, on parle alors d'azoospermie obstructive et des causes sécrétoires, les plus fréquentes, accompagnées d'un défaut de la spermatogenèse, on parle alors d'azoospermie non-obstructive.

Anomalies liées à la morphologie

Ces anomalies sont observables en effectuant un spermocytogramme. Plusieurs classifications ont été établies. Cependant, c'est la classification de David modifiée (**Table** : 1.13) 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.

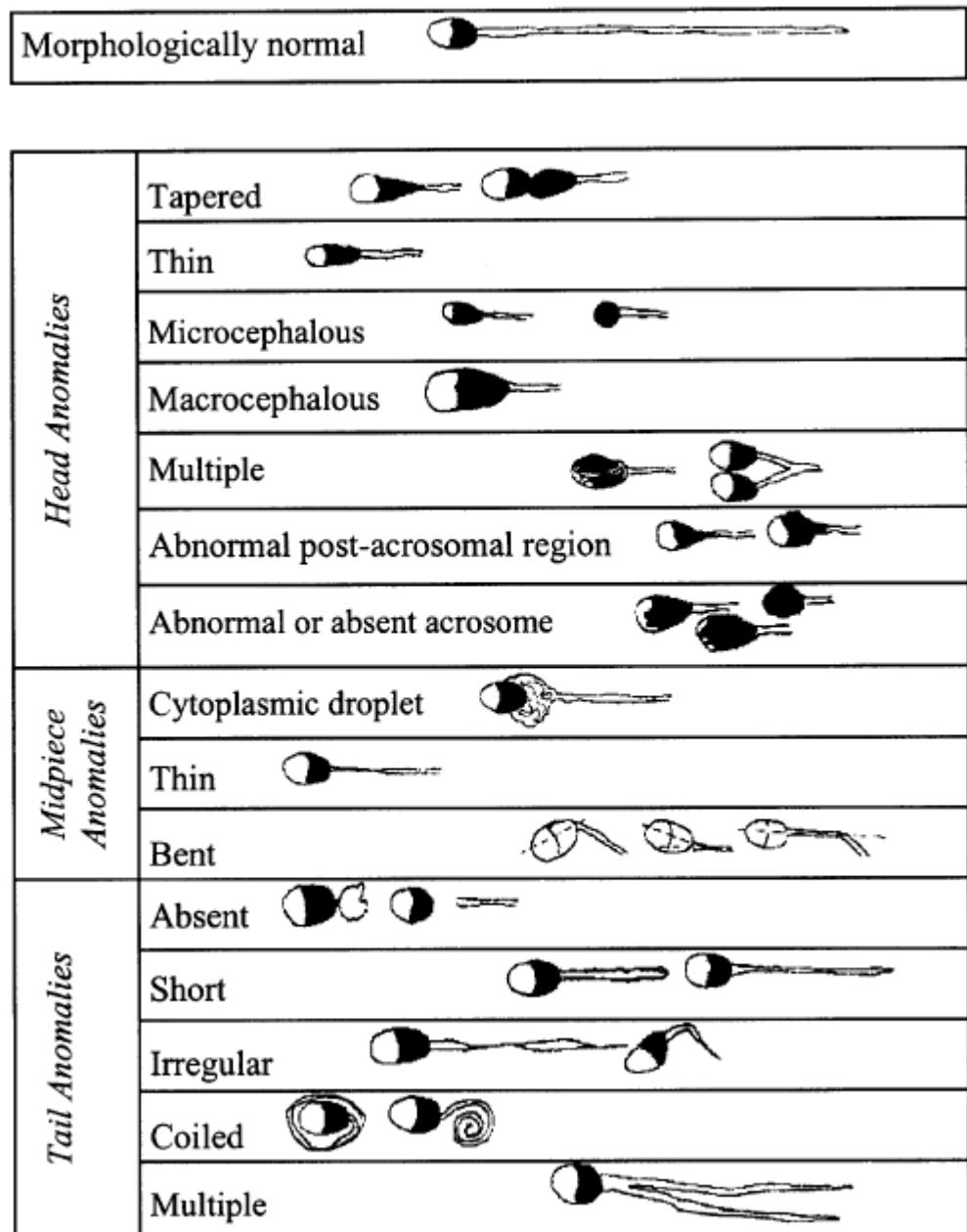


Figure 1.13 – *Classification morphologique de spermatozoïdes humains normaux et anormaux d'après [68].*

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 [69, 70]. 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 [71].

1.4.2 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 [72].

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 [73]. 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** : 1.14). 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) [74].

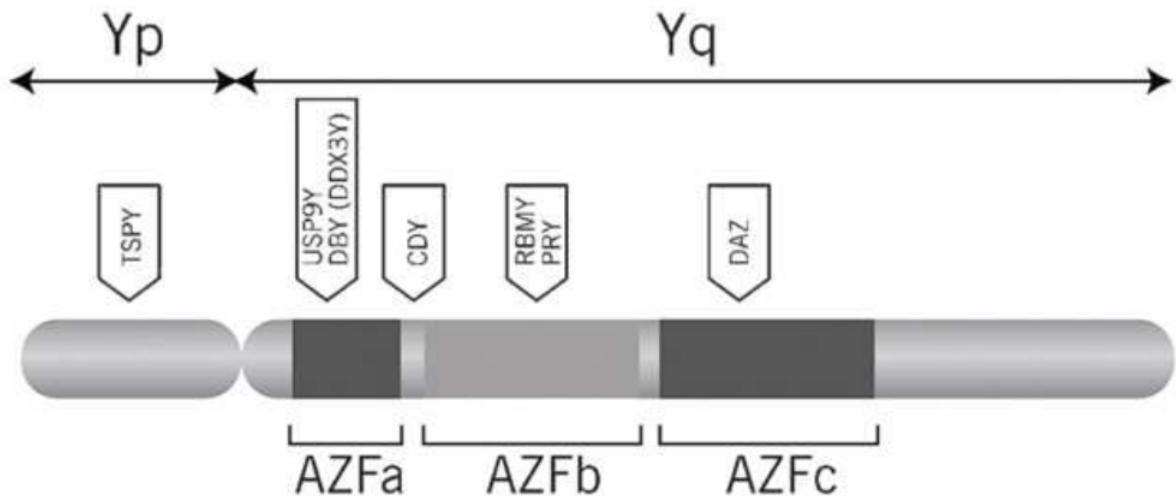


Figure 1.14 – Représentation schématique du chromosome Y adapté d'après [75] : Visualisation la région AZF ainsi que des trois sous-régions AZF a, b, c et des principaux gènes compris dans chacune des sous-régions.

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 [76].
 - a. **Syndrome de Klinefelter** : le syndrome de Klinefelter (ou 46, 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) [77] mais elle est environ 50 fois supérieure chez les patients infertiles azoospermiques [78].
 - 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 : 1.15 - 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 [79]. Les secondes (**Figure : 1.15 - 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 [75].

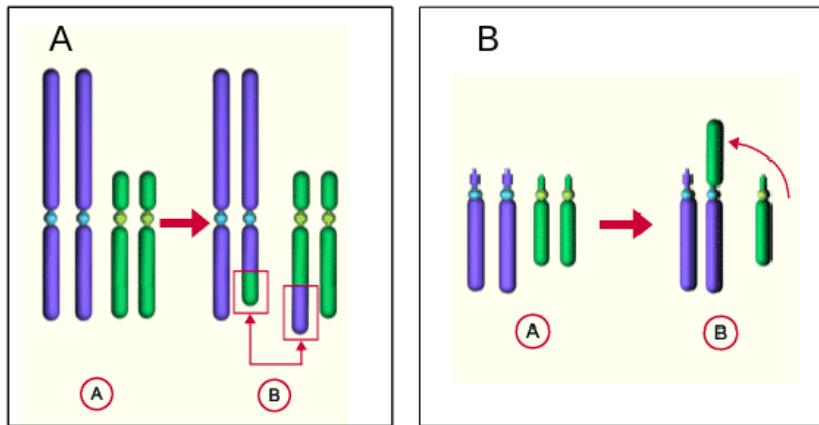


Figure 1.15 – Les différents types de translocation d'après embryology.ch : **A** : La translocation réciproque. **B** : La translocation robertsonniène

- 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 [80].
 - 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 [81] 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 [82].

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** : 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 ; de même une étude de 2016 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 [83]. De même, plusieurs variants dans le gène *TEX111* et *SYC1* ont été décrits comme entraînant un arrêt de la méiose [84–86].
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 [87], 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 [88]. Depuis, d'autres études ont permis d'associer d'autres variants sur ce même gène à ce phénotype [89]. Des anomalies du gène *AURKC* seraient ainsi responsables d'environ 83.7% des cas macrozoospermie chez des patients non apparentés [88]. 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 [90].
 - 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 [91–93]. 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*) [94] dont la protéine va, au cours de la spermatogenèse, fusionner avec les vésicules proacrosomales pour former l'acrosome [94, 95]. 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 [96]. cf globo
 - c. **Spermatozoïdes acéphaliques** : Ce phénotype rapporté plusieurs fois [97–99] 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* [100] qui avait précédemment été décrit comme localisant à la jonction noyau / flagelle du spermatozoïde [101].

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... [102, 103]. 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 [103]. 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 [104, 105].
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 Ca²⁺. Ce processus est médié par une protéine spécifique du spermatozoïde, la *phospholipase C Zeta 1* (PLCζ1) codée par le gène *PLCZ1* [106, 107]. Plusieurs cas d'échec d'activation ovocytaire ont été liés à l'absence ou à la mauvaise localisation de la protéine PLCζ1. 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 [108] et un peu plus tard une mutation homozygote chez deux frères consanguins [109].

1.5 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.

1.5.1 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 [110]. 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 [111] 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 *SYCP3*, *SOHLH1* et *NR5A1* entraînant tous les trois un phénotype d'azoospermie [112–114].

1.5.2 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 [115–118], mais aussi la réponse physiologique à plusieurs types de stimuli tel que la prise de certains médicaments [119].

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 [120]. 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.

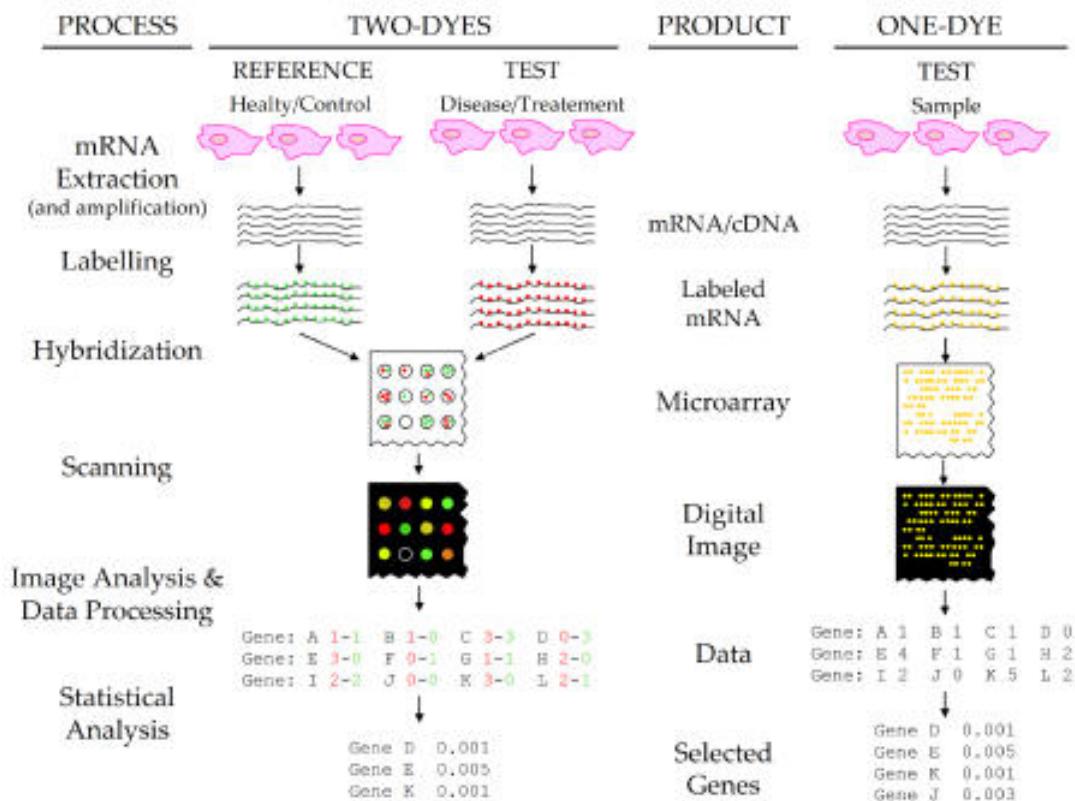


Figure 1.16 – Représentation schématique des méthodes d'analyse d'expression génique par puce à ADN d'après [121] : Présentation des méthodes à double et à simple colorant, respectivement à gauche et à droite. Pour les analyses à double colorant, une seule puce et nécessaire, les échantillons de la référence 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 [122] malgré le “bruit de fond” causé par l’hybridation non spécifique dont elle souffre (**Figure : 1.17**).

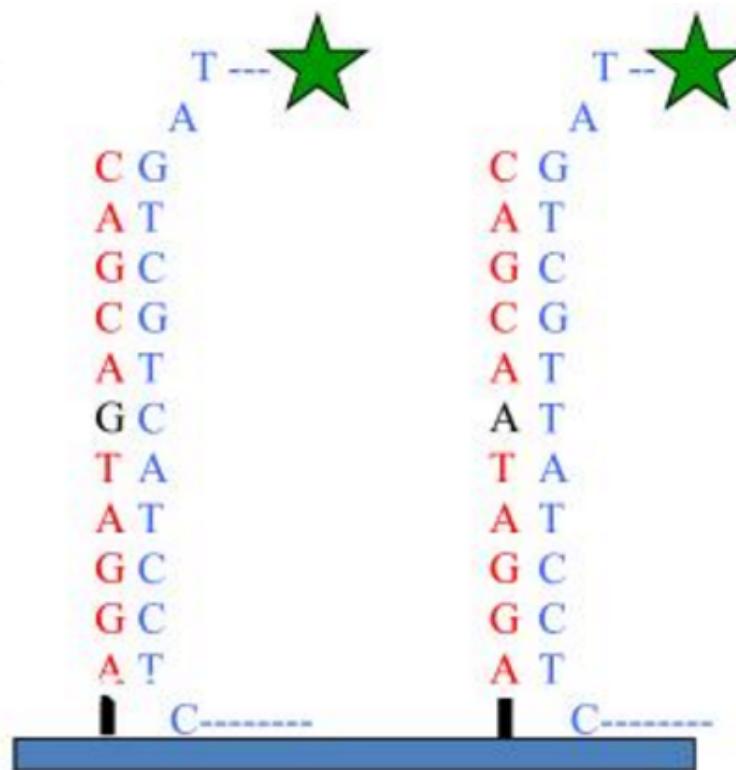


Figure 1.17 – Méthode de génotypage par discrimination allélique par hybridation d'après [123] : Des sondes complémentaires à chacun des allèles sont positionnées sur la puce. L'ADN génomique fragmenté et labélisé est mis en contact de la puce. Après nettoyage de la puce, l'analyse du signal émis par 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 [124]. Son principe est très similaire à celui utilisé dans les puces à expression (**Figure : 1.16**) 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.

1.5.3 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 [125].

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 [126]. 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 : 1.18**) ; 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 [127]. La couverture est définie comme le nombre de *reads* qui, après l'étape d'alignement, se chevauchent les uns les autres au sein d'une région génomique spécifique. Par exemple, une couverture de 30x pour le gène XXXX signifie que chaque nucléotide de ce gène est chevauché par au moins 30 *reads* distincts.

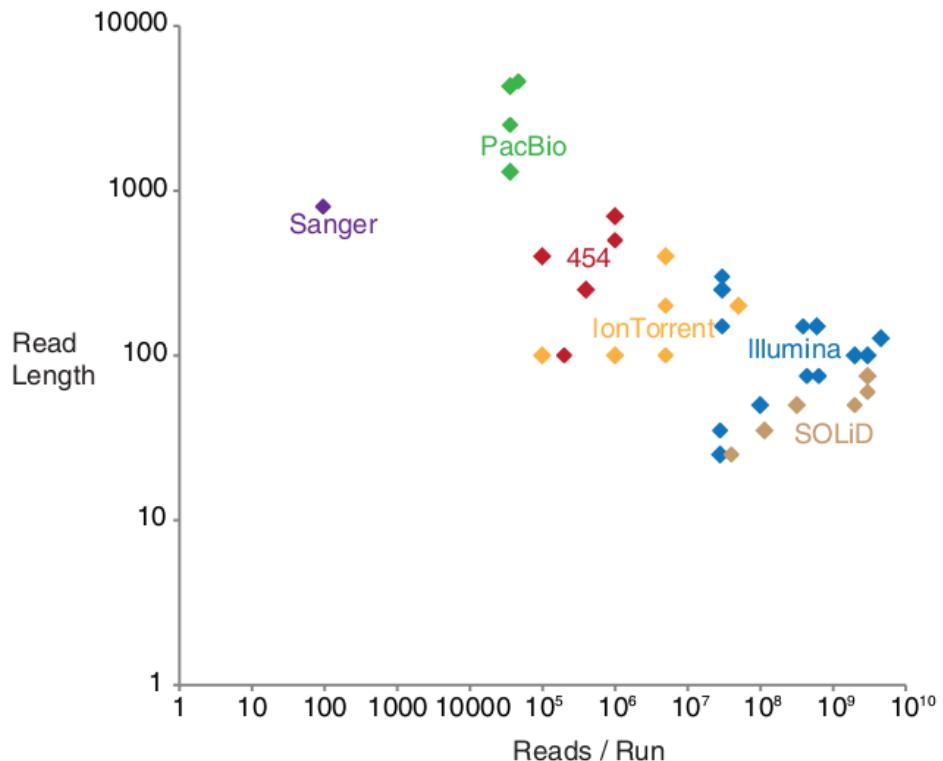


Figure 1.18 – Présentation de la taille des reads et du nombre de reads par run en fonction de la technologie de séquençage utilisée d'après [128] : Chaque point représente une plateforme de séquençage, la couleur détermine la marque du séquenceur.

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 [129]. 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 [130, 131]. 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 [131]. 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 [131].

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 : 1.19 - 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 : 1.19 - 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 : 1.19 - 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.

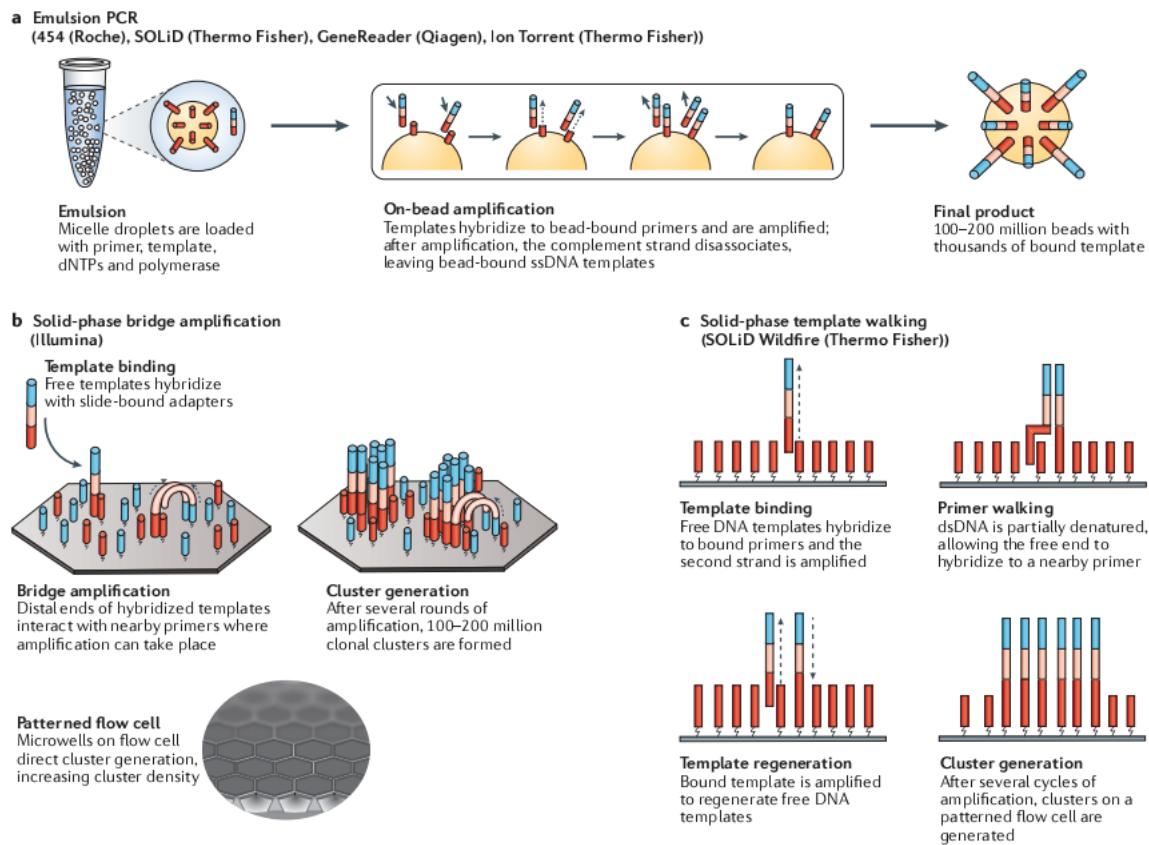


Figure 1.19 – Présentation des différentes stratégies d'amplification de l'ADN dans le cadre du NGS d'après [132] : a : PCR en émulsion. b : amplification par pont. c : amplification par modèle mobile.

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 [132] :
 - a. **Terminaison par cycle réversible, cyclic reversible termination (CRT)** (**Figure : 1.20**) : Cette méthode est caractérisée par l'utilisation de molécules terminatrices auxquelles le groupement 3' – OH est modifié de sorte à éviter l'elongation [133], on parlera de groupement 3' – bloqué. Une amorce liée au fragment d'ADN permettra l'initialisation du processus de polymérisation. À chaque cycle, un mix comprenant l'ensemble des quatre désoxynucléotides (dNTPs), préalablement labélisés par un fluorophore 3' – bloqué, est mis en contact du fragment. Après l'incorporation d'un unique dNTP au fragment, les dNTPs non liés sont éliminés et la nature du dNTP ajouté est identifiée grâce à son fluorophore. Le fluorophore et le groupement 3' – bloqué sont retirés permettant ainsi à un nouveau cycle de commencer.

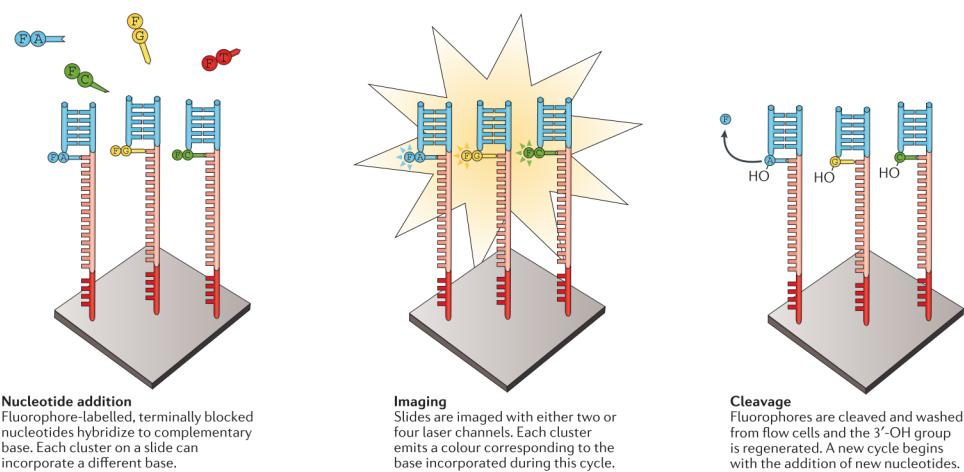


Figure 1.20 – Séquençage CRT tel qu'il est effectué par Illumina d'après [132] : a : ajout d'un dNTP labellisé par un fluorophore 3'-bloqué. b : identification du dNTP ajouté grâce au fluorophore. c : le fluorophore est clivé du dNTP et le groupement 3'-OH est reformé à partir du groupement 3'-bloqué permettant ainsi l'elongation.

b. **Addition de nucléotides uniques (SNA)** (**Figure : 1.21**) : 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.

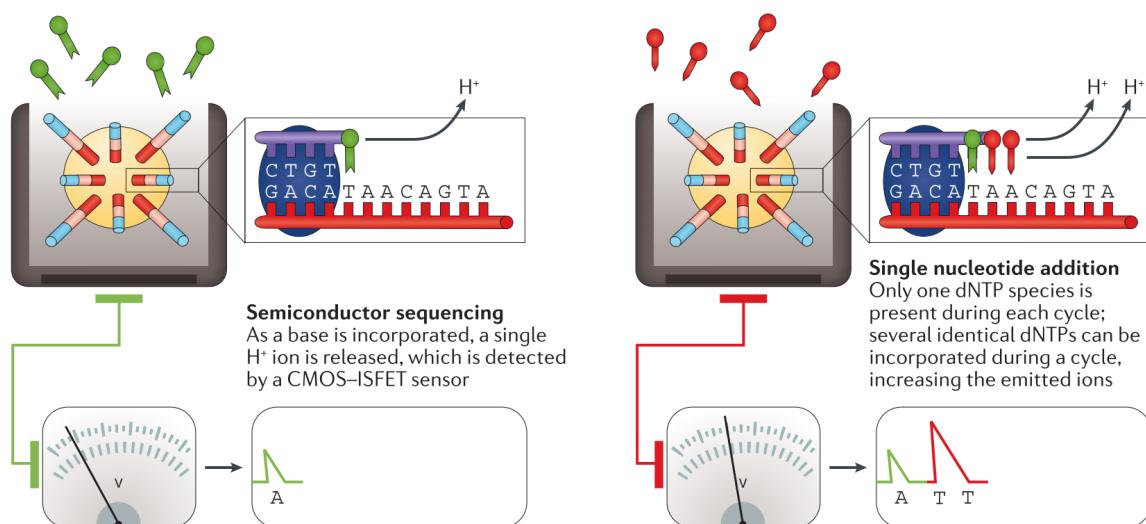


Figure 1.21 – Séquençage SNA tel qu'il est effectué par Ion Torrent d'après [132] : a : mise en présence du patron d'ADN à séquencer avec un mix contenant un seul type de dNTP, si le dNTP est complémentaire au patron, il se fixe et libère un proton permettant d'identifier la liaison. b : dans le cas 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 : 1.22**) : Par définition, cette méthode est basée sur l'hybridation et la ligation de l'ADN à une sonde liée à un fluorophore [134]. 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.

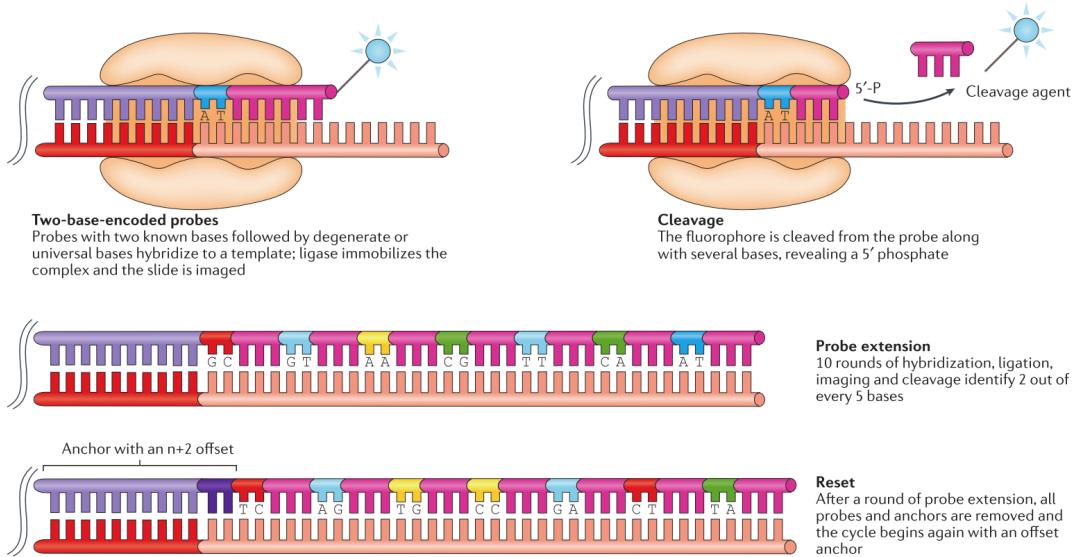


Figure 1.22 – Séquençage SBL tel qu'il est effectué par SOLiD d'après [132] : **a** : dans cette phase d'initialisation, un ensemble de sondes composées de deux nucléotides connus suivis de bases dégénérées et de bases universelles est reliée en 5' à un fluorophore. De fait qu'il n'y ait que quatre fluorophores 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.

1.6 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 [135, 136]. Ces techniques ont été appliquées dans différents contextes, notamment la métagénomique [137], la détection de SNPs [138] et de variants structuraux [139, 140] mais également dans des études portant sur la méthylation de l'ADN [141], l'analyse de l'expression des ARNs messagers [142], dans la génétique du cancer [143] et la médecine personnalisée [144]. Cependant, pour l'ensemble de ces applications, la grande quantité de données générées par chaque analyse pose plusieurs défis informatiques [145]. 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 [146, 147]. 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.

1.6.1 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 : 1.18**). 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 [148]. En plus de SNP, ce format permet de mettre en évidence des variants structuraux [149].

Le format FASTQ

Le format FASTQ (**Figure : 1.23**) 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 [150]. 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 [150].

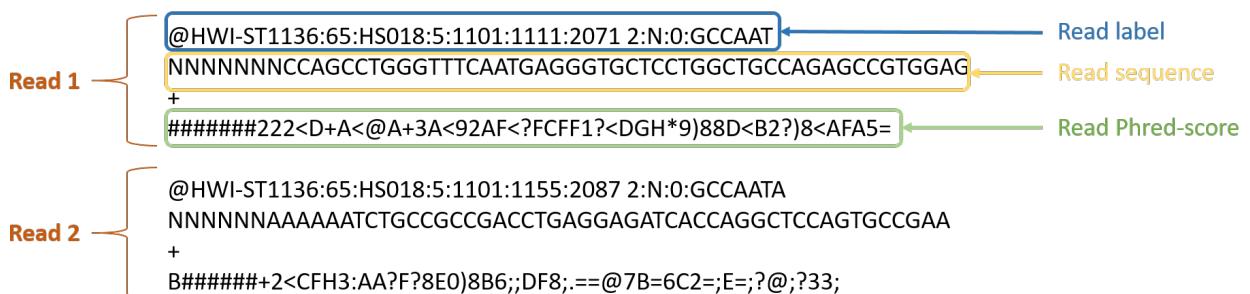


Figure 1.23 – Présentation d'un fichier FASTQ : Chaque *read* présent au sein d'un fichier FASTQ est composé d'un label, d'une séquence et d'un score de qualité associé à chaque nucléotide de la séquence.

1.6.2 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 : 1.24**). Elle constitue probablement l'étape la plus importante de l'analyse des données issues du séquençage haut débit [151] car elle est la base sur laquelle repose l'ensemble des étapes effectuées en aval, notamment l'appel des variants [152]. 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 [153] qui pourront entraîner l'alignement d'un même *read* à plusieurs régions du génome [154]. De nombreux aligneurs ont émergé afin de répondre au mieux à cette problématique tel que Bowtie [155], Bowtie2 [153], BWA [156], NovoAlign, MAGIC [157]. 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 [158–160].

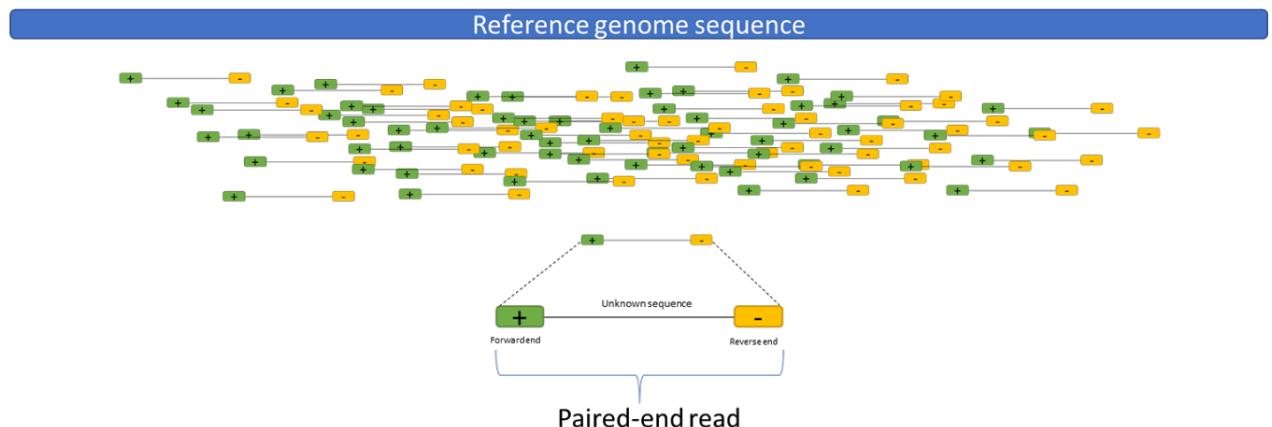
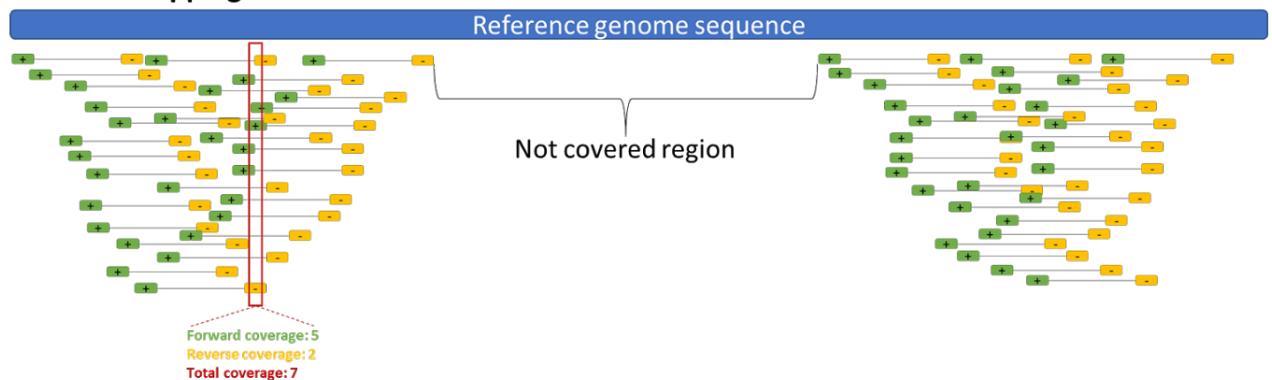
A: Before mapping**B: After mapping**

Figure 1.24 – 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.

1.6.3 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 [152, 161, 162]. 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 1.25 – 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 des hétérozygote. Lorsqu'à une position donnée, peu de *reads* portent un variant, la cause est souvent une erreur de séquençage.

De nombreux logiciels d'appels des variants, ou *caller*, basés sur des algorithmes différents ont émergé ces dernières années pour répondre à cette problématique. Parmi les plus connus on note SAMtools [163], Genome Analysis Tool Kit - HaplotypeCaller (GATK-HC) [164], Freebayes, SOAPindel et TVC. Les quatre premiers cités, peuvent être utilisés pour analyser des données provenant de tout type de plateforme de séquençage contrairement à TVC qui a été développé spécifiquement pour les données provenant de Ion Proton. Les données issues de NGS peuvent présenter un taux d'erreur important. Ce taux d'erreur est multifactoriel et 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 [127]. 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 [165]. 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 : 1.26**) :

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.

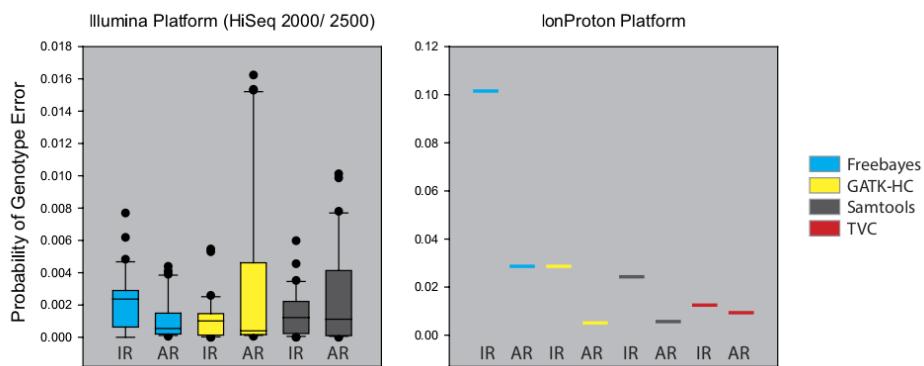


Figure 1.26 – Représentation des erreurs d'appel de type IR et AR en fonction de la plateforme de séquençage et du logiciel d'appel d'après [165] : Pour la plateforme Illumina, on peut voir que Freebayes favorise les appels variant-homozygote tandis que GATK-HC et Samtools favorisent les appels hétérozygotes. 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 [166–168]. 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* [167]. 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* [165].

1.6.4 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 [169]. 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 : 1.28**) :

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 [170]. 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) [171] 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) [172], SnpEff [173] ou encore ANNOVAR

[174] vont prédire l'impact qu'aura un variant sur les différents transcrits qu'il chevauche. D'autres logiciels tel que SIFT [175], PROVEAN [176], Polyphen2, ou encore CADD vont, eux, chercher à prédire la pathogénicité de ce variant, c'est à dire la probabilité que ce variant soit délétère pour la fonction de la protéine. Bien que cette information soit importante, elle est à pondérer, étant donné le peu de concordance qu'il existe entre les prédictions de ces différents logiciels (**Figure : 1.27**).

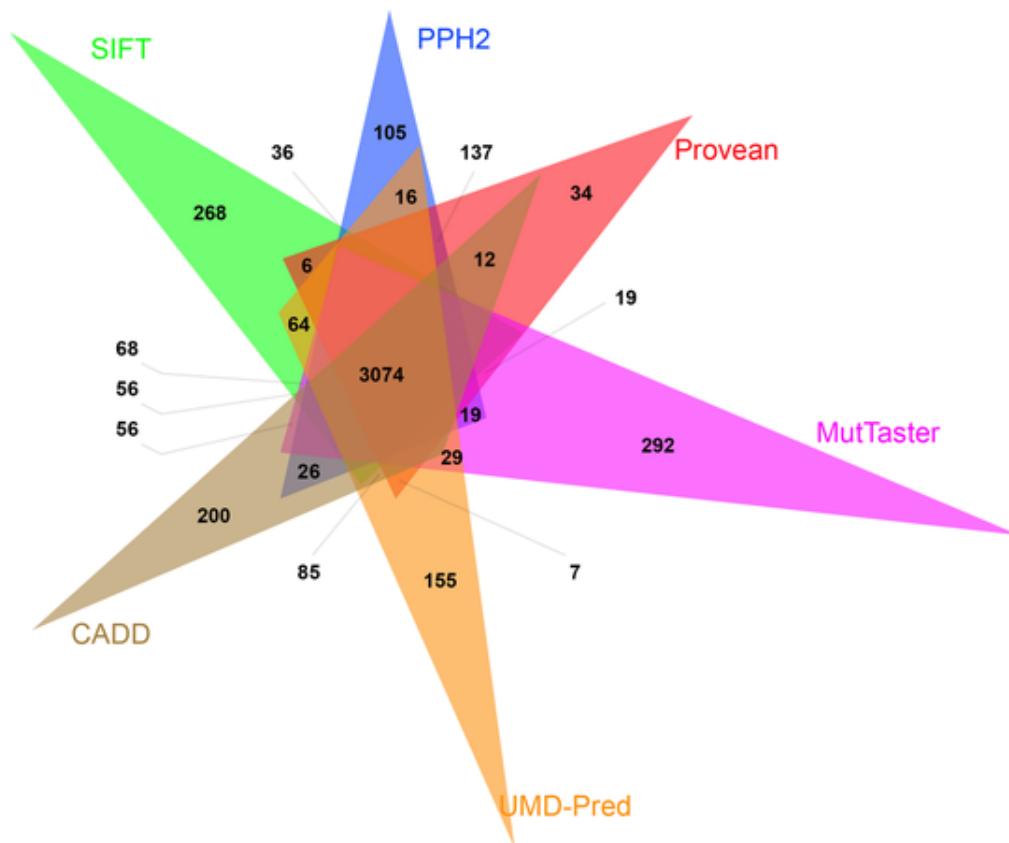


Figure 1.27 – Diagramme de Venn des prédictions de pathogénicité de variants de six logiciels d'après [177] : Les six logiciels utilisés sont : CADD [178] (marron), SIFT [175] (vert), PolyPhen2 [179] (bleu), Provean [176] (rouge) MutationTaster [180] (violet) et UMD-Predictor [181] (orange).

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 A Nalysis THrough Evolutionary Relationships* (PANTHER) [182] 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) [183] 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 [184] ou encore le pLI [171]. 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.

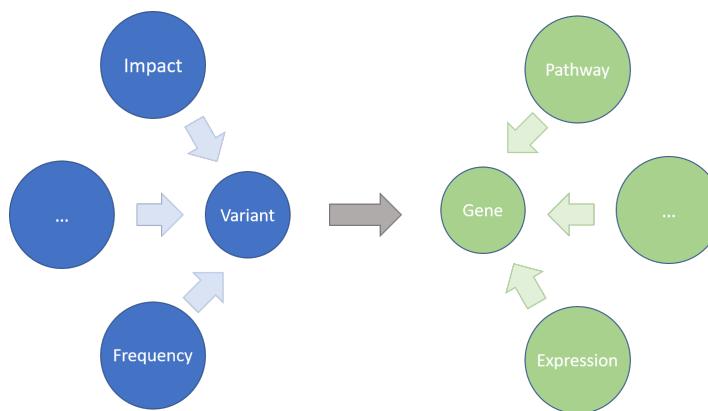


Figure 1.28 – Représentation simplifiée du processus d'annotation : On peut observer deux niveaux d'annotation, le premier est l'annotation des variants consistant à ajouter un maximum d'information sur un variant spécifique (sa fréquence, son impact...). La deuxième est au niveau du gène, consistant 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.

1.6.5 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 variant 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'a aidant 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 [185] bien que les transcrits du *Consensus Coding Sequence project* (CCDS) soient bien représentés par ces trois listes [187]. De même, pour une même liste de gène, de nombreuses différences seront observées en fonction du ou des logiciels de prédition utilisés [177, 185].

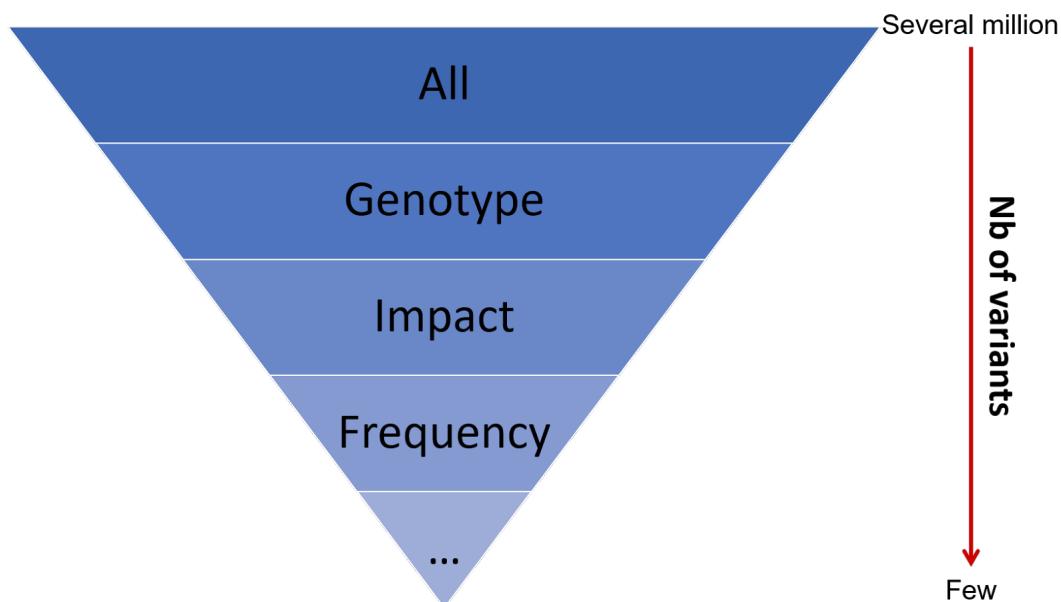


Figure 1.29 – Représentation simplifiée du processus de filtrage des variants : L'ensemble des annotations ajoutées lors de l'étape précédente servent alors de support pour filtrer (ou non) les variants. Il est par exemple commun de filtrer les variants ayant une forte fréquence dans la population générale 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 filtrés.

1.6.6 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" [188]. 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 : 1.30**). 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 [189]. 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 [190]. 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.

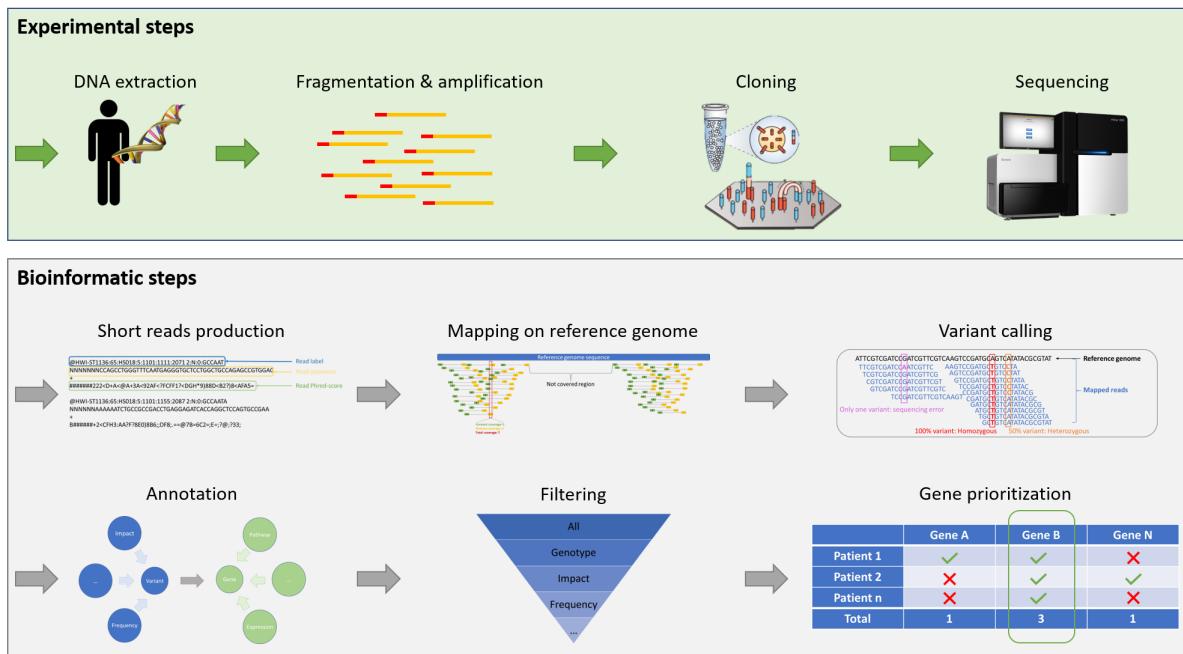


Figure 1.30 – Récapitulatif des différentes étapes du séquençage NGS dans le cadre d'une étude phénotype-génotype : L'ADN est d'abord extrait, puis fragmenté. Les fragments sont liés à des adaptateurs puis amplifiés. Ces amplifiats sont alors isolés et soumis à une amplification clonale (schéma de l'amplification clonale est adapté d'après [132]). 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 variants et leur génotype seront appelés puis annotés. Ces annotations serviront ensuite pour filtrer les variants jugés non pertinent dans le cadre de l'étude, les variants / gènes restant seront ensuite prioriser de sorte à identifier le/les variant(s) responsable(s) du phénotype.

1.7 Problématique : Un patient, 50.000 variants, à la fin il ne peut en rester qu'un. 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 [191].

L'objectif de ma thèse a ainsi été double. Ma première mission a été de mettre en place une stratégie permettant l'analyse des données de séquençage haut débit de patients et patientes souffrant d'infertilité afin de mettre en évidence les variant responsables de leur phénotype. J'ai pour cela développé un pipeline d'analyse des données NGS que j'ai pu tester sur sept cas familiaux d'individus infertiles. J'ai également utilisés ce pipeline sur une cohorte de 15 femmes souffrant d'un syndrome déficience méiotique ovocytaire ainsi que sur une dernière cohorte de 78 patients présentant des anomalies morphologiques multiples au niveau des flagelles spermatiques.

La seconde mission de ma thèse a été de contribuer à la caractérisation moléculaire du gène *DPY19L2* dont la délétion homozygote est responsable du phénotype de globozoospermie, un phénotype rare mais sévère de teratozoospermie menant à l'infertilité masculine.

Au cours de cette thèse j'ai donc eu l'occasion de participer à l'ensemble des analyses permettant l'identification d'un gène comme responsable d'un phénotype en commençant par l'analyse bioinformatique des données de séquençage haut débit en finissant par la caractérisation moléculaire d'un gène.

L'ensemble de ce manuscrit a été rédigé sous l'environnement de développement Rstudio [192] en language 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 script développés pendant ma thèse sont consultables sur la page GitHub : <https://github.com/tkaraouzene/>.

Problématique : Un patient,
50.000 variants, à la fin il ne peut
en rester qu'un. 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 [191].

L'objectif de ma thèse a ainsi été double. Ma première mission a été de mettre en place une stratégie permettant l'analyse des données de séquençage haut débit de patients et patientes souffrant d'infertilité afin de mettre en évidence le variant responsables de leur phénotype. J'ai pour cela développé un pipeline d'analyse des données NGS que j'ai pu tester sur sept cas familiaux d'individus infertiles. J'ai également utilisés ce pipeline sur une cohorte de 15 femmes souffrant d'un syndrome déficience méiotique ovocytaire ainsi que sur une dernière cohorte de 78 patients présentant des anomalies morphologiques multiples au niveau des flagelles spermatiques.

La seconde mission de ma thèse a été de contribuer à la caractérisation moléculaire du gène *DPY19L2* dont la délétion homozygote est responsable du phénotype de globozoospermie, un phénotype rare mais sévère de teratozoospermie menant à l'infertilité masculine.

Au cours de cette thèse j'ai donc eu l'occasion de participer à l'ensemble des analyses permettant l'identification d'un gène comme responsable d'un phénotype en commençant par l'analyse bioinformatique des données de séquençage haut débit en finissant par la caractérisation moléculaire d'un gène.

L'ensemble de ce manuscrit a été rédigé sous l'environnement de développement Rstudio [192] en language 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 script développés pendant ma thèse sont consultables sur la page GitHub : <https://github.com/tkaraouzene/>.

CHAPITRE 2

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

En 2011, les bases moléculaires d'environ 3700 pathologies à transmission mendélienne avaient été élucidées. Cependant, pour une quantité équivalente de pathologies Mendéliennes (ou suspectées de l'être) cette cause reste un mystère [190]. Avec plusieurs centaines de pathologies caractérisées depuis 2010 [193], les séquençages WGS et WES ont, depuis leur émergence, révolutionné les méthodes de recherche dans le cadre d'étude phénotype-génotype en permettant de manière rapide et à moindre coût le séquençage de la quasi-totalité des gènes humains. Dès lors, le défi de ces analyses n'est plus le séquençage de l'ADN mais l'interprétation des données massives produites. En effet, l'un des plus grands challenges des analyses phénotype-génotype réalisées par WES réside dans l'analyse de l'importante quantité de variants portée par chaque individu s'élevant à plusieurs dizaines de milliers lorsque l'on compare avec le génome de référence. Même après avoir retiré les variants retrouvés fréquemment dans la population générale, des méthodes additionnelles sont nécessaires pour prédire, parmi les variants restant, lesquels induisent des conséquences fonctionnelles sérieuses afin de les prioriser [194]. De nombreux logiciels tel que Variant Effect Predictor [172], SnpEff [173] ou encore ANNOVAR [174] permettent d'identifier quels sont les variants qui ont un effet tronquant sur la protéine. Cependant, avec en moyenne 165 variants homozygotes ayant un effet tronquant retrouvés dans chaque exome [194] ces méthodes, bien qu'efficaces sont souvent insuffisantes.

D'autres logiciels tel que Exomiser [195] vont, à partir d'une liste de variants déjà appelés, effectuer les étapes d'annotation, de filtrage et de priorisation. Malgré l'efficacité de ces logiciels, aucun d'entre eux ne couvre l'ensemble des étapes allant de l'alignement des *reads* à la priorisation des variants, la plupart ayant pour point de départ une liste de variants appelés en amont. Ils ne contrôlent donc en aucune manière les étapes d'alignement et d'appel des variants. Or, comme il a été dit plus tôt, ces deux étapes constituent la base de l'analyse. Ce chapitre décrit à la fois la constitution d'un pipeline d'analyse des données de séquençage exomique recouvrant l'ensemble des étapes allant de l'alignement des séquences à la priorisation des variants ainsi que son utilisation dans le cadre de la recherche de mutations entraînant différents phénotypes d'infertilité d'une part au sein de cas familiaux et, d'autre part, au sein d'une large cohorte d'individus non apparentés présentant tous un phénotype MMAF. L'ensemble des scripts de ce pipeline ont été écrits en Perl v5.10.1 et sont récupérables à l'adresse suivante : <https://github.com/tkaraouzene/ExSQLibur>.

Les graphiques ont été réalisés en R v3.4.1 [196] notamment grâce aux packages ggplot2 [197] et cowplot <https://CRAN.R-project.org/package=cowplot>.

2.1 Méthode : Description du pipeline

2.1.1 L'alignement des *reads*

Comme expliqué plus tôt, l'étape d'alignement a pour objectif de repositionner l'ensemble des *reads* d'un individu le long d'un génome de référence. Cette étape peut ainsi être comparée à la reconstruction d'un puzzle dans lequel chaque *read* peut être assimilé à l'une des pièces tandis que le génome de référence serait ici le modèle (**Figure : 1.24**).

L'ensemble de nos exomes ayant été réalisé en *paired-end*, les deux extrémités de chaque fragment sont séquencées. Chaque *end* d'un même *read* peut donc être considérée comme un *read* à part entière qui est aligné **indépendamment** le long du génome de référence ; l'information fournie par le *paired-end* n'étant utilisée qu'*a posteriori* en tant que critère qualité. Au sein de notre pipeline, cette étape est effectuée par le logiciel MAGIC [157], qui, dans le cadre de nos études, s'est basé sur la version hg19 / GHRC37 du génome de référence. Suite à cet alignement, plusieurs critères sont observés afin de filtrer les *reads* présentant une faible qualité d'alignement.

Ainsi, le premier de ces filtres consiste à tout d'abord filtrer l'ensemble des *reads* dupliqués, c'est à dire les *reads* ayant des séquences parfaitement identiques, ceux-ci étant souvent le résultat d'un excès d'amplification au moment des PCRs effectuées en amont. De la même manière, afin d'éviter toute ambiguïté au moment de l'interprétation des résultats, l'ensemble des *reads* s'étant aligné sur plusieurs régions du génome est aussi filtré. Une fois cela fait, nous vérifions la "compatibilité" des deux *ends* composant chacun des *reads* restant. Un *read* est dit compatible lorsque les deux *ends* qui le composent s'alignent face à face (une sur le brin sens du génome de référence et l'autre sur le brin anti-sens) et couvrent une zone ne faisant pas plus de 3 fois la taille médiane de l'insert. Les *reads* dont les deux *ends* se sont alignées mais ne remplissant pas ces conditions seront dits "non compatibles", ceux dont une seule des deux *ends* s'est alignée seront appelés "orphelins" et enfin ceux pour lesquels aucune des deux *ends* ne se sont alignées sont appelés "non-alignés". L'ensemble des *reads* "non-compatibles", "orphelins" et "non-alignés" sont, en raison de leur faible qualité, filtrés et donc non considérés pour les analyses en aval. Les *reads* ayant passé l'ensemble des critères qualité mentionnés précédemment seront, eux, utilisés pour effectuer l'appel des variants.

2.1.2 L'appel des variants

Si l'alignement des séquences peut être comparé à la reconstruction d'un puzzle, l'appel des variants pourrait lui être vu comme un jeu des sept erreurs, au cours duquel, pour chaque position couverte, les différences entre la séquence de l'individu séquencé et le génome de référence seront listées et appelées variants.

Comme nous l'avons vu plus ci-dessus, il est fortement conseillé d'effectuer l'appel des variants en tenant compte de l'aligneur choisi [152, 161, 162]. C'est pourquoi, nous avons développé notre propre algorithme d'appel des variants spécialement conçu pour l'analyse des données de MAGIC. Ainsi, l'appel des variants sera directement basé sur quatre comptages (R_+ , R_- , V_+ et V_-) fournis directement par MAGIC pour chaque position suffisamment couverte :

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

Ainsi, les sommes : $R_+ + V_+$ et $R_- + V_-$ indiqueront respectivement la couverture d'une position en ne tenant que des *reads forward* et *reverse*. En fonction de ces couvertures, nos appels seront classés en trois catégories :

1. **Les appels *double strand* (DS)** : ils qualifient les positions ayant une couverture ≥ 10 sur **les deux strands**. Ces appels sont ceux ayant la meilleure qualité. Le choix de la valeur 10 comme critère de couverture minimum est basée sur l'analyse de nos données et sera explicité plus tard dans le manuscrit.
2. **Les appels *single strand* (SS)** : ces appels définissent les positions pour lesquelles **un des deux strands** présente une couverture ≤ 10 . Dans ce cas, ce *strand* est ignoré et l'appel est effectué uniquement en utilisant le second *strand*.
3. **Les appels *non strand* (NS)** : les positions NS sont celles pour lesquelles la couverture est ≤ 10 sur **les deux strands**. Aucun appel n'est effectué à ces positions qui **ne sont pas conservées dans la suite des analyses**.

Ensuite, pour chaque position couverte, des appels indépendants seront effectués pour chaque *strand* de telle sorte que, pour chacune de ces positions si :

1. 0 à 20% des *reads* portent un variant, la position est appelée **homozygote référence**.
2. 20 à 40% des *reads* portent un variant, l'appel sera considéré comme **ambigu bas**. Cette région est elle-même subdivisée en deux sous-région, 20 à 30% et 30 à 40%.

3. 40 à 75% des *reads* portent un variant, la position est appelée **hétérozygote**.
4. 75 à 85% des *reads* portent un variant, l'appel sera considéré comme **ambigu haut**. Cette région est elle-même subdivisée en deux sous-région, 75 à 80% et 80 à 85%.
5. 85 à 100% des *reads* portent un variant, la position est appelée **homozygote variant**.

Pour les positions DS, la concordance des appels fournis par chaque *end* est ensuite vérifiée. Cette vérification de la concordance des appels entre les *reads forward* et *reverse* a pour principal intérêt de filtrer les erreurs systématiques pouvant survenir lors du processus de séquençage. Par exemple, les séquenceurs Illumina vont avoir tendance à “se tromper” à la position T des motifs GGT [198]. De fait, cette erreur devrait *a priori* se produire uniquement lors du séquençage dans un seul des deux sens, celui contenant ce motif. Les *reads* alignés sur le brin complémentaire contiendront dès lors la séquence correcte. C'est pourquoi, afin de limiter les erreurs d'appels au maximum, nous effectuons, pour chaque position DS, des appels indépendants sur les deux sens. Ainsi, un variant sera considéré (**Figure : 2.1**) :

1. **Homozygote référence**, si les deux appels sont homozygotes références, ou, un des appels est homozygote référence et l'autre se situe dans la sous-région 20-30% de la région ambigu bas.
2. **Hétérozygote**, si les deux appels sont hétérozygotes, ou, si l'un des appels est hétérozygote et l'autre se situe dans la sous-région 30-40% de la région ambigu bas ou bien dans la sous-région 75-80% de la région ambigu haut.
3. **Homozygote variant**, si les deux appels sont homozygotes variants, ou, un des appels est homozygote variant et l'autre dans la sous-région 75-85% de la région ambigu haut
4. **Ambigu**, si les deux appels sont ambigus bas ou s'ils sont tous les deux ambigus haut.
5. **Discordant**, pour toutes les combinaisons restantes.

Pour les positions SS, l'appel final correspondra directement à l'appel effectué sur l'unique *strand* suffisamment couvert. Ces variants, bien que conservés seront, en raison des erreurs dont ils peuvent être la source, considérés comme de faible qualité. Les appels ambigus et discordants seront, eux, filtrés.

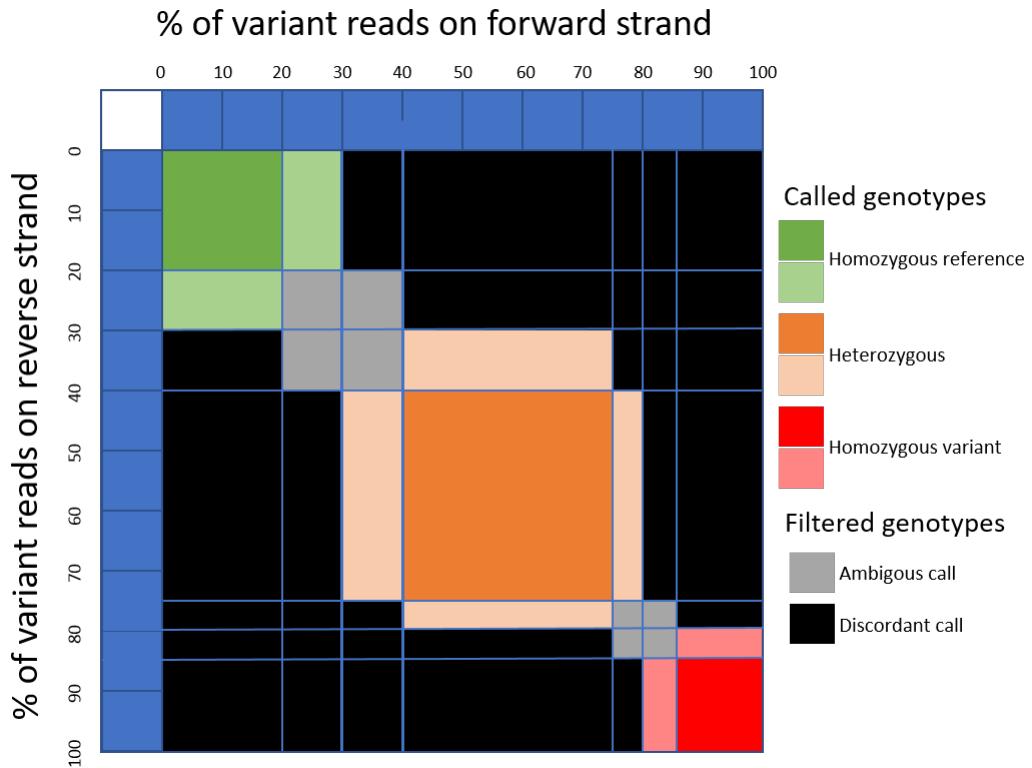


Figure 2.1 – Détail de l'appel général effectué pour les appels DS : Illustration de l'appel des génotypes effectué en fonction du pourcentage de *reads* variants observés sur chacun des deux *strands* de séquençage (*forward* et *reverse*). Le génotype général est appelé si les génotypes des deux brins sont les mêmes ou si l'un des deux est dans la zone ambiguë adjacente au premier. Les zones vertes correspondent aux appels homozygotes références, les zones orange, hétérozygotes et rouges homozygotes variants. Les zones grises sont les appels ambigus tandis que les noires sont les appels discordants. Ces deux derniers appels ne sont pas conservés dans la suite des analyses.

2.1.3 L'annotation

Chaque variant retenu sera ensuite annoté par le logiciel *Variant Effect Predictor* (VEP) [172] qui nous indiquera pour chaque variant la conséquence que celui-ci aura sur la séquence codante de l'ensemble des transcrits Ensembl qu'il chevauche (**Figure : 2.2, Table : 2.1**). Dans le cas d'une substitution faux-sens, c'est à dire entraînant le changement d'un seul acide-aminoé de la séquence protéique, nous utilisons les prédictions fournies par SIFT et PolyPhen afin d'estimer leur pathogénicité. Ensuite, nous ajoutons, pour chaque gène, son expression tissulaire en nous basant sur les données Ensembl [199] générées par le projet Illumina BodyMap qui recense les données RNAseq des gènes humains pour 16 tissus différents. Puis nous ajoutons, lorsque celle-ci est disponible, la fréquence du variant dans les bases de données ExAC [171], ESP600 (Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA) et 1000Genomes [170] donnant ainsi une estimation de sa fréquence dans la population générale. De même, la particularité de ce pipeline est que chaque variant qu'elle a identifié alimente une base de données interne pouvant par la suite servir de contrôle lors de l'analyse d'individus présentant un phénotype différent de ceux étudiés précédemment. L'intérêt d'une telle base, par rapport aux bases de données tel que ExAC, est qu'elle permet d'utiliser, comme contrôle, des individus ayant subi le même protocole de séquençage et la même analyse bio-informatique, permettant ainsi de mieux identifier, et donc filtrer, les erreurs systématiques pouvant arriver à chacune des étapes.

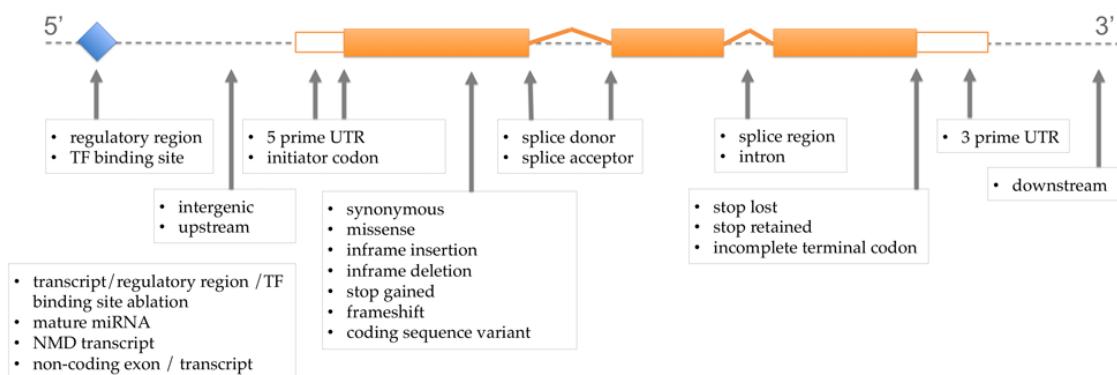


Figure 2.2 – Listes des différentes conséquences prédites par VEP et leur positionnement sur le transcrit d'après : Variant Effect Predictor web site

Table 2.1 – Liste simplifiée des conséquences prédites par VEP avec leur description et impact associée

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

2.1.4 Le filtrage des variants

L'étape de filtrage est primordiale si l'on souhaite analyser de manière efficace les données provenant de WES. C'est pourquoi, elle occupe une place importante dans notre pipeline. L'intégralité des paramètres de cette étape peut être modifiée par l'utilisateur pour faire correspondre les critères de filtres aux besoins de l'étude. Afin de rendre son utilisation la plus efficace possible, nous avons souhaité définir des paramètres par défauts pertinents dans la plupart des études de séquençage exomique de sorte que, sauf si le contraire est spécifié, les filtres suivants seront appliqués :

1. **Filtre 1 : L'union des variants** : quand des individus présentent un lien de parenté et présentent le même phénotype, seuls les variants observés chez l'ensemble des individus sont conservés.
2. **Filtre 2 : Génotype des variants** : ce pipeline d'analyse a été développé, avant tout, pour la recherche de variants impliqués dans des pathologies à transmission récessive. C'est pourquoi, dans le cadre d'étude d'individus présentant un historique de consanguinité, l'ensemble des variants hétérozygotes sont filtrés. En revanche, dans le cas d'individus issus d'unions non consanguines nous procédons à la recherche de variants hétérozygotes composites, c'est à dire **au moins deux variants hétérozygotes différents situés sur chacun des deux allèles du même gène d'un patient**. Dès lors, bien que les variants soient différents, les deux allèles sont altérés rendant possible l'apparition de phénotype récessif. Malheureusement, dans le cadre des séquençages WES et WGS, il est impossible de connaître le "phasage" de ces variants, c'est à dire que l'on ne peut déterminer si deux variants hétérozygotes sont situés sur le même allèle ou sur deux allèles différents (**Figure : 2.3**). Pour cela, une analyse familiale permettant de suivre la ségrégation des variants est nécessaire.
3. **Filtre 3 : Les transcrits "non pertinents"** : au cours de nos analyses, nous nous sommes concentrés uniquement sur les transcrits codants pour une protéine. Ainsi, l'ensemble des transcrits annotés comme étant non codants ont été filtrés tout comme ceux annotés comme étant NMD (*nonsense-mediated decay*). En effet, ce mécanisme a pour but de contrôler la qualité des ARNm cellulaires chez les eucaryotes [200] en éliminant les ARNm qui comportent un codon stop prématué [201] pouvant être le résultat d'une erreur de transcription, d'une mutation ou encore d'une erreur d'épissage. Il est donc peu probable que les variants présents sur des transcrits annotés NMD soient responsables du phénotype. Dès lors, ces transcrits ont été également filtrés. Ainsi, l'ensemble des variants impactant **uniquement** des transcrits non codants et / ou annotés NMD sont filtrés.
4. **Filtre 4 : Impact du variant** : afin de ne conserver que les variants ayant le plus de risque d'avoir un effet délétère sur la protéine, seuls sont conservés ceux impactant la séquence codante d'un transcript. De plus, les variants synonymes ne sont pas conservés (excepté ceux se trouvant proches des régions d'épissage) car ceux-ci n'ont aucun effet sur la séquence protéique. Pour les variants faux-sens

(changement d'un seul acide-amidé de la séquence protéique) il est plus difficile de trancher, dès lors, seuls ceux étant prédis comme *tolerated* par SIFT [175] et comme *benign* par Polyphen [179] sont filtrés.

5. **Filtre 5 : Fréquence des variants** : la fréquence d'un variant dans la population générale est un moyen rapide d'avoir une prédiction fiable de l'effet délétère ou non de celui-ci. En effet, il est peu probable qu'un variant retrouvé fréquemment dans la population générale soit causal d'une pathologie sévère. C'est pourquoi, l'ensemble des variants ayant une fréquence $\geq 1\%$ dans l'une des trois bases de données que sont ExAC, ESP et 1KG est filtré.
6. **Filtre 6 : Présence des variants dans la cohorte contrôle** : le filtre utilise les variants composant la base de données interne du pipeline et permet de filtrer l'intégralité des variants homozygotes retrouvés chez les patients séquencés ne présentant **pas** le même phénotype que le patient analysé. Comme dit plus tôt, ce filtre se révèle particulièrement intéressant lorsque plusieurs patients porteurs de phénotypes différents ont subi le même protocole de séquençage. Ainsi l'ensemble des variants faux-positifs résultant d'artéfacts liés aux différentes étapes en amont de l'analyse bio-informatique pourra alors être filtré. De même ce filtre permet de mettre en évidence les variants propres à une population lorsque des patients provenant de la même région géographique et ne présentant toujours pas le même phénotype sont comparés.

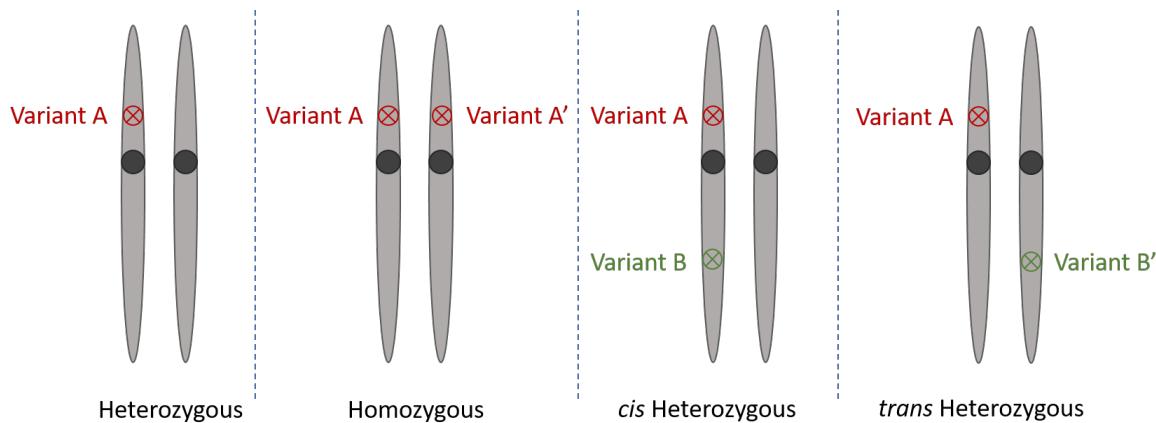


Figure 2.3 – Représentation schématique des phasages de deux variants avec les génotypes associés : Un variant est homozygote lorsque le **même** variant est présent sur les deux allèles d'un gène et hétérozygote lorsqu'il est présent sur **un seul** des deux allèles. On parle d'hétérozygotes *cis* lorsque deux variants hétérozygotes **dif- férents** sont positionnés sur **le même allèle** et d'hétérozygote *trans* (ou composite) lorsque ces deux variants hétérozygotes sont positionnés sur **deux allèles différents**. En WES et en WGS il est impossible de différencier les hétérozygotes *cis* des hétérozygotes *trans*.

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

Cette partie se concentre sur l'analyse bio-informatique des résultats des séquençages exomiques de 13 individus infertiles provenant de 7 familles. Pour 6 familles, 2 frères atteints ont été analysés et pour la septième, un seul des frères atteint a été séquencé et l'ADN du deuxième frère a été disponible *a posteriori* 2.2.

1. **Famille FAM** : cette famille est composée de 2 frères azoospermes. Comme nous avons pu le voir, l'azoospermie est un phénotype d'infertilité masculine caractérisé par l'absence de spermatozoïdes dans l'éjaculat. Des 13 patients de cette étude, les frères Ghs44 et Ghs45 sont les deux seuls à ne pas avoir été séquencés au Génopole d'Évry.
2. **Famille FF** : les spermatozoïdes des 2 frères de cette famille sont caractérisés par leur incapacité à féconder l'ovocyte malgré leur morphologie et leur mobilité normales.
3. **Famille MMAF1-5** : ici nous avons 5 familles dont l'ensemble des membres séquencés présentent un phénotype MMAF (*multiple morphological abnormalities of the sperm flagella*). Ce syndrome se caractérise par la présence d'une majorité de spermatozoïdes présentant une mosaïque d'anomalies morphologiques du flagelle.

Table 2.2 – Tableau récapitulatif des familles séquencées et de leur phénotype

Family	Individuals	Phenotype	Year	Place
AZ	Ghs44, Ghs45	Azoospermia	2012	Mount Sinai Institut
FF	Ghs113, Ghs117	Fertilization failure	2014	Genoscope (Evry)
MMAF1	Ghs56, Ghs58	MMAF	2014	Genoscope (Evry)
MMAF2	Ghs59, Ghs60	MMAF	2014	Genoscope (Evry)
MMAF3	Ghs62, Ghs130	MMAF	2014	Genoscope (Evry)
MMAF4	Ghs119, Ghs63	MMAF	2014	Genoscope (Evry)
MMAF5	Ghs131	MMAF	2014	Genoscope (Evry)

2.2.1 Résultats des différentes étapes de l'analyse

Résultat de l'alignement

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

La quantité de *reads* composant les exomes de chaque individu peut varier en fonction de plusieurs paramètres et n'est donc pas égale pour chaque patient bien que l'ordre de grandeur reste le même avec une médiane de 87747080 *reads*. Seuls les deux frères *Ghs44* et *Ghs45* de la famille AZ se distinguent avec près de 3 fois plus de *reads* que les autres patients. Cette différence peut être expliquée car ces deux patients sont les deux seuls à avoir été séquencés au Mount Sinaï Institut. Or leur protocole d'amplification précédant le séquençage contient un nombre de cycles de PCR supérieur à ceux appliqués au Génopole d'Évry où ont été séquencés les autres patients. Il faut noter que ce nombre plus important de *reads* n'est en rien le reflet d'une meilleure qualité. En effet, celui-ci est causé par une grande quantité de *reads* dupliqués qui seront pour la plupart filtrés au cours des analyses ultérieures (**Table : 2.2, Figure : 2.4 - A**).

La première étape du contrôle qualité des *reads* consiste à filtrer ceux ne s'étant pas alignés sur le génome. Ces *reads* sont extrêmement minoritaires puisqu'ils ne représentent qu'entre 1.2 et 5.5 % des *reads* de nos individus (**Figure : 2.4 - B**).

Parmi les *reads* s'étant correctement alignés sur le génome, seuls les *reads* présentant des *ends* compatibles sont conservés. Pour rappel, ces *reads* représentent ceux pour lesquels une des deux *ends* s'est alignée sur le *strand forward* tandis que l'autre s'est alignée sur le *strand reverse* et que la distance qui les sépare n'est pas supérieure à trois fois celle de l'insert. Dans nos données, les *reads* remplissant ces conditions sont majoritaires puisqu'ils représentent environ 91.3 % des *reads* s'étant correctement alignés tandis que les *reads* non-compatibles et orphelins représentent respectivement 6.2 et 2.8 % de ces mêmes *reads* (**Figure : 2.4 - C**).

La dernière étape de ce contrôle qualité consiste à analyser le nombre de sites sur lesquels se sont alignés les *reads*. En effet, certaines zones du génome étant dupliquées, l'une des problématiques des *short-reads* est qu'il est possible que ceux-ci s'alignent à plusieurs régions différentes du génome. Afin d'éviter toute ambiguïté, seuls ceux s'étant alignés sur un site unique sont conservés pour la suite des analyses. Ces *reads* représentent entre 92.3 et 96.9 % des *reads* ayant passé les précédents filtres (**Figure : 2.4 - D**).

Ainsi, à la fin de ces trois étapes de contrôle qualité des *reads*, environ 84.8 % d'entre eux sont conservés soit un total d'environ 74409523 *reads* par patients.

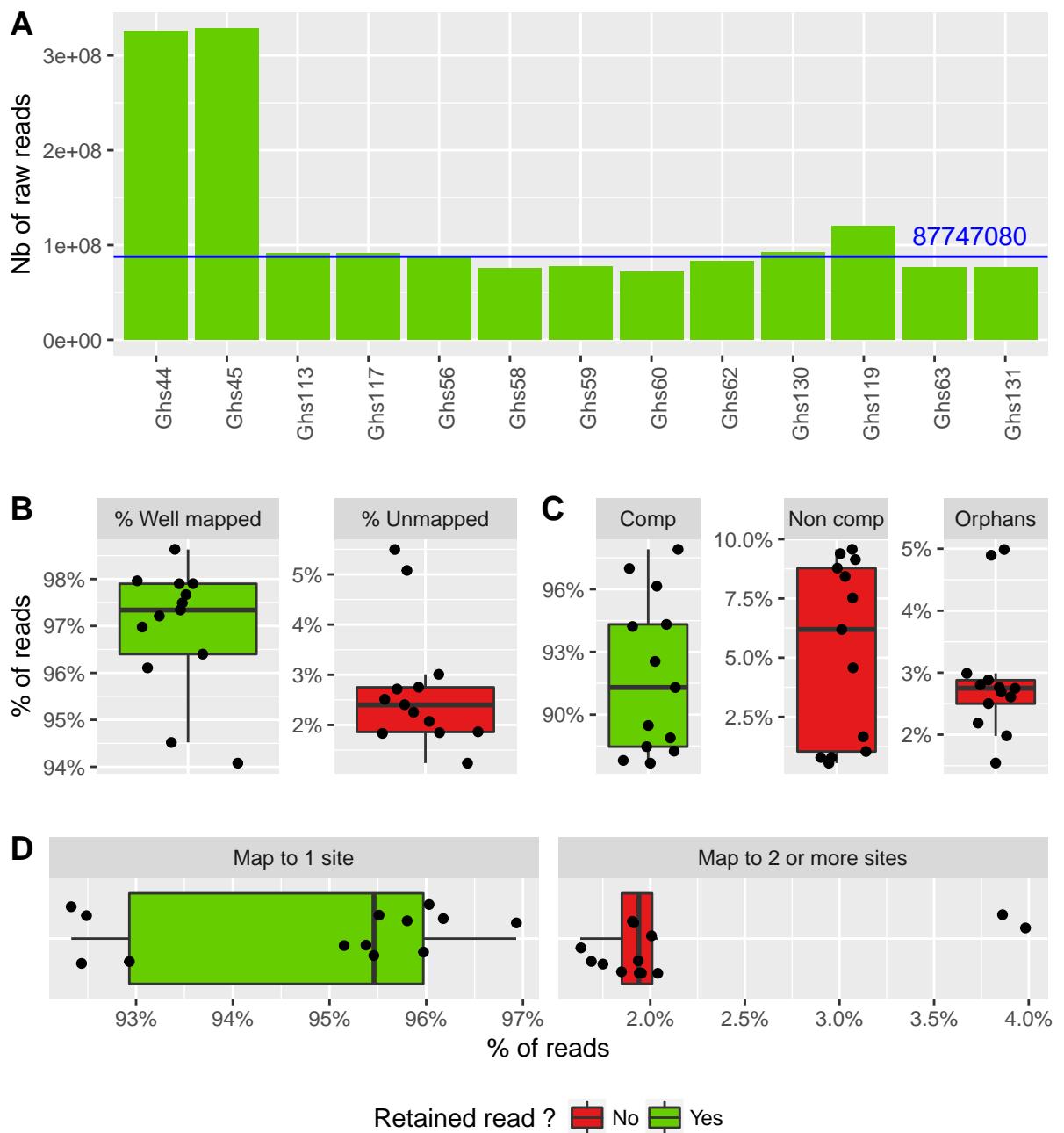


Figure 2.4 – Processus simplifié du contrôle qualité des reads :
Pour chacun des graphiques, les *reads* représentés en vert sont conservés tandis que ceux en rouge sont filtrés. **A** : Quantité de *reads* bruts générés pour chaque patient au cours de l'étape de séquençage. La médiane des *reads* est représentée en bleue. **B** : Pourcentage, pour chaque individu, de *reads* s'étant alignés correctement et ne s'étant pas alignés sur le génome de référence. **C** : Distribution pour chaque patient des *reads* compatibles (Comp), non compatibles (Non comp) et orphelins (Orphans). **D** : Présentation pour chaque *read* du nombre de sites auxquels ils s'alignent

L'appel des variants

Une fois l'alignement effectué, il faut identifier les positions présentant des différences avec le génome de référence et leur assigner un génotype. Tout d'abord, afin de conserver uniquement les positions auxquels il sera possible d'effectuer un appels de qualité, il est nécessaire de filtrer les positions ne présentant pas une couverture suffisante.

Pour cela, nous avons fait varier notre critère de filtre de couverture minimum de 1 à 100 afin. Ainsi, nous avons pu quantifier, pour chacune des valeurs de filtre, la quantité d'appels DS, SS et NS d'observer la quantité de position. On peut ainsi constater qu'à partir de 10, la quantité d'appel NS augmente considérablement, tandis que le nombre d'appel DS et SS chute. Ainsi, au-delà cette valeur, un nombre important de positions ne respectent plus les critères qualité et sont ainsi filtrées (**Figure : 2.5**).

Ensuite, le génotype étant dépendant du pourcentage de *reads* variants à une position donnée, il est nécessaire, afin de calibrer les bornes de notre algorithme d'appel, de connaître la répartition globale de ces proportions. Comme attendu, nous observons des pics à 0, 50 et 100% de *reads* variants pour une position donnée, ces trois pourcentages correspondant respectivement aux appels "homozygote référence", "hétérozygote" et "homozygote variant". La démarcation séparant les appels hétérozygote et homozygote variant est relativement claire. La distinction de ces deux génotypes ne pose donc pas de problème, la région d'ambiguïté allant de 75 à 85% de *reads* variants ne concernant qu'une minorité de positions. Ce n'est cependant pas le cas pour celle séparant les appels homozygote référence et hétérozygote. En effet, on peut constater que la zone d'ambiguïté allant de 20 à 40% de *reads* variants concerne une part non négligeable des positions couvertes. Les positions SS situées dans cette région seront systématiquement filtrées. Pour les positions DS, celles pour lesquelles **un seul** des *strand* se situe dans cette région alors que l'autre se trouve dans une des régions adjacentes (homozygote référence ou hétérozygote) seront conservées. Les autres seront filtrées elles aussi (**Figure : 2.6**).

Dans nos données, les appels SS sont majoritaires et représentent environ 46.7 % de nos appels (contre 37 % d'appels DS). Au vu de l'importance de ces appels, nous avons fait le choix de les conserver afin de ne pas filtrer une quantité trop importante de données. Ces appels seront cependant considérés comme étant de faible qualité, de fait, leurs analyses et interprétation seront plus précautionneuses. En revanche, au vu de la trop grande incertitude de l'appel des variants NS, ceux-ci sont systématiquement filtrés éliminant ainsi entre 10.3 et 23.7 % des positions appelées pour chaque patient (**Figure : 2.7 - A**).

De même, les appels discordant et ambigus sont filtrés, soit environ 13.9 % des variants DS. Il est intéressant de noter que bien que les variants *single strand* (SS) soient conservés, on peut s'attendre à ce qu'environ 8.8%, soit le pourcentage de DS discordants, soient aberrants, ceux-ci n'ayant pu subir le même contrôle que les SS

(Figure : 2.7 - B).

Pour l'ensemble des variants ayant passé les filtres énoncés ci-dessus, c'est à dire les variants SS et les variants DS avec appels concordants, le génotype est déterminé en fonction du pourcentage de *reads* portant le variant à cette position. Ainsi, pour chaque individu nous avons pu établir une liste de SNVs et d'indels avec leur génotype associé. Pour chacun de nos 13 patients, les ordres de grandeur du nombre de variants appelés sont identiques. Ainsi pour chaque patient nous avons appelé environ 43246 variants hétérozygotes (40721 SNVs et 2525 indels) et 34290 variants homozygotes (32497 SNVs et 1793 indels) (Figure : 2.7 - C).

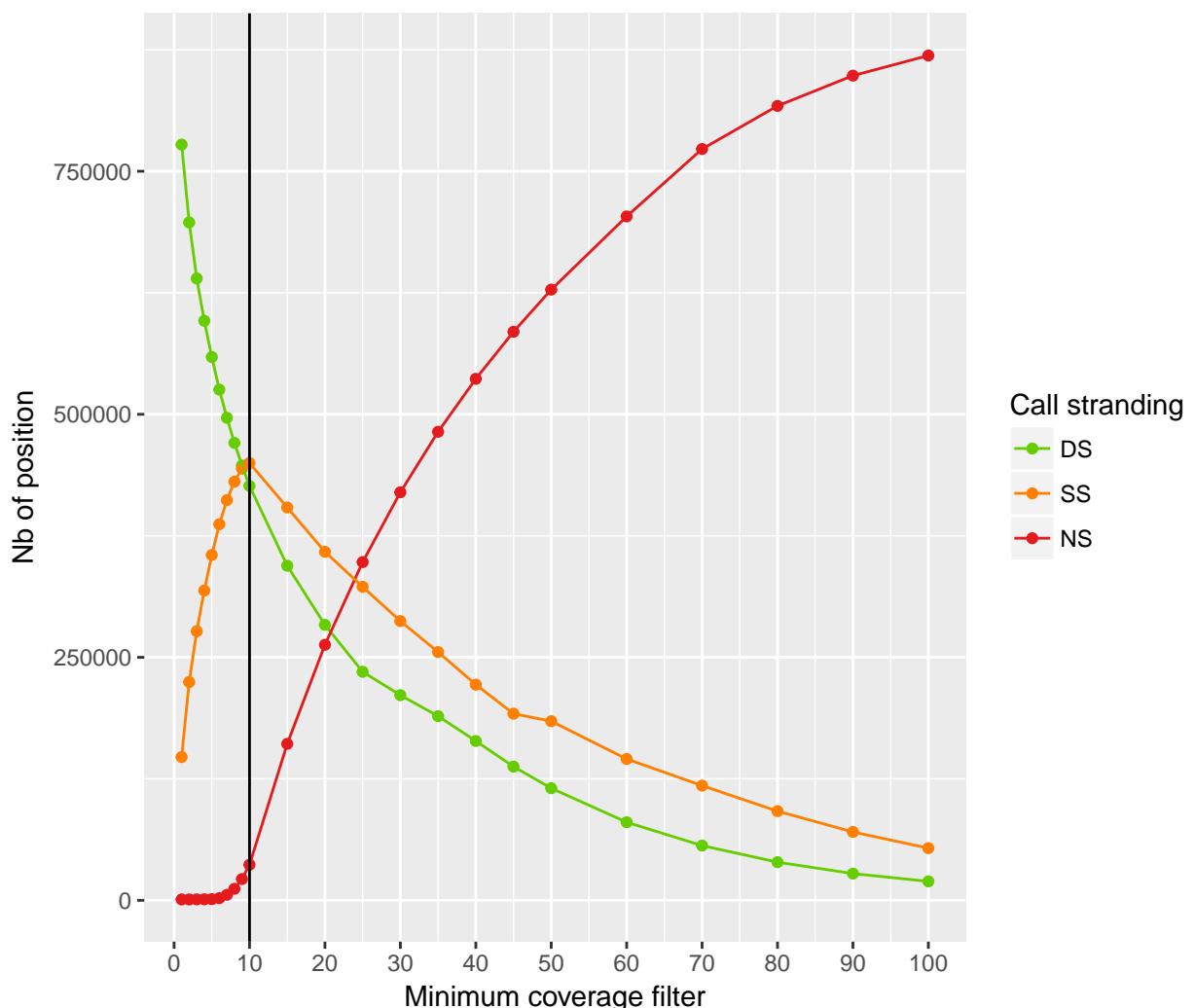


Figure 2.5 – Détermination de la valeur de couverture minimale : Quantification du nombre médian d'appels DS, SS et NS pour chaque valeur de couverture minimum testée. On note une augmentation du nombre d'appel NS à partir de 10 tandis qu'à partir de cette même valeur les appels DS et SS diminuent

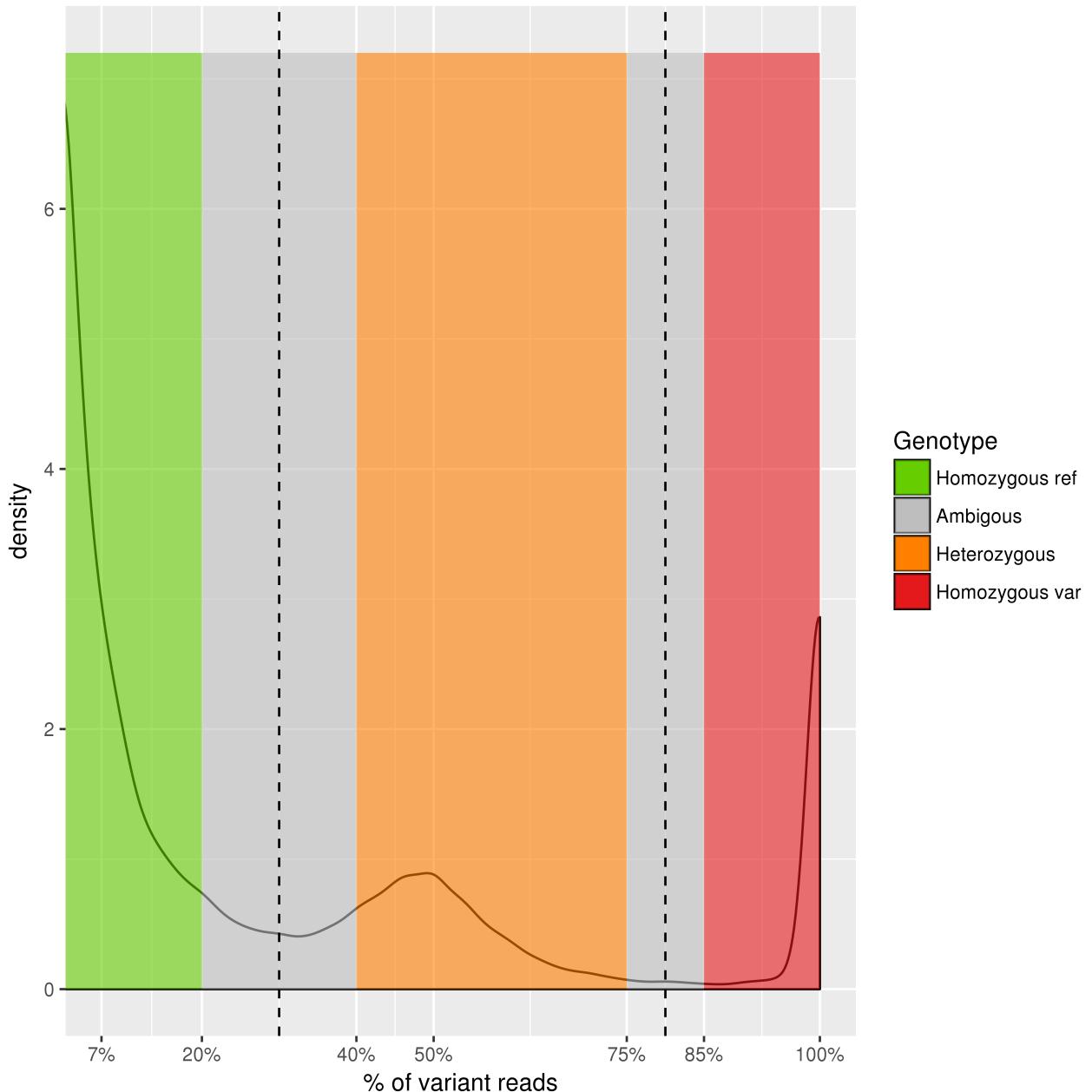


Figure 2.6 – Densité de répartition du pourcentage de reads variants pour chaque position couverte : Quantification du nombre médian d'appels DS, SS et NS pour chaque valeur de couverture minimum testée. On note une augmentation du nombre d'appel NS à partir de 10 tandis qu'à partir de cette même valeur les appels DS et SS diminuent

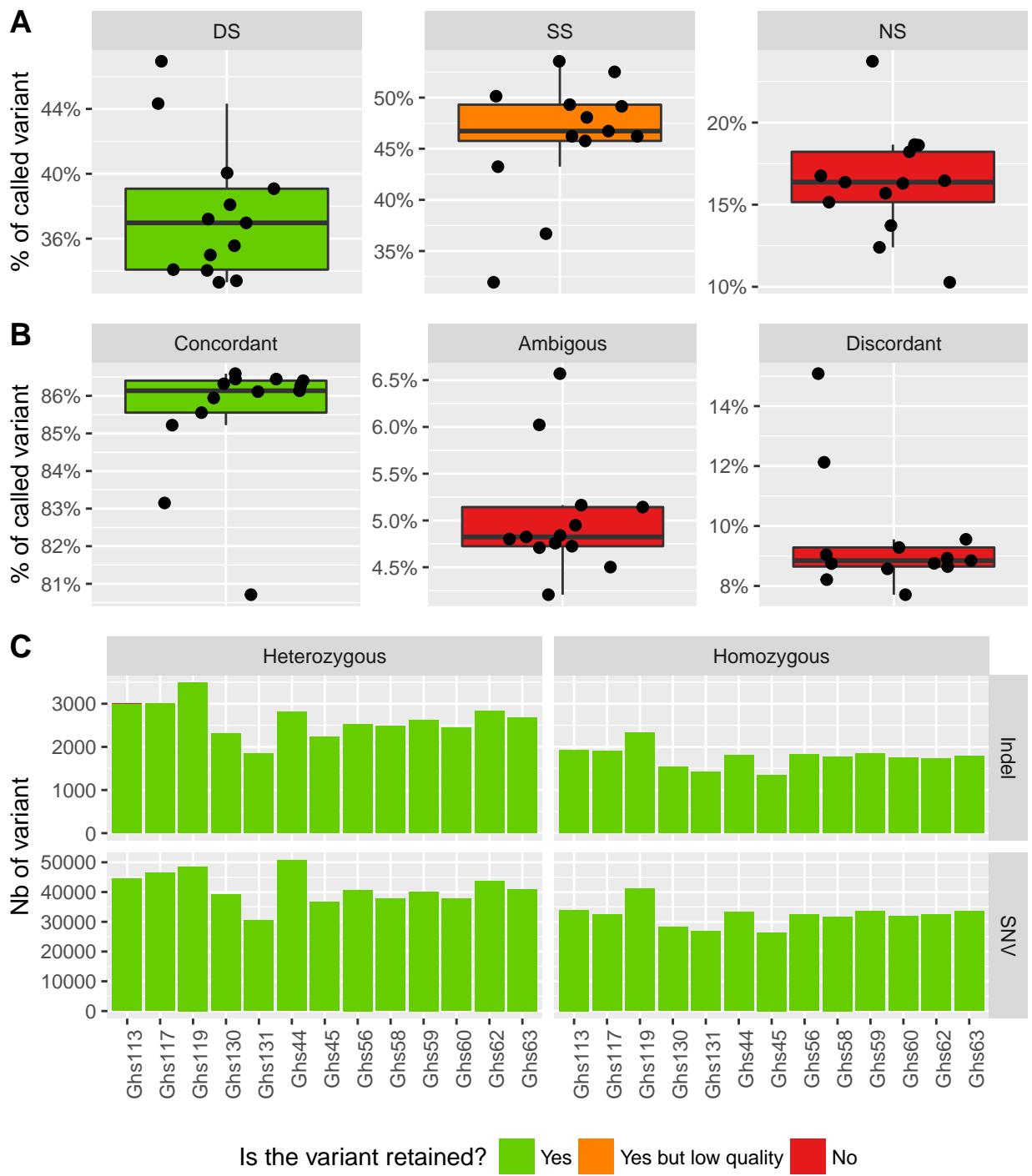


Figure 2.7 – Contrôle qualité des variants appelés : Pour chacun des graphiques, les variants représentés en vert et en orange sont conservés tandis que ceux en rouge sont filtrés. **A** : Distribution du stranding des appels pour chaque patient. **B** : Comparaison des appels entre les deux ends des variants appelés DS. **C** : Distribution des SNVs et indels en fonction de leur génotype pour chaque patient (représentés par une barre).

Nous avons ensuite cherché à comparer les résultats obtenus par notre pipeline avec ceux obtenus par des logiciels déjà existant. Pour cela nous avons aligné les données de nos 13 patients avec le logiciel BWA [156] et effectué l'appel à l'aide de GATK-HC [164]. Ainsi, nous avons pu constater que 78847 variants étaient appelés par notre pipeline avec un génotype hétérozygote ou homozygote tandis que 106701 était appelés par BWA + GATK-HC parmi soit 27854 de plus que par notre pipeline (**Figure : 2.8 - A**).

Cette différence n'est pas surprenante puisque, afin de ne sélectionner que les variants de confiance nous avons appliqué des critères de filtre stringent. On peut ainsi noter que parmi ces 27854 variants, 1644 d'entre-eux sont retrouvés comme discordant par notre pipeline, 6456 sont retrouvés comme ambigus et 15508 comme NS. Ainsi, on peut estimer que 84.8% des variants que nous n'avons pas appelés sont appelés par BWA + GATK-HC avec un génotype soit erroné soit de faible confiance (**Figure : 2.8 - B**).

Néanmoins, en ce qui concerne les variants retrouvés à la fois par le duo BWA + GATK-HC et notre pipeline, 99.63% des appels sont identiques aux deux procédures. (**Table : 2.3**).

Table 2.3 – Comparaison des génotypes des variants appelés par les deux procédures

Magic + in-house algo	BWA + GATK-HC	Count	Percent
Heterozygous	Heterozygous	36225	55.51
Homozygous	Homozygous	28793	44.12
Homozygous	Heterozygous	219	0.34
Heterozygous	Homozygous	18	0.03

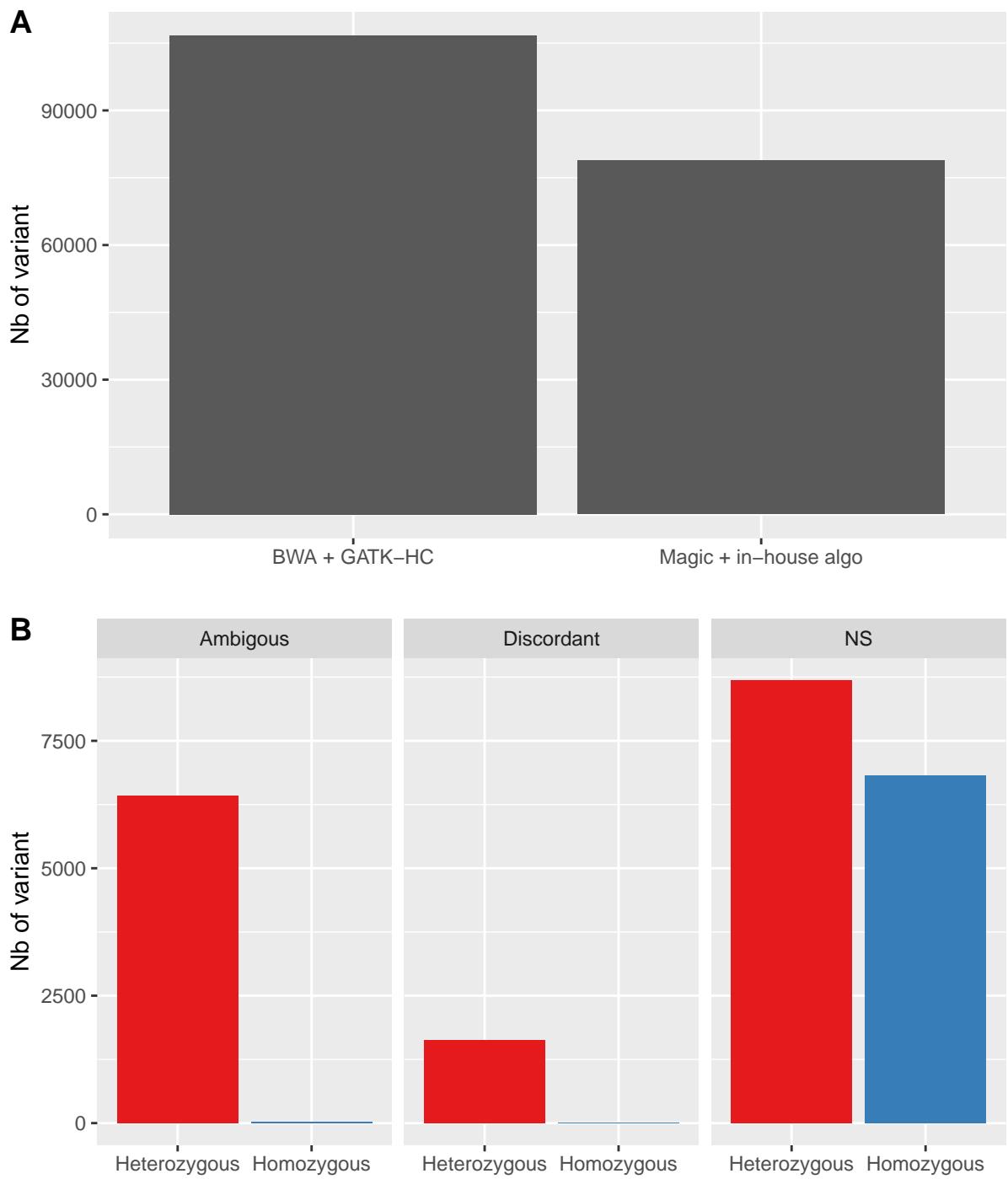


Figure 2.8 – Comparaison des variants obtenus par MAGIC + notre algorithme d'appel et BWA + GATK-HC : A : Nombre de variants hétérozygotes ou homozygotes observé par les deux méthodes. B : Quantification du nombre de variant appellés discordants, ambiguës ou NS par notre méthode retrouvés par BWA + GATK-HC

L'annotation des variants

Après avoir annoté nos variants, nous avons pu constater que pour chaque patient 24794 gènes sont en moyenne affectés par au moins un variant homozygote pour en moyenne 121843 transcrits (soit environ 5 transcrits par gène) (**Figure : 2.9 - A**). Il faut noter que parmi ces gènes se trouvent à la fois des gènes codants pour des protéines et d'autres non codants.

Chaque variant affectera l'ensemble des transcrits qu'il chevauche, ainsi un même variant pourra impacter plusieurs transcrits. Ces impacts sont ensuite classés par VEP en quatre catégories qui sont, de la plus délétère à la moins délétère : *HIGH*, *MODERATE*, *LOW*, *MODIFIER* (**Table : 2.1**).

Comme attendu, les variants ayant un impact tronquant se retrouvent être les moins fréquents chez chacun de nos patients. Ceci est d'autant plus flagrant pour l'impact *HIGH* qui regroupe, entre autres, les variants créant un codon stop ou causant un décalage du cadre de lecture (**Table : 2.1**). Ceux-ci se retrouvent, par rapport aux autres impacts, en quantité extrêmement faible puisqu'ils ne représentent en moyenne que 0.23 % des variants. Cependant, bien que ce pourcentage soit faible, cela représente tout de même une moyenne de 131 variants *HIGH* hétérozygotes par patients et 114 variants *HIGH* homozygotes par patient (**Figure : 2.9 - B**).

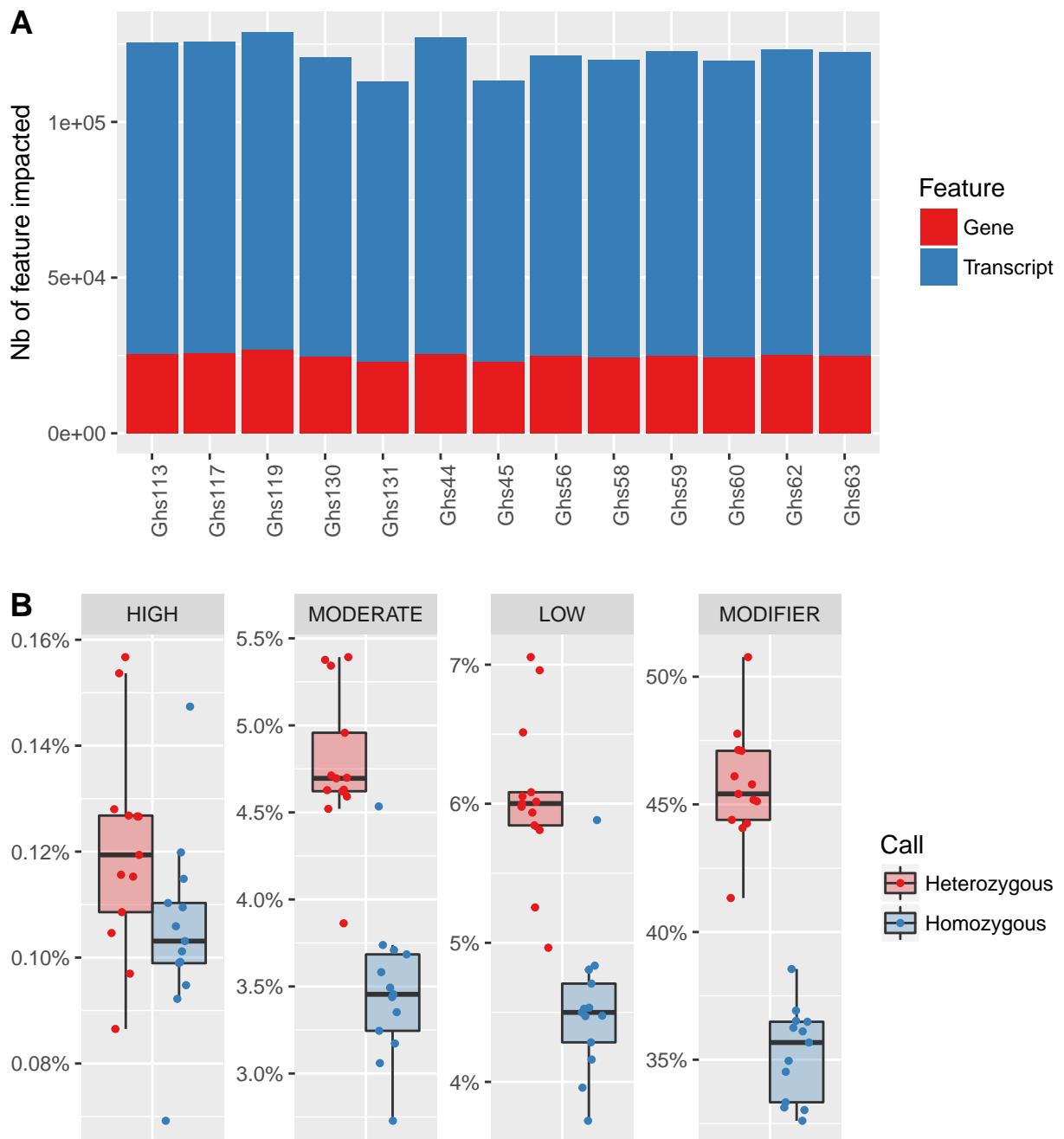


Figure 2.9 – Annotation des variants : A : Quantification du nombre de gènes (en bleu) / transcrits (en rose) impactés par au moins un variant pour chaque patient chacun représenté par une barre. B : Distribution des impacts HIGH MODERATE LOW et MODIFIER en fonction des patients et du statut du variant.

Le filtrage des variants

Les étapes précédentes nous ont permis de mettre en évidence pour chaque patient une liste de variants passant l'ensemble de nos critères qualité. Ces variants ont dès lors pu être annotés nous permettant notamment d'avoir connaissance de leurs impacts sur les différents transcrits qu'ils chevauchent ou encore leur fréquence dans la population générale. Désormais, afin de ne conserver que les variants ayant la plus forte probabilité d'être responsable du phénotype de ces patients, nous avons appliqué successivement les six filtres précédemment décrits.

1. **Filtre 1 : L'union des variants** : dans cette étude nous analysons les données génétiques de 7 familles composées de 1 à 2 frères. Nous avons donc émis l'hypothèse que **le phénotype de chacun des frères d'une même famille était dû à une cause génétique commune**. C'est pourquoi, seul les variants observés chez l'ensemble des membres d'une même famille furent conservés. Ainsi ce filtre a permis de filtrer entre 17703 et 41944 variants pour chacun des patients (**Figure : 2.12 - A**).
2. **Filtre 2 : Génotype des variants** : ici, nous avons émis l'hypothèse d'une transmission récessive du phénotype. Ainsi, seuls les variants homozygotes ont été conservés, filtrant en moyenne 45089 variant par individu soit une moyenne de 56.5% de leurs variants (**Figures : 2.7 - C, 2.12 - A**).
3. **Filtre 3 : Impact du variant** : ce filtre se basant à la fois sur les prédictions VEP mais aussi, dans le cas de variants faux-sens, sur les prédictions SIFT et PolyPhen permet de ne conserver que les variants ayant les effets les plus délétères. Ce filtre est, de prime abord, le plus efficace puisqu'il permet de filtrer à lui seul environ 88.3% des variants de chaque individu.
4. **Filtre 4 : Les transcrits “non pertinents”** : cette étape de filtre permet de filtrer systématiquement entre 13712 et 17992 transcrits différents par patients. Sont considérés comme non-pertinents les transcrits ne codant pas pour une protéine et ceux annotés étant dégradés (*nonsense mediated decay* (NMD)). Cependant, un même variant pouvant impacter à la fois des transcrits “non pertinents” **et** des transcrits “pertinents”, seuls ceux impactant **uniquement** des transcrits “non pertinents” sont filtrés, soit une moyenne de 1797 variants par individu (**Figure : 2.10**).

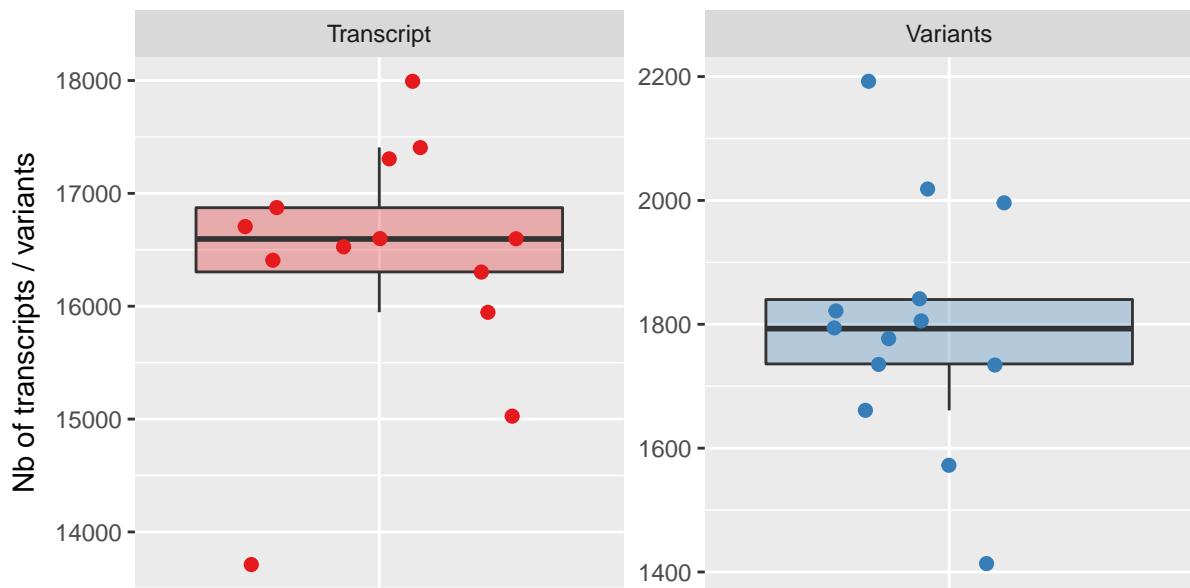


Figure 2.10 – Filtrage des transcrits jugés “non pertinents” et des variants les chevauchant : Pour chaque patient nous avons filtré les transcrits jugés “non pertinents” pour l’analyse, c’est à dire ceux ne codant pas pour une protéine et ceux annoté NMD (boîte rouge). Dès lors, l’intégralité des variants chevauchant **uniquement** des transcrits non pertinents sont systématiquement filtrés (boîtes bleue). Les autres sont conservés.

5. **Filtre 5 : Fréquence des variants :** filtrer systématiquement les variants retrouvés avec une fréquence ≥ 0.01 dans l’une des trois bases de données que sont ExAC, 1KG et ESP6500 permet de filtrer entre 25126 et 29797 variants par patient.
6. **Filtre 6 : Présence des variants dans la cohorte contrôle :** L’ensemble des variants répertoriés au sein de ces 3 phénotypes a été confronté à des listes de variants identifiés chez d’autres patients analysés par notre pipeline et présentant des phénotypes différents. Dès lors, l’ensemble des variants retrouvés à l’état homozygote chez au moins un des individus de la cohorte contrôle sera filtré de notre liste de variants. On peut cependant noter que les variants retrouvés chez les patients *Ghs44* et *Ghs45* de la famille AZ n’ont pas été confrontés à ceux observés dans notre cohorte de 15 femmes présentant des anomalies de développement ovocytaire. En effet, ces deux phénotypes peuvent être causés par un même gène impliqué dans la division méiotique (**Figure : 2.11**).

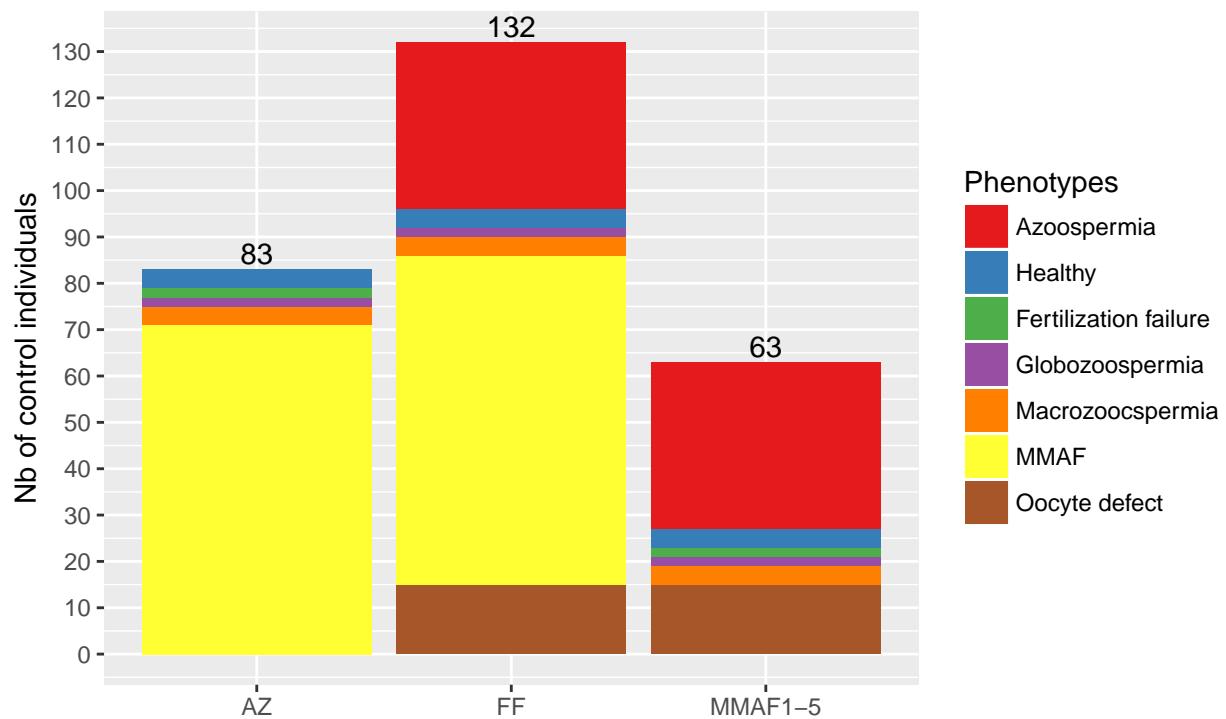


Figure 2.11 – Nombre d’individus et leur phénotypes composant la cohorte contrôle de chaque famille : Ici, chaque barre représente une famille et sa hauteur est déterminée par le nombre d’individus composant la cohorte contrôle à laquelle elle a été confrontée. Le nombre total de contrôles utilisés par famille est inscrit au-dessus de chaque barre. Les couleurs déterminent les phénotypes des individus de la cohorte contrôle. Chaque individu de cette cohorte a été séquencé en WES par notre équipe. Afin d’être considéré comme “contrôle” et intégrer cette cohorte, un individu doit être sain ou présenter un phénotype d’infertilité suffisamment différent de la famille étudiée. Par exemple, un individu MMAF pourra servir de contrôle aux familles AZ et FF mais pas aux familles MMAF1-5.

Comme on pouvait s'y attendre, ces six filtres ont un pouvoir discriminant extrêmement différent. En effet, tandis que le filtre “*Transcript relevance*” (filtre n°4) élimine en moyenne 3.9% des variants de chaque individu, le filtre “*Variant impact*” (filtre n° 3) élimine, lui, jusqu'à 90.1% de ces mêmes variants. Cette différence n'est pas surprenante. En effet, comme nous l'avions vu plus tôt, les variants considérés comme *MODIFIER* par VEP qui regroupent entre autres les variants chevauchant les séquences UTRs et introniques (**Table** : 2.1) représentent en moyenne 81% des variants de nos patients, or, ceux-ci sont tous filtrés. On peut également constater l'importance de la cohorte contrôle puisqu'elle permet retirer entre 76.5 et 88.4% des variants de chaque individu (**Figure** : 2.12 - A).

Cependant, regarder uniquement le pourcentage de variants filtrés par chaque filtre révèle une information partielle. En effet, dans ce cas de figure, on observe la quantité de variant éliminée par chaque filtre indépendamment les uns des autres. Ainsi, un même variant peut donc être filtré par plusieurs filtres. Dès lors, il faut également analyser la quantité de variants filtrée **spécifiquement** par chaque filtre. Ainsi, on peut constater que le classement des filtres en fonctions de leur stringence reste quasiment identique. Il est tout de même intéressant de noter que désormais le filtre “*Variant impact*” apparaît moins efficace que les filtres “*Ctrl*” et “*Genotype*” en filtrant spécifiquement une moyenne de 253 variants par individu contre 423 pour le filtre génotype et 882 pour le filtre “*Ctrl*”. Ainsi, ce dernier devient celui filtrant spécifiquement le plus de variants avec entre 364 et 1060 variants spécifiquement filtrés par patients confirmant ainsi l'importance de ce filtre dans nos analyses. Aussi, les filtres “*Transcript relevance*”, “*Union*” et “*Frequency*” apparaissent désormais comme étant anecdotiques en comparaison des trois autres filtres puisqu'ils filtrent au maximum 43 variants spécifiques (**Figure** : 2.12 - B).

Après avoir appliqué l'ensemble de ces filtres, seuls quelques variants subsistent, nous permettant d'obtenir une liste de restreinte gènes pour chaque famille et ainsi de tirer des conclusions quant au(x) variant(s) responsable(s) du phénotype de chacune d'entre elles. Ces travaux ont ainsi pu mener à l'écriture de trois articles dont je suis co-auteur.

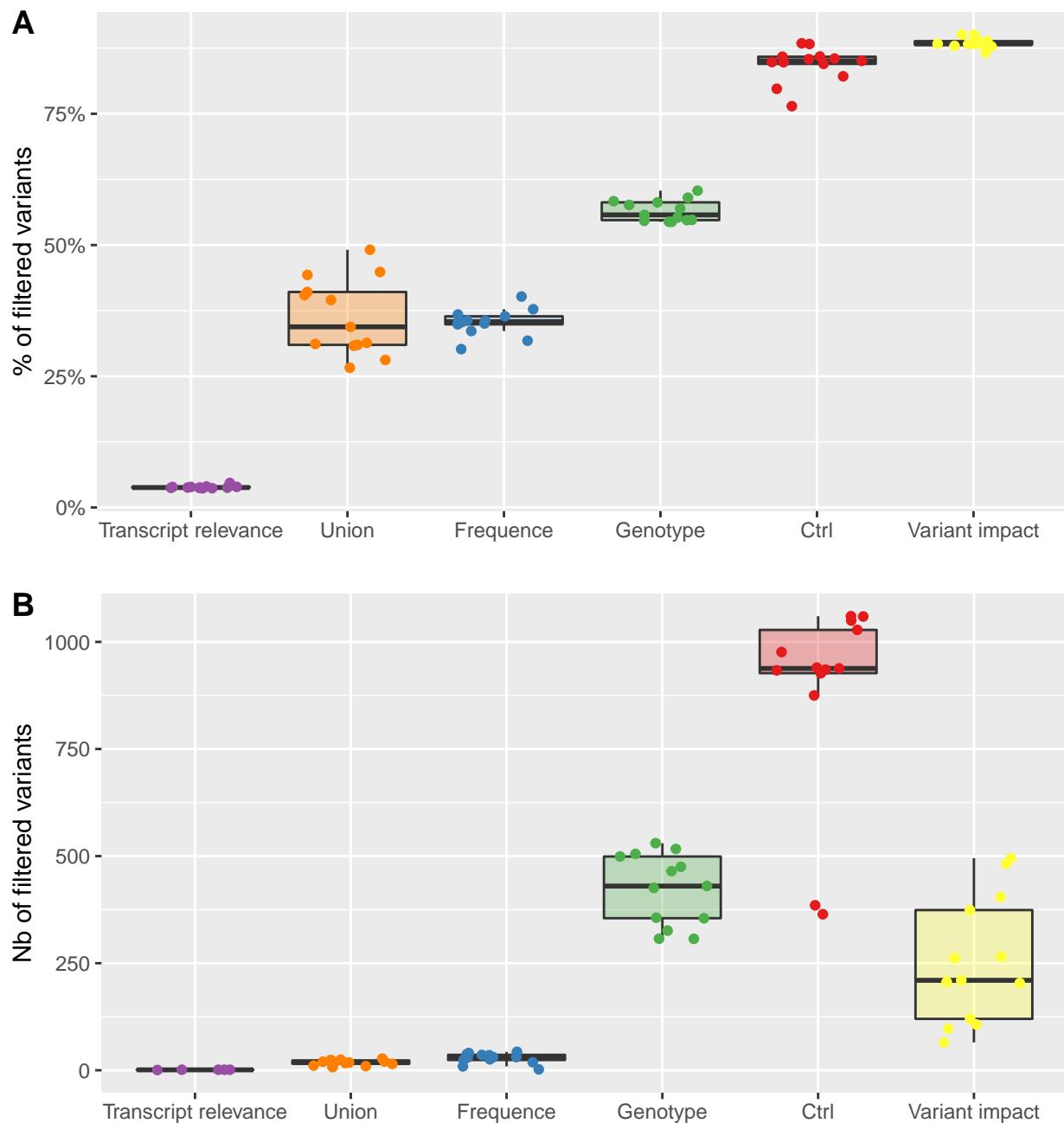


Figure 2.12 – Comparaison de l'efficacité de chacun des six filtres utilisés : A : Comparaison du pourcentage de variants filtrés par chacun des six filtres indépendamment les uns des autres pour chaque patient (représenté par les points). Dès lors, un même variant peut-être filtré par plusieurs filtres. B : Comparaison du nombre de variants filtrés spécifiquement par chacun des filtres. Ici, un variant ne peut-être filtré que par un seul filtre.

2.2.2 Article n°1

SPINK2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes

Kherraf ZE*, Christou-Kent M*, Karaouzène T, Amiri-Yekta A, Martinez G, Vargas AS, Lambert E, Borel C, Dorphin B, Aknin-Seifer I, Mitchell MJ, Metzler-Guillemain C, Escoffier J, Nef S, Grepillat M, Thierry-Mieg N, Satre V, Bailly M, Boitrelle F, Pernet-Gallay K, Hennebicq S, Fauré J, Bottari SP, Coutton C, Ray PF, Arnoult C

* Co-premiers auteurs

EMBO Molecular Medicine, Mai 2017

Contexte et objectifs

L'oligospermie, comme l'azoospermie sont des phénotypes d'infertilité masculine liées à la quantité de spermatozoïdes présents dans l'éjaculat. Les différentes études publiées ces dernières années montrent que les microdélétions du chromosome Y sont retrouvées chez 10% des hommes avec une azoospermie non-obstructive et chez 5% des patients avec une oligozoospermie sévère [74]. Ces taux bien qu'élevés ne représentent qu'une infime partie des cas d'azoospermie et d'oligospermie, suggérant l'implication de nombreux autres gènes dans ce phénotype.

Entre 2005 et 2014 deux frères issus d'une union consanguine ont demandé des conseils médicaux auprès de différentes cliniques d'infertilité après deux ans d'échec reproductif. Ces deux frères étant mariés à des femmes non-apparentées, une cause féminine fût exclue et les recherches ont été concentrées sur l'analyse des deux frères. Tous deux présentaient de sévères défauts de production de spermatozoïdes, l'un des frères présentant une azoospermie non-obstructive et l'autre une oligozoospermie extrême (<1 Million de spermatozoïde/ ml). Au vu de la similarité du phénotype et de leur lien de parenté, l'hypothèse d'une cause génétique commune fut émise. L'analyse de leur caryotype et du locus AZF du chromosome Y ne révélant aucune anomalie, la procédure d'un séquençage WES fut décidée.

Dans ce contexte, l'objectif de mon travail a été d'effectuer l'ensemble des analyses des données WES obtenues après leur séquençage afin de mettre en évidence une mutation homozygote commune pouvant expliquer le déficit spermatique des deux frères. Dans un second temps, j'ai mis en place le protocole de génotypage des souris au locus du gène *Spink2* permettant d'identifier les souris sauvages *Spink2^{+/+}* des souris KO *Spink2^{-/-}*. Pour finir, afin d'estimer l'importance des variants du gène *SPINK2* comme cause d'infertilité masculine chez l'humain, j'ai également contribué au séquençage Sanger de la séquence codante de *SPINK2* d'une partie des 611 patients séquencés dans cette étude.

SPINK2 deficiency causes infertility by inducing sperm defects in heterozygotes and azoospermia in homozygotes

Zine-Eddine Kherraf^{1,†}, Marie Christou-Kent^{1,†}, Thomas Karaouzene¹, Amir Amiri-Yekta^{1,2,3}, Guillaume Martinez¹, Alexandra S Vargas¹, Emeline Lambert¹, Christelle Borel⁴, Béatrice Dorphin⁵, Isabelle Aknin-Seifer⁶, Michael J Mitchell⁷, Catherine Metzler-Guillemain⁷, Jessica Escoffier¹, Serge Nef⁴, Mariane Grepillat¹, Nicolas Thierry-Mieg⁸, Véronique Satre^{1,9}, Marc Bailly^{10,11}, Florence Boitrelle^{10,11}, Karin Pernet-Gallay¹², Sylviane Hennebicq^{1,13}, Julien Faure^{2,12}, Serge P Bottari^{1,14}, Charles Coutton^{1,9}, Pierre F Ray^{1,2,‡,*}  & Christophe Arnoult^{1,‡}

Abstract

Azoospermia, characterized by the absence of spermatozoa in the ejaculate, is a common cause of male infertility with a poorly characterized etiology. Exome sequencing analysis of two azoospermic brothers allowed the identification of a homozygous splice mutation in *SPINK2*, encoding a serine protease inhibitor believed to target acrosin, the main sperm acrosomal protease. In accord with these findings, we observed that homozygous *Spink2* KO male mice had azoospermia. Moreover, despite normal fertility, heterozygous male mice had a high rate of morphologically abnormal spermatozoa and a reduced sperm motility. Further analysis demonstrated that in the absence of Spink2, protease-induced stress initiates Golgi fragmentation and prevents acrosome biogenesis leading to spermatid differentiation arrest. We also observed a deleterious effect of acrosin overexpression in HEK cells, effect that was alleviated by *SPINK2* coexpression confirming its role as acrosin inhibitor. These results demonstrate that *SPINK2* is necessary to neutralize proteases during their cellular transit toward the acrosome and that its deficiency induces a pathological continuum ranging from oligoasthenoteratozoospermia in heterozygotes to azoospermia in homozygotes.

Keywords azoospermia; exome sequencing; genetics; infertility; spermatogenesis

Subject Categories Genetics, Gene Therapy & Genetic Disease; Urogenital System

DOI 10.15252/emmm.201607461 | Received 13 December 2016 | Revised 14 April 2017 | Accepted 26 April 2017

Introduction

The World Health Organization estimates that 50 million couples worldwide are confronted with infertility. Assisted reproduction technologies (ART) initiated 35 years ago by Nobel Prize Winner Robert Edwards have revolutionized the practice of reproductive medicine, and it is now estimated that approximately 15% of couples in Western countries seek assistance from reproductive clinics for infertility or subfertility. Despite technological breakthroughs and advances, approximately half of the couples concerned still fail to achieve a successful pregnancy even after repeated treatment cycles. Alternative treatment strategies should therefore be

¹ Genetic Epigenetic and Therapies of Infertility, Institute for Advanced Biosciences, Inserm U1209, CNRS UMR 5309, Université Grenoble Alpes, Grenoble, France

² CHU de Grenoble, UF de Biochimie Génétique et Moléculaire, Grenoble, France

³ Department of Genetics, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran

⁴ Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva 4, Switzerland

⁵ Laboratoire d'Aide Médicale à la Procréation, Centre AMP 74, Contamine-sur-Arve, France

⁶ Laboratoire de Biologie de la Reproduction, Hôpital Nord, Saint Etienne, France

⁷ Aix Marseille Univ, INSERM, GMGF, Marseille, France

⁸ Univ. Grenoble Alpes / CNRS, TIMC-IMAG, Grenoble, France

⁹ CHU de Grenoble, UF de Génétique Chromosomique, Grenoble, France

¹⁰ Department of Reproductive Biology and Gynaecology, Poissy General Hospital, Poissy, France

¹¹ EA 7404 GIG, Université de Versailles Saint Quentin, Montigny le Bretonneux, France

¹² Grenoble Neuroscience Institute, INSERM 1216, Grenoble, France

¹³ CHU de Grenoble, UF de Biologie de la procréation, Grenoble, France

¹⁴ CHU de Grenoble, UF de Radioanalyses, Grenoble, France

*Corresponding author. Tel: +33 4 76 76 55 73; E-mail: pray@chu-grenoble.fr

†These authors contributed equally to this work

‡These authors contributed equally to this work as senior authors

envisioned to improve ART success rate, especially for patients impervious to usual assisted reproductive technologies. Improvement in treatment efficiency essentially depends upon an accurate diagnosis and the characterization of the molecular etiology of the defect. These efforts to better characterize infertility subtypes should first be concentrated on the most severe defects since they generally have a poor prognosis and affected patients would benefit the most from new treatments. Moreover, the most severe phenotypes are more likely to be caused by monogenic defects which are easier to identify. As such, the genetic exploration of non-obstructive azoospermia (NOA), the absence of spermatozoa in the ejaculate due to a defect in spermatogenesis, should be considered a priority. NOA is a common cause of infertility found in approximately 10% of the couples assessed for infertility. Although a genetic etiology is likely to be present in most cases of azoospermia, only a few defective genes have so far been associated with this pathology accounting for a minority of cases. At present, only chromosomal abnormalities (mainly 47XXY, Klinefelter syndrome identified in 14% of cases) and microdeletions of the Y chromosome are routinely diagnosed, resulting in a positive genetic diagnosis in < 20% of azoospermia cases (Tuttmann et al., 2011). The evolution of sequencing technologies and the use of whole-exome or whole-genome sequence (WES/WGS) analysis paves the way to a great improvement in our ability to characterize the causes of genetically heterogeneous pathologies such as NOA.

Spermatogenesis can be subdivided into three main steps: (i) multiplication of diploid germ cells; (ii) meiosis, with the shuffling of parental genes and production of haploid cells; and (iii) spermiogenesis, the conversion of round spermatids into one of the smallest and most specialized cells in the body, the spermatozoa. NOA is expected to be mainly caused by failures in steps 1 and 2, and it is indeed what has been observed in a majority of cases so far. Very recently, defects in six genes were linked to azoospermia in man. Most of these genes code for meiosis-controlling proteins such as TEX11, TEX15, SYCE1, or MCM8, and the absence of the functional proteins induces a blockage of meiosis (Tuttmann et al., 2011; Maor-Sagie et al., 2015; Okutman et al., 2015; Yang et al., 2015; Yatsenko et al., 2015). Another WES analysis of two consanguineous families identified likely causal mutations in TAF4B and ZMYND15 (Ayhan et al., 2014). Study of *Taf4b* KO mice showed that homozygous mutant males are subfertile with extensive pre-meiotic germ cell loss due to altered differentiation and self-renewal of the spermatogonial stem cell pool, thus illustrating that pre-meiotic block induces NOA. More surprisingly, *ZMYND15* codes for a spermatid-specific histone deacetylase-dependent transcriptional repressor and its absence in mice induced a significant depletion of late-stage

spermatids (Yan et al., 2010) suggesting that NOA can also be induced by post-meiotic defects.

Here, WES analysis of two brothers with NOA led to the identification of a homozygous truncating mutation in the *SPINK2* gene coding for a Kazal family serine protease inhibitor. Studying KO mice, we observed that homozygous KO animals also suffered from azoospermia thus confirming the implication of *SPINK2* in NOA. Furthermore, we observed that *SPINK2* is expressed from the round-spermatid stage onwards thus confirming that post-meiotic anomalies can result in NOA. We suggest that *SPINK2* is necessary to neutralize the action of acrosomal proteases shortly after their synthesis and before they can be safely stored in the acrosome where they normally remain dormant until their release during the acrosome reaction. We also show that in the absence of *SPINK2*, protease-induced stress initiates Golgi fragmentation contributing to the arrest of spermatid differentiation and their shedding from the seminiferous epithelium. The characterization of the molecular pathophysiology of this defect opens several novel therapeutic perspectives which may allow the restoration of a functional spermatogenesis.

Results

Medical assessment of two brothers with defective sperm production

Two French brothers (Br1 and Br2), born from second cousin parents (Fig 1A), and their respective wives sought medical advice from infertility clinics in France (Chatellerault, Tours, Poissy, and Grenoble) between 2005 and 2014 after 2 years of unsuccessful attempts to spontaneously conceive. Analyses of their ejaculates (Fig 1B; $n = 5$) evidenced the absence of spermatozoa for the first brother (Br1) and a very low concentration (0–200,000/ml, mean 126,000/ml $n = 5$) for the second (Br2). Moreover, all spermatozoa were immotile and presented an abnormal morphology (pin-shaped head devoid of acrosome; detached flagella) and were not suitable for *in vitro* fertilization (IVF) with intracytoplasmic sperm injection (ICSI). Interestingly, ejaculates of both brothers presented a significant concentration of germ cells ($8.6 \times 10^6 \pm 6.2 \times 10^6$ /ml and $9.0 \times 10^6 \pm 7.0 \times 10^6$ /ml for Br-1 and Br-2, respectively) likely corresponding to spermatids. As Br1 and Br2 both present a severe default of sperm production with a high number of spermatids in the ejaculate, we believe that they present the same phenotype, likely caused by the same genetic defect. A normal karyotype was observed for both brothers (46,XY), and no deletions of the Y

Figure 1. Azoospermia in two consanguineous brothers.

- A Genetic tree of the studied family showing affected brothers Br1 and Br2 illustrating the consanguinity of the parents (P1 and P2).
- B Comparisons of ejaculate volume ($n = 5$) and spermograms ($n = 5$) of brothers Br1 and Br2 with those of fertile controls ($n = 35$) evidence the absence of mature sperm and the presence of round cells in the ejaculates. Data represent mean \pm SEM. P-values are $P = 4 \times 10^{-4}$ (a), $P = 0.6$ (b, non-significant), and $P = 4 \times 10^{-5}$ (c); statistical differences were assessed using t-test.
- C, D Testis sections from a fertile control (D) patient Br1 stained with periodic acid-Schiff (PAS). The lumen of tubules from the control is large and mature sperm are present (C), whereas the lumen of most of seminiferous tubules from patient Br1 is filled with non-condensed and early condensed round spermatids and no mature sperm are observed. Scales bars, 100 μ m.
- E, F In the fertile control (E) seminiferous tubule cross sections, spermatogonia (Sg), spermatocytes (Sc) and spermatids (RS) are regularly layered, whereas the different types of spermatogenic cells are disorganized in patient Br1 (F). Scales bars, 100 μ m.

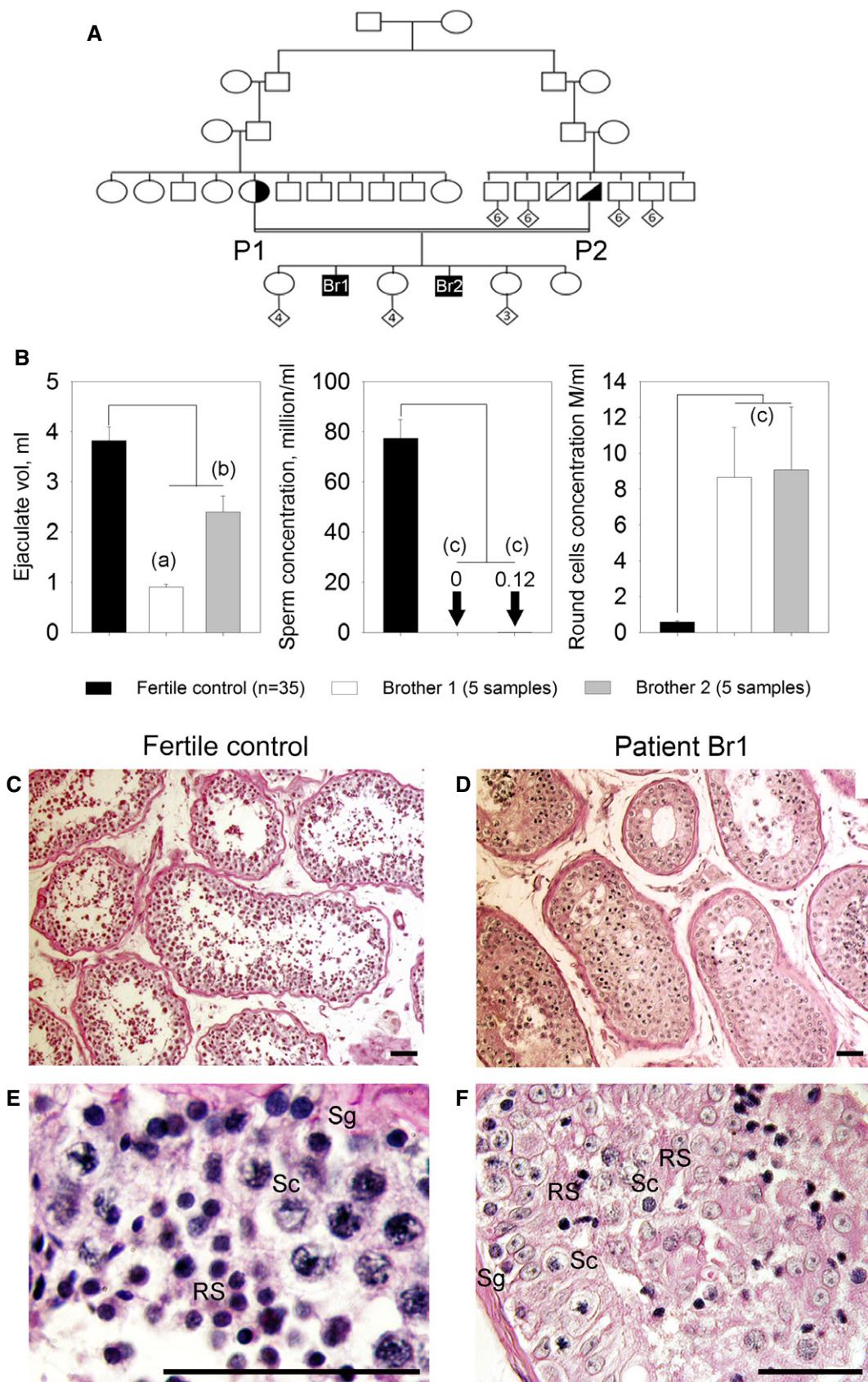


Figure 1.

chromosome were observed at the AZF loci. Testis sperm extraction was carried out twice for Br1 in 2008 and 2014. Each time the recovery was unsuccessful (although a few spermatozoa were observed in fixed dilacerated testicular tissues) suggesting a diagnosis of post-meiotic NOA. Histological analysis of seminiferous tubules obtained from Br1 biopsies showed: (i) a disorganization of the structure of the tubules; (ii) that the lumen of the seminiferous tubules were filled with immature germ cells, an indication of intense desquamation of the germinal epithelium; and (iii) a reduced number of round spermatids, with an overrepresentation of early round spermatids (Fig 1C–F). Brother Br2 has only had spermograms for diagnostic purposes which did not show any ICSI-compatible spermatozoa and has not been able to attempt ART.

Whole-exome sequencing identifies a homozygous truncating mutation in SPINK2

Since the brothers were married to unrelated women, we excluded the possibility of a contributing female factor and focused our research on the brothers. Given the familial history of consanguinity, we postulated that their infertility was likely caused by a common homozygous mutation. We proceeded with WES to identify a possible genetic defect(s) which could explain the observed azoospermia. After exclusion of common variants, both brothers carried a total of 121 identical missense heterozygous variants (none appearing as obvious candidate) and only five identical homozygous variants common to both brothers (Appendix Table S1). Among these different genes, only the Chr4:57686748G>C SPINK2 variant was described to be predominantly expressed in human testis (Appendix Fig S1A) as well as in mouse testis (Appendix Fig S1B). The mutation was validated by Sanger sequencing in both brothers (homozygous) and their parents (heterozygous) (Fig 2A). SPINK2 thus appeared as the best candidate to explain the human condition. The variant Chr4:57686748G>C was not present in > 121,000 alleles analyzed in the ExAC database (<http://exac.broadinstitute.org>) and could have an effect on RNA splicing. SPINK2 is located on chromosome 4 and contains four exons (Fig 2B). The gene codes for a Kazal type 2 serine protease inhibitor also known as an acrosin–trypsin inhibitor. The Ensembl expression database (www.ensembl.org) predicts the presence of four transcripts. We studied the expression of the different transcripts in human testis by RT–PCR, and only one band was present corresponding to NM_021114, ENST00000248701, which codes for a protein of 9.291 kDa consisting of 84 amino acids (Fig 2C). All nucleotide sequences herein refer to this transcript. The identified mutation, c.56-3C>G, is located three nucleotides before exon 2 and may create a new splice acceptor site, leading to a frame-shift and premature stop codon in exon 2 and the generation of an abnormal transcript (T1) and/or to the skipping of exon 2 (44 nt) giving rise to an early stop codon at the beginning of exon 3 and the generation of another abnormal transcript (T2) (Fig 2B). To validate these hypotheses, RT–PCR was performed on testicular extract from Br1. Two bands were observed (Fig 2C) and sequenced after isolation of each band following gel electrophoresis. Sequence analysis demonstrated that the bands corresponded to T1 and T2, demonstrating that both abnormal transcripts were present in the patient's testis (Fig 2D). Since the protease inhibitor and binding sites of the protein are coded mostly by exon 3, it is expected that the truncated proteins corresponding to T1 and T2 transcripts are not functional

(Appendix Fig S2). Sequencing of Br1's transcripts therefore confirms that the identified splice variant abrogates the production of a full-length protein thereby confirming its role as a deleterious mutation.

Importance of SPINK2 variants as a cause for human infertility: sequence analysis of a cohort of infertile men with an altered spermatogenesis

We sequenced SPINK2 whole coding sequences of 611 patients affected by azoo- or oligozoospermia (210 patients with azoospermia, 393 subjects with oligozoospermia and 8 with unspecified cause). Only one variant, identified in patient 105 (P105), was not described in ExAC and was likely deleterious (Appendix Table S2). This variant, c.1A>T (Fig EV1A), abrogates the SPINK2 start codon and was present heterozygously in P105, a man with oligozoospermia. An alternate start site could potentially be used in the middle of exon 3 allowing the synthesis of a truncated protein of 2 kDa lacking the reactive site and disulfide bonds, both known to be crucial for SPINK2 function (Fig EV1B). However, overexpression of the mutated gene in HEK cells did not produce any portion of the SPINK2 protein indicating that the putative alternative start site is not functional and that the alteration of the initial start site does not permit the synthesis of any part of the SPINK2 protein. This was evidenced by transfecting HEK293 cells with a plasmid containing the full human SPINK2 ORF sequence with the c.1A>T mutation and a C-terminus DDK-tagged. Extracted proteins were loaded onto a 20% acrylamide gel and detected with anti-DDK or anti-SPINK2 antibodies (Fig EV1C). P105 and his wife, born from non-consanguineous parents, experienced a 5-year period of infertility before giving birth to a healthy boy conceived spontaneously. They sought medical advice 2 years after their son's birth to initiate a second pregnancy. Sperm analysis resulted in the diagnosis of oligozoospermia associated with a reduced percentage of progressive motile spermatozoa (Table EV1). The patient's sperm morphology was assessed with Harris–Shorr staining using the modified David's classification and showed that 34–39% of sperm had a normal morphology ($n = 2$). The main defects observed were abnormal acrosome (34–39%) and defective neck–head junction (40–46%), defects that are similar to those observed in patient Br2.

This analysis indicates that SPINK2 defects are extremely rare with an allelic frequency of approximately 1/1,200 in the cohort of infertile men analyzed. The rarity of SPINK2 variants and the fact that P2, the father of Br1 and Br2, also harboring a heterozygous mutation, presents in a milder phenotype than P105 could indicate that SPINK2 haploinsufficiency induces a milder phenotype of oligozoospermia with an incomplete penetrance on infertility.

Homozygous *Spink2* KO mice have azoospermia due to a spermiogenesis blockade at the round-spermatid stage

In order to confirm that the absence of SPINK2 leads to azoospermia, homozygous *Spink2* KO ($^{−/−}$) mice were obtained and their reproductive phenotype was studied. We first performed qRT–PCR on *Spink2*^{+/+} and *Spink2*^{−/−} testis mRNA extracts to validate the absence of *Spink2* mRNA and thus of protein. Contrary to what was observed in WT littermates, we observed no *Spink2* amplification in KO males, confirming *Spink2* deficiency (Appendix Fig S3). Males

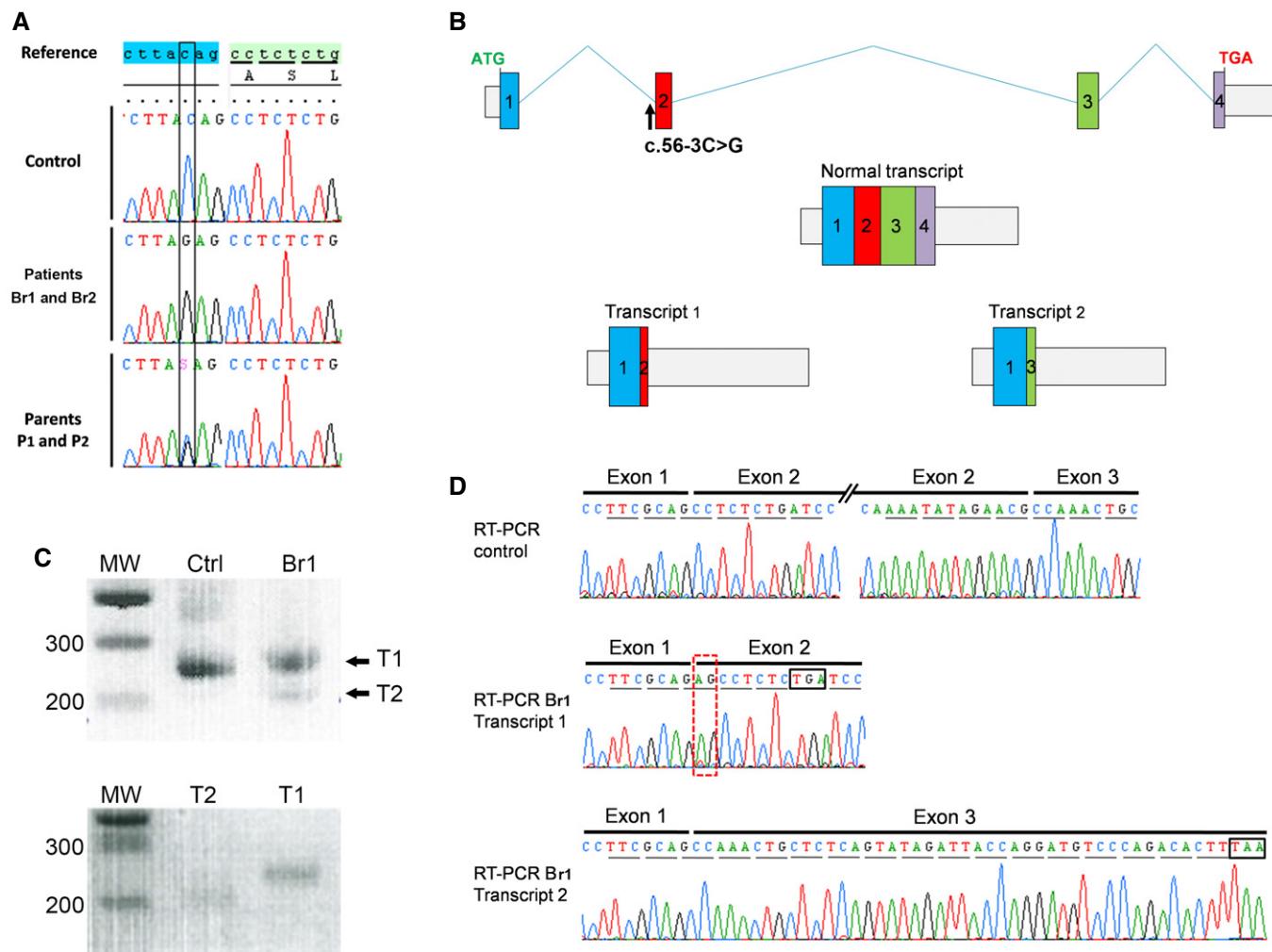


Figure 2. Identification of a SPINK2 variant (c.56-3C>G) by exome sequencing and its consequences on splicing and translation.

- A The identified variant, homozygous in patients 1 and 2 and heterozygous in their parents, is located three nucleotides before exon 2 and creates an AG that immediately precedes the original AG splice acceptor site.
- B If recognized during splicing, this new acceptor site is expected to add two nucleotides (AG) at the beginning of exon 2, inducing a frameshift leading to a stop codon 3 amino acids later (transcript 1). The non-recognition of the abnormal acceptor site is expected to induce the skipping of exon 2 (transcript 2). The first stop codon can be observed 15 codons after the mis-inserted exon 3.
- C RT-PCR of mRNA extracts from fertile control (Ctrl) and the brother Br1. Results show one band for Ctrl. The sequencing of this band showed that it corresponds to transcript NM_021114. For Br1, two bands were present, named T1 and T2. Bottom gel shows T1 and T2 after gel isolation.
- D Transcripts T1 and T2 were collected and sequenced: T1 showed the insertion of an additional AG (red-dashed rectangle) leading to a premature stop codon (black box), whereas transcript T2 showed that exon 2 had been excised; these two transcripts correspond to the expected transcripts 1 and 2 from panel (B). Stop codons are shown in black boxes.

were completely infertile, whereas no reproductive defects were observed in females (Fig 3A1). Homozygous KO mice had comparatively smaller sized testes and a testis/body weight ratio half that of their wild-type (WT) littermates [3.63 ± 0.21 in WT and 1.77 ± 0.03 in KO (Fig 3A2)]. Furthermore, there was a complete absence of spermatozoa in *Spink2*^{-/-} caudal epididymis (Fig 3A3) which only contained round cells likely corresponding to round spermatids and multinucleated cells, known as symblasts. Histological studies of KO seminiferous tubules stained with periodic acid-Schiff (PAS) revealed the presence of germ cells up to the early round-spermatid stage but condensed and elongated spermatids and

mature spermatozoa were completely absent, contrary to WT (Fig 3B1 and C1). The lumen of the seminiferous tubules of *Spink2*^{-/-} males contained round cells and symblasts (Fig EV2A and B), a result in agreement with observations of the cellular content of the cauda epididymis, which showed the presence of round cells only (Fig EV2C). In contrast to what was observed in WT (Fig 3B2), sections of caudal epididymis confirmed the absence of spermatozoa and the presence of symblasts and round cells (Fig 3C2). Comparing PAS staining of *Spink2*^{+/+} and *Spink2*^{-/-} seminiferous tubules, we noticed that contrary to WT, *Spink2*^{-/-} round spermatids did not contain an acrosomal vesicle, suggesting

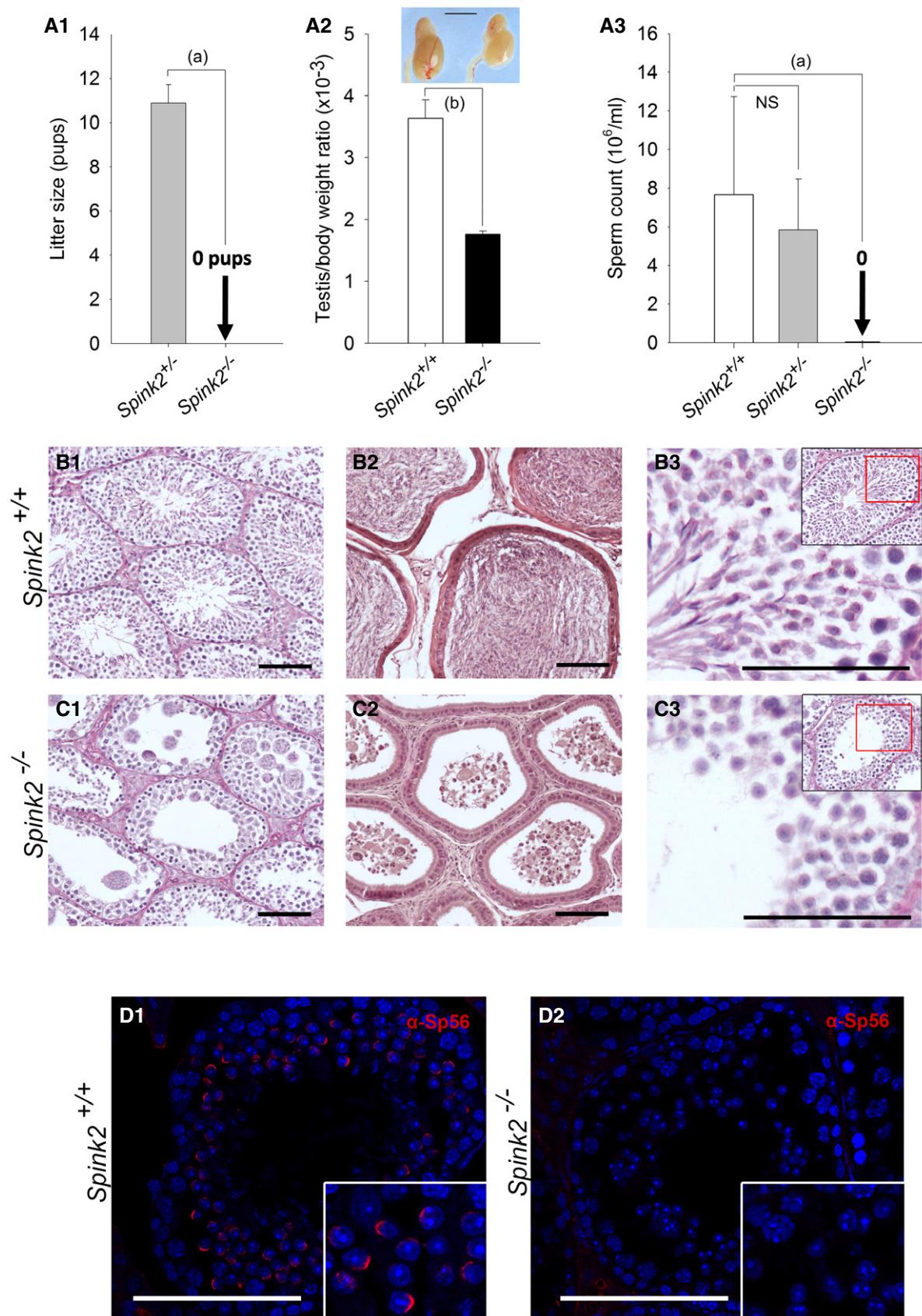


Figure 3.

Figure 3. *Spink2*^{-/-} males are infertile and azoospermic, and spermatogenesis presents a post-meiotic blockade.

- A1 Litter size of *Spink2*^{-/-} and *Spink2*^{+/+} males mated with wild-type females ($n = 5$).
- A2 Testis/body weight ratio for WT and *Spink2*^{-/-} mice ($n = 6$) and morphology and size of wild-type and *Spink2*^{-/-} testes of male siblings. Scale bar, 5 mm.
- A3 Sperm concentrations from the cauda epididymis of wild-type, *Spink2*^{+/+}, and *Spink2*^{-/-} male testes ($n = 10$).
- B, C Histological comparisons of testis and epididymis from WT and *Spink2*^{-/-} mice. (B1, C1) Periodic acid–Schiff (PAS) staining of seminiferous tubule cross sections shows complete spermatogenesis in WT (B1) contrary to *Spink2*^{-/-} mice (C1), where condensed, elongated spermatids and mature sperm are absent. (B2, C2) Sections of epididymis stained with eosin/hematoxylin. In the lumen of tubules from WT mice, mature sperm are present (B2), whereas only round cells and multinucleated spheroblasts occupy the lumen of tubules from *Spink2*^{-/-} mice (C2). (B3, C3) Enlargement of seminiferous tubule sections stained with PAS evidences deep pink staining in round spermatids, which corresponds to the acrosome in WT mice (B3), whereas round spermatids from *Spink2*^{-/-} mice present no deep pink staining, indicating that the acrosome is not formed (C3). Scale bars, 100 μ m.
- D1, D2 Immunofluorescence experiments using an anti-Sp56 antibody (red staining) confirm the presence of the acrosome in seminiferous tubule sections from WT contrary to those from *Spink2*^{-/-} mice, where no staining is observed. Scale bars, 100 μ m.

Data information: Data represent mean \pm SEM. P-values are $P = 1 \times 10^{-5}$ (a) and $P = 1 \times 10^{-4}$ (b); statistical differences were assessed using t-test. NS, not statistically significant.

that the absence of Spink2 prevents acrosome biogenesis (Fig 3B3 and C3). This point was confirmed by immunofluorescent staining using the Sp56 antibody, a specific marker of the acrosome (Kim *et al.*, 2001) (Fig 3D1 and D2). We then identified the spermatogonia using an anti-PLZF antibody (Zhang *et al.*, 2014) (Fig EV3A and B) and observed no significant difference in the median number of spermatogonia per tubule ($n = 30$) between *Spink2*^{+/+} and *Spink2*^{-/-} mice (Fig EV3C). These results indicate that the absence of Spink2 does not impact spermatogonial survival but leads to an early arrest of round-spermatid differentiation. Overall, the *Spink2*^{-/-} mouse phenotype perfectly mimics the human condition and confirms that SPINK2 deficiency is involved in human azoospermia.

SPINK2 is an acrosomal protein

In order to further investigate the molecular pathogenicity of this SPINK2-dependent azoospermia, we determined the localization of SPINK2 in human and mouse testis. We first verified the specificity of a SPINK2 antibody through Western blot (WB) and immunofluorescence (IF) experiments on HEK293 cells overexpressing human SPINK2. In Western blots, the SPINK2 antibody recognized three bands of less than 17 kDa weight, likely corresponding to oligomeric complexes (Appendix Fig S4A). No bands appeared in non-transfected cells. Moreover, the overexpressed SPINK2 featured a DDK-tag which was recognized by an anti-DDK-tag antibody revealing three bands of identical molecular mass (Appendix Fig S4B). No bands were observed when the primary antibody was omitted. SPINK2 expression was also studied by IF and confocal microscopy. Transfected cells displayed a cytoplasmic staining, whereas no staining was observed in non-transfected cells (Appendix Fig S4C). Taken together, these results demonstrate the specificity of this antibody in WB and IF experiments. Next, the localization of SPINK2 was determined by IF in human and mouse seminiferous tubule cross sections and in mature sperm (Fig EV4). In mouse, SPINK2 was present in the acrosomal vesicle from the beginning of the acrosome's biogenesis at the round-spermatid stage as indicated by a colocalization with Sp56, a marker of the acrosome (Fig EV4A and B). In accordance with the results shown in Fig 3D2, no SPINK2 staining was observed in *Spink2*^{-/-} testis cross sections (Fig EV4C). A similar localization was observed for SPINK2 in human seminiferous tubule sections (Fig EV4D). Finally, we observed that SPINK2 remains present in the acrosome of human and mouse mature spermatozoa (Fig EV4E and F).

Ultrastructure of *Spink2*^{-/-} round spermatids shows that fusion of proacrosomal vesicles is hampered and that the Golgi apparatus is fragmented

We showed that SPINK2 is located in the acrosome and that its absence prevents acrosome biogenesis. To understand the reasons for the absence of acrosome biogenesis, we performed transmission electron microscopy (EM) to study the ultrastructure of round spermatids from *Spink2*^{-/-} males (Fig 4). In wild-type round spermatids, proacrosomal vesicles generated by the Golgi apparatus docked in a specialized area of the nuclear envelope (NE) and fused together to form a giant acrosomal vesicle (Fig 4A). Contrary to WT, in *Spink2*^{-/-}, the proacrosomal vesicles generated by the Golgi apparatus of round spermatids were mostly unable to fuse (Fig 4B2, white arrowheads), likely explaining the absence of acrosome biogenesis. Moreover, the Golgi apparatus from *Spink2*^{-/-} animals produced abnormal proacrosomal vesicles of irregular sizes (Fig 4B2) and showed a considerable disorganization with a decreased proportion of flattened membrane stacks (Fig 4B2) displaying shorter lengths (Fig 4C). Acrosome biogenesis is dependent on the simultaneous synthesis of vesicles by the Golgi apparatus and the modification of the nuclear envelope (NE) facing the Golgi apparatus, with tight apposition of both nuclear membranes and aggregation of a nuclear dense lamina (NDL) on the nuclear side of the inner nuclear membrane (Kierszenbaum *et al.*, 2003). In *Spink2*^{-/-} round spermatids, the densification of the NE appears to occur normally and the NDL is clearly visible in EM (Fig 4B2). Using IF, the modification of the NE facing the Golgi apparatus was followed with an anti-Dpy19l2 antibody. We indeed had previously shown that Dpy19l2 participates in linking the acrosome to the nucleus and that it is located in the nuclear membrane facing the forming acrosome (Pierre *et al.*, 2012) and is thus a component of this specialized area of the nuclear envelope. In costaining experiments using anti-Dpy19l2 and anti-GM130 antibodies to stain the nuclear envelope facing the acrosomal vesicle (evidenced by the NDL in EM) and the Golgi apparatus, respectively, we found that in WT round spermatids, the Golgi apparatus is either located immediately in front of the NDL in the early phase of acrosome biogenesis or, at a slightly later stage, lies adjacent to it (Fig 4D1 and D2). In contrast to WT, the Golgi apparatus of *Spink2*^{-/-} round spermatids was positioned randomly around the nucleus, often found on the opposite side of the NDL (Fig 4D3–D6) indicative of a disruption of the polarity of the NDL and of the Golgi apparatus, which should both be located at the apical face of the round spermatid.

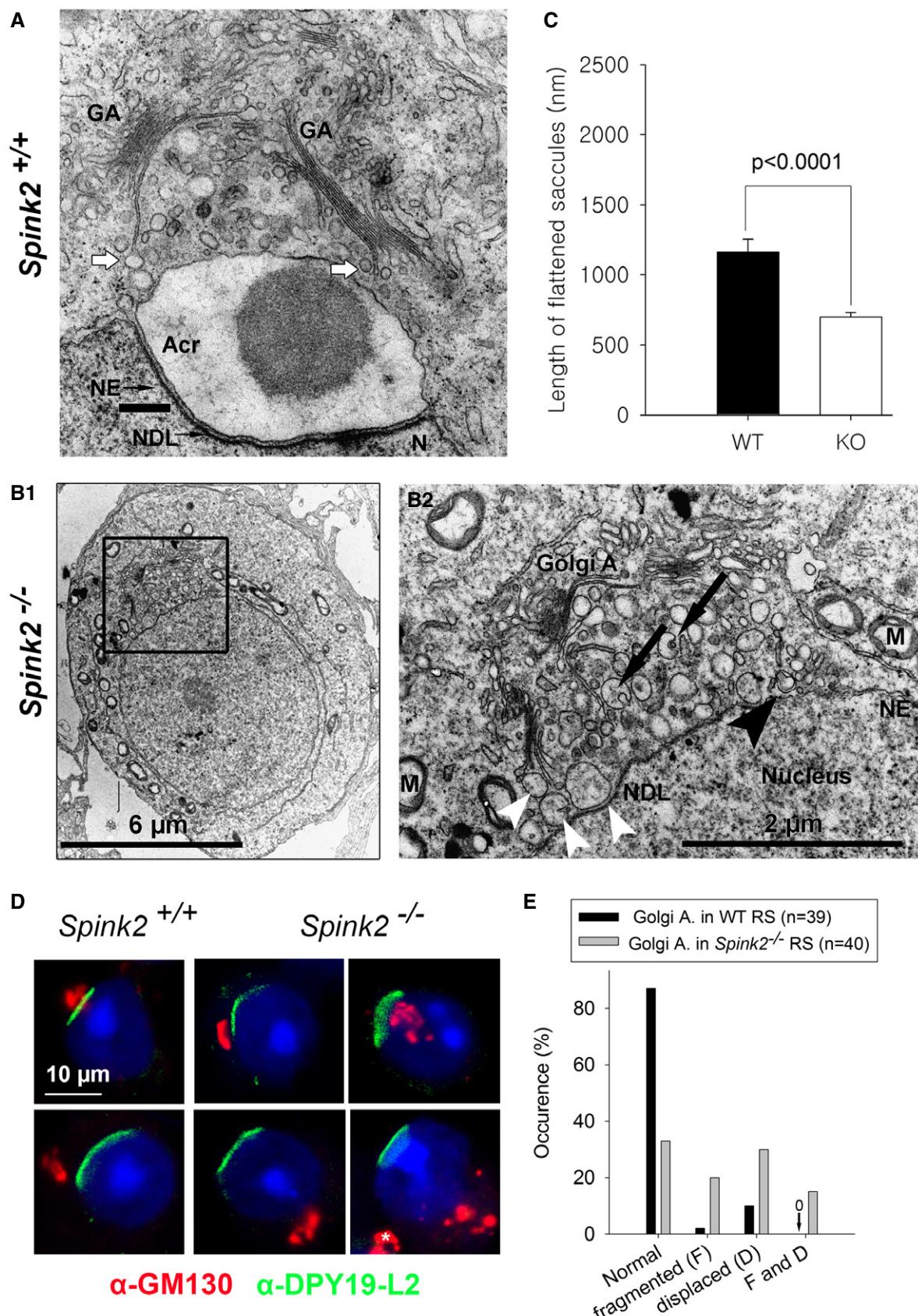


Figure 4.

Figure 4. Lack of Spink2 prevents the fusion of proacrosomal vesicles and induces a disorganization of the Golgi apparatus.

- A Partial section of a WT round spermatid observed by EM showing the early biogenesis of the acrosome (Acr) due to the continuous formation and aggregation of small vesicles (white arrows) coming from the Golgi apparatus (GA). The nuclear envelope (NE) facing the acrosome has a specific organization and is associated with the nuclear dense lamina (NDL). N, nucleus. Scale bar, 400 nm.
- B Ultrastructure of the Golgi apparatus in *Spink2*^{-/-} round spermatid observed by EM. (B1) Ultrastructure of a *Spink2*^{-/-} round spermatid observed at low magnification. The black box corresponds to the Golgi apparatus and is enlarged in (B2). (B2) In the absence of Spink2, vesicles do not aggregate at the nuclear envelope although modification of the NE and formation of the NDL occur. Unfused vesicles of different sizes accumulate in the cytoplasm with very few docking on the nuclear envelope (white arrowhead). Moreover, the GA shows disorganization with strong decrease or absence of stacks of flattened membranes. Finally, microautophagy-like structures and vesicles with a double membrane (black arrowhead) are observed around the GA (black arrows). M, mitochondria. Scale bars, 6 μm (B1) and 2 μm (B2).
- C The length of flattened saccules is statistically reduced in *Spink2*^{-/-} round spermatids (WT saccules, n = 74; and KO saccules, n = 136). Data represent mean ± SEM; the statistical difference was assessed with t-test, P-value as indicated.
- D Absence of Spink2 induces Golgi apparatus fragmentation and mislocalization. (D1, D2) IF experiments using an anti-Dpy19l2 antibody marking the specific NE facing the NDL (green staining) and an anti-GM130 antibody marking the cis-Golgi (red staining) show that the Golgi apparatus (GA) is a compact structure and located either in front of the NDL or close to it in WT round spermatids (normal). (D3–D6) In contrast, similar double staining of round spermatids from *Spink2*^{-/-} mice shows that only one-third of GA are compact and normally placed (D3) and the other GA are either displaced (D4), fragmented (D5), or both (D6). In panel (D6), white asterisk corresponds to a GA belonging to a different cell.
- E Quantification of the morphology and the relative localization of the GA and Dpy19l2 staining in WT (n = 40) and *Spink2*^{-/-} (n = 39) round spermatids.

Disjunction of the Golgi apparatus and of the NDL was also observed in EM (Fig EV5A). Moreover, anti-GM130 staining in *Spink2*^{-/-} round spermatids appeared disseminated and punctuated, confirming the disorganization of the Golgi apparatus and indicating a fragmentation of the organelle (Fig 4D4–D6).

Interestingly, EM observations of *Spink2*^{-/-} round spermatids showed the presence of multivesicular bodies, a known biomarker of microautophagy (Li *et al.*, 2012) (Fig EV5B). These latter structures strongly suggest that the absence of Spink2 activates an uncharacterized self-degradation pathway. Visual signs of the initial events of microautophagy occurring at the Golgi apparatus level are the engulfment of vesicles (Fig 4B2, black arrows) and the presence of already engulfed vesicles (Fig 4B2, black arrowhead). We note that the thorough examination of round spermatids on EM images did not reveal any detectable signs of morphological hallmarks of apoptosis such as chromatin condensation, fragmentation of the plasma membrane, and the presence of apoptotic bodies. Moreover, no differences in DNA damage were observed between WT and *Spink2*^{-/-} round spermatids when assessed by terminal deoxynucleotidyl transferase (TdT)-mediated deoxyuridine triphosphate (dUTP)-nick-end labeling (TUNEL) test (Appendix Fig S5). Altogether, these results suggest that the absence of Spink2 at the round-spermatid stage does not activate the apoptotic pathway.

Rescue of acrosin-induced cell proliferation defects by coexpression with SPINK2

During spermatid differentiation, several enzymes, involved in sperm penetration through the protective layers surrounding the oocytes, accumulate in the acrosomal vesicle. Among these different enzymes, several proteases have been described to play a key role, including acrosin, believed to be the main acrosomal protease (Liu & Baker, 1993). Acrosin, a trypsin-like protease, is synthesized in the reticulum as a zymogen (proacrosin), transits through the Golgi apparatus, and accumulates in the acrosomal vesicle. Autoactivation of acrosin is pH-dependent and occurs at a pH > 6 (Meizel & Deamer, 1978) leading to sequential N-ter and C-ter cleavages of the proacrosin (46 kDa), eventually giving active forms of acrosin with lower weights of 20–34 kDa (Baba *et al.*, 1989; Zahn *et al.*, 2002). Since the pH of both the endoplasmic reticulum and the Golgi apparatus is greater than 6 (Rivinoja *et al.*, 2012), we postulated that

Spink2, as a serine peptidase inhibitor, prevents acrosin autoactivation in these cellular compartments, thus preventing cellular stress induced by uncontrolled protease activation. Such stress would cause cellular defects including Golgi apparatus destabilization and defective acrosome biogenesis leading to spermatid differentiation arrest. To test this hypothesis, heterologous expressions of human C-terminus DDK-tagged proacrosin (ACR), SPINK2, or both were carried out in HEK293 cells and the kinetics of cell proliferation were followed using xCELLigence Real-Time Cell Analysis (RTCA) technology for the different conditions. It is worth noting that no members of the SPINK family are reported to be expressed in HEK293 cells. Analyses of kinetics showed that proacrosin expression quickly led to cell proliferation arrest and detachments in contrast to what was observed in the control condition (Fig 5A and B). Interestingly, cells showed a normal proliferation when SPINK2 was coexpressed with proacrosin (Fig 5A and B), therefore demonstrating that cell stress and damages induced by the proacrosin were prevented by SPINK2 coexpression. The presence of the different overexpressed proteins was verified in the different conditions by Western blotting using the SPINK2 antibody (Fig 5C), an anti-acrosin (Fig 5D), and the anti-DDK (Fig 5E) antibodies. In extracts of HEK293 cells transfected with proacrosin only and revealed with an anti-acrosin antibody (Fig 5D), two bands were present at around 46 and 34 kDa. The latter (red arrowhead) likely corresponds to the active form of acrosin resulting from the cleavage of proacrosin upon autoactivation. This band was not present when acrosin was coexpressed with SPINK2 or in non-transfected cells (control). Moreover, a closer inspection of the band around 46 kDa in the extracts of cells transfected with proacrosin only or proacrosin + SPINK2 shows that this band is of lower MW and was less intense in “acrosin” extract compared to “acrosin + SPINK2” cell extract, showing the process of successive cleavages occurring during proacrosin autoactivation (Zahn *et al.*, 2002). Similar results were obtained with the anti-DDK antibody (Fig 5E). It is worth noting that anti-DDK antibody immunodecorates the zymogen form only and not the active form of acrosin because the C-terminus containing the DDK-tag is cleaved upon autoactivation. Western blot results thus demonstrate that coexpression of proacrosin with SPINK2 prevented its autoactivation. We can thus conclude that in the absence of a serine peptidase inhibitor, proacrosin can autoactivate and induces a cellular stress leading to

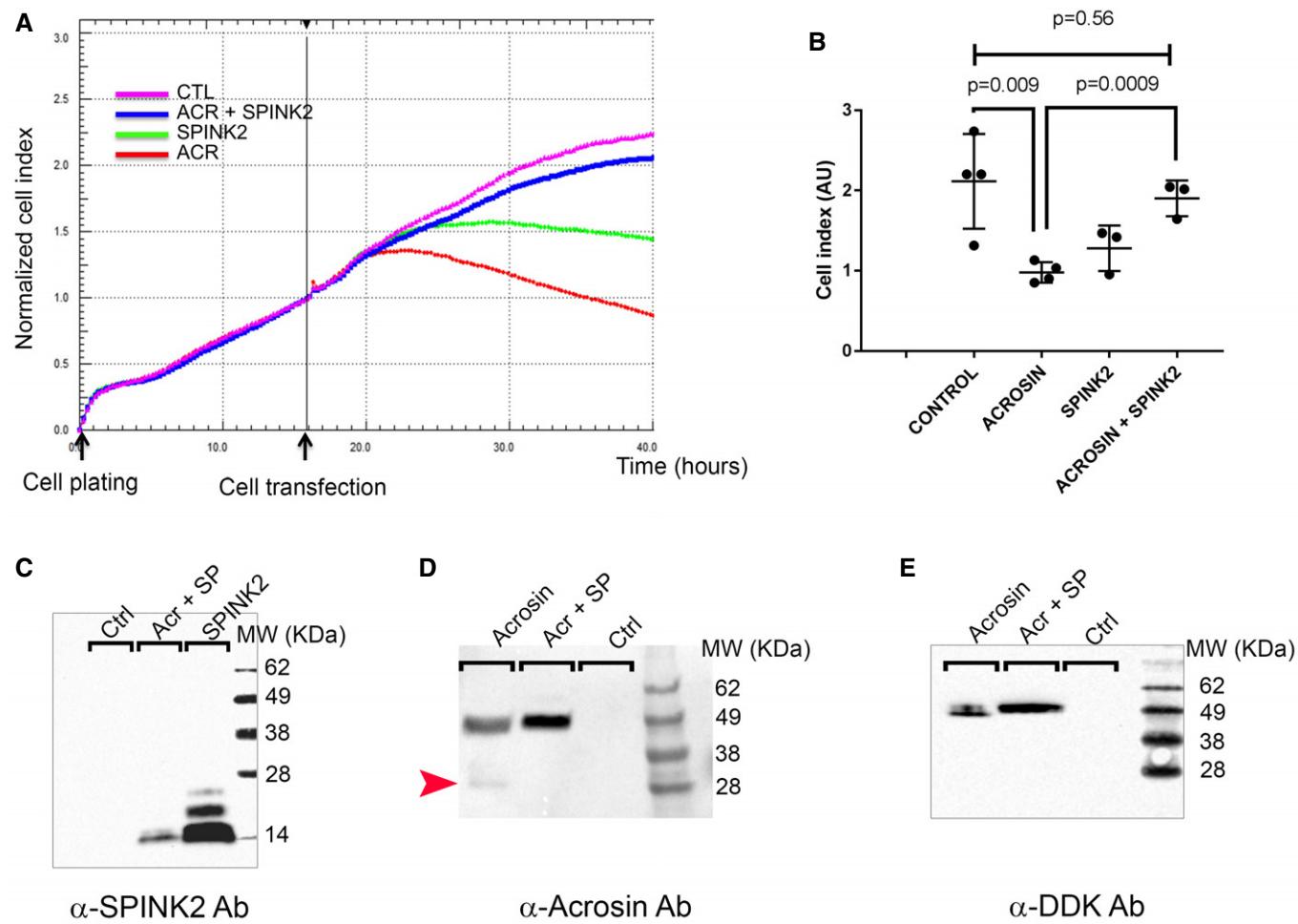


Figure 5. Heterologous expression of proacrosin in HEK293 cells induces acrosin activation and cell proliferation arrest, a phenotype rescued by SPINK2 coexpression.

- A Representative kinetics of HEK293 cell proliferation measured with Real-Time Cell Analysis (RTCA) technology in different conditions as indicated. Each point corresponds to the mean of four technical replicates measured simultaneously. Black arrows indicate the time of cell plating ($t = 0$ h) and introduction of the different plasmids in the cell chambers ($t = 16$ h).
- B Scatter plots showing the mean and SD of the cell index measured at 40 h after plating (corresponding to cell proliferation and detachment) in different transfection conditions and measured for three independent biological replicates. Statistical differences were assessed using *t*-test, *P*-values as indicated.
- C Western blot using an anti-SPINK2 antibody showing the expression of SPINK2 in cell extracts of HEK293 cells transfected with different plasmids containing SPINK2 (SP) or acrosin and SPINK2.
- D Representative Western blot using an anti-acrosin antibody. In extracts of HEK293 cells transfected with proacrosin only (lane "acrosin"), two bands were observed, one at around 34 kDa and corresponding to the active form of acrosin (red arrowhead) and one at 46 kDa and corresponding to the zymogen form, whereas in extracts of HEK293 cells transfected with proacrosin and SPINK2 (lane "Acr + SP"), only the zymogen form was observed. Equal protein loading was verified by stain-free gel technology (Taylor & Posch, 2014) and Western blots against tubulin (Appendix Fig S6). Note that the zymogen form in lane "acrosin" has a slightly lower mass and that the band is less intense than that in lane "Acr + SP".
- E Representative Western blot using an anti-DDK antibody showing the expression of the proacrosin zymogen form in HEK293 cells transfected with different plasmids as indicated. Note that once more, the zymogen form in the lane "acrosin" has a slightly lower mass and that the band is less intense than that in lane "Acr + SP". Similarly, equal protein loading was verified by stain-free gel technology (Appendix Fig S6).

cell proliferation arrest and cell detachment, a phenotype similar to that observed in round spermatids from *Spink2*^{-/-} males.

SPINK2 haploinsufficiency induces sperm defects with incomplete penetrance in man

Only one additional subject, P105, was identified with a *SPINK2* heterozygous deleterious variant, and we cannot be sure that this

variant is the cause of the patient's oligozoospermia. Two arguments could in fact suggest that *SPINK2* haploinsufficiency is not deleterious: (i) Br1 and Br2's father is *SPINK2* heterozygous and has conceived six children spontaneously, and unfortunately, we could not obtain sperm samples to characterize this man's sperm parameters; and (ii) because heterozygous *Spink2*^{+/−} male mice are fertile, they did not produce litters of reduced size (Fig 3A). We however carried out a detailed characterization of *Spink2*^{+/−} and *Spink2*⁺⁺

sperm parameters to address the question of the impact of SPINK2 haploinsufficiency on mouse spermatogenesis. Heterozygous males displayed a significant increase in teratozoospermia (Fig 6A). Abnormal spermatozoa showed non-hooked heads, isolated heads, or a malformed base of the head (Fig 6B). Moreover, sperm motility of heterozygous males was impaired with lower total and progressive motility (Fig 6C). We note that the observed defects are very similar to those observed in the heterozygous patient P105 (Table EV1). We can therefore conclude that in mice, *SPINK2* haploinsufficiency induces asthenoteratospermia with no alteration of reproductive fitness, whereas in man it leads to oligoteratozoospermia with variable expressivity and infertility with an incomplete penetrance.

Discussion

SPINK family emerges as an important family for human genetic diseases

SPINK proteins are serine protease inhibitors containing one or several Kazal domains which interact directly with the catalytic domains of proteases blocking their enzymatic activity (Rawlings *et al.*, 2004). The Kazal domain structure contains three disulfide bonds which are highly conserved. Different SPINK proteins are specifically expressed in different tissues and inhibit a number of serine proteases, such as secreted trypsin in the pancreas, acrosin in sperm, or kallikrein in the skin. Downregulation of the activity of different SPINK proteins leads to severe pathologies such as chronic pancreatitis and Netherton syndrome. In the pancreas, trypsin is produced as an inactive zymogen to prevent cell damage, yet the trypsinogen is occasionally able to autoactivate. This protease activity is then blocked by SPINK1. Chronic pancreatitis can be triggered by mutations of *SPINK1* that decrease or suppress its trypsin inhibitor function, leading to cell distress (Chen *et al.*, 2000; Witt *et al.*, 2000). In the skin, kallikrein-related peptidases are controlled by SPINK5 and unopposed kallikrein-peptidase activity due to *SPINK5* deficiency leads to Netherton syndrome, a severe skin disease (Furio & Hovnanian, 2014). SPINK6 and SPINK9 are also expressed in the skin, and altered expression levels are associated with atopic dermatitis or psoriasis (Redelfs *et al.*, 2016). The other members of the SPINK family, including SPINK2, have not yet been associated with a human pathology. Here, we have clearly demonstrated that the absence of SPINK2 induces azoospermia, a severe infertility phenotype, emphasizing the importance of this family in human pathologies.

Role of SPINK2 during spermiogenesis

We have shown that SPINK2 is located in the acrosomal vesicle in round spermatids and remains present in mature spermatozoa, suggesting that this protein is necessary for spermiogenesis and sperm survival. SPINK proteins are known to control protease activities in different tissues (Witt *et al.*, 2000; Rawlings *et al.*, 2004; Ohmuraya *et al.*, 2012; Furio & Hovnanian, 2014) and since SPINK2 is located in the acrosome, it very likely neutralize acrosomal proteases before their release prior fertilization. Several proteases have been described to be present in the acrosome (Arboleda & Gerton, 1987; Kohno *et al.*, 1998; Cesari *et al.*, 2004). Among these,

acrosin (Acr) was the first to be described and is the acrosomal protein which has been the most studied. Acrosin is present in the acrosome as a zymogen called proacrosin (Huang-Yang & Meizel, 1975) which is predicted to be activated during the acrosome reaction (Brown & Harrison, 1978) upon a rise in acrosomal pH to 7 which induces pH-dependent proacrosin autoactivation (Baba *et al.*, 1989). Before the acrosome reaction, at least two mechanisms prevent autoactivation: The first is the acrosomal acidic pH which is below 5, which blocks autoactivation of proacrosin (Meizel & Deamer, 1978); and the second is the presence in the sperm of a non-fully characterized proacrosin conversion inhibitor of 12 kDa which has been purified from boar acrosome (Kennedy *et al.*, 1982). The presented results strongly suggest that this protein is in fact SPINK2. Proacrosin is however produced in the endoplasmic reticulum and transits through the Golgi apparatus, two cellular compartments with a pH of approximately 7 and 6.5, respectively. In these compartments, autoactivation of proacrosin is thus possible and would result in the release of active acrosin within these apparatuses. We therefore believe that SPINK2, which transits through the same cellular compartments, quenches this premature protease activity and prevents the described cascade of events leading to azoospermia. This hypothesis is supported by heterologous expression experiments: We have indeed demonstrated that proacrosin expression in HEK293 cells induces (i) autoactivation of proacrosin and (ii) cellular proliferation arrest and cell detachment. Moreover, cellular toxicity of proacrosin expression is prevented by SPINK2 coexpression, showing the ability of SPINK2 to inhibit acrosin activity.

One of the most striking effects of SPINK2 deficiency is the fragmentation of the Golgi apparatus, a key organelle for protein processing and translocation, in particular for membrane proteins. The notable strong desquamation of the germinal epithelium may be due to severe changes in membrane protein composition resulting from a defective Golgi apparatus function.

Impact of SPINK2 deficiency

We have shown that the absence of SPINK2 in round spermatids leads to several subcellular defects targeting the process of proacrosomal vesicle formation by the Golgi apparatus. The observed abnormalities include the disorganization and delocalization of the Golgi apparatus, the presence of vesicles of various sizes, and the absence of proacrosomal vesicle fusion. The absence of SPINK2 likely allows proacrosin autoactivation within the reticulum and the Golgi apparatus compartments, leading to the above-described subcellular defects. It was previously shown that transgenic expression of porcine proacrosin in mice led to post-meiotic cell death and oligozoospermia, supporting the hypothesis that unbalanced expression of acrosin/Spink2 is deleterious (O'Brien *et al.*, 1996). Interestingly, we have demonstrated that the cell responds to this stress by activating a microautophagy-like pathway: First, we showed that larger vacuoles engulfed small vacuoles, likely leading to the observed heterogeneity in vacuole size in the vicinity of the Golgi apparatus; and secondly, multivesicular bodies, a hallmark of microautophagy (Li *et al.*, 2012), were clearly observed within *Spink2*^{-/-} round spermatids, whereas they were never observed in WT. Furthermore, the lack of various SPINK proteins induces autophagy-induced cell death in regenerating Hydra (Chera *et al.*, 2009)

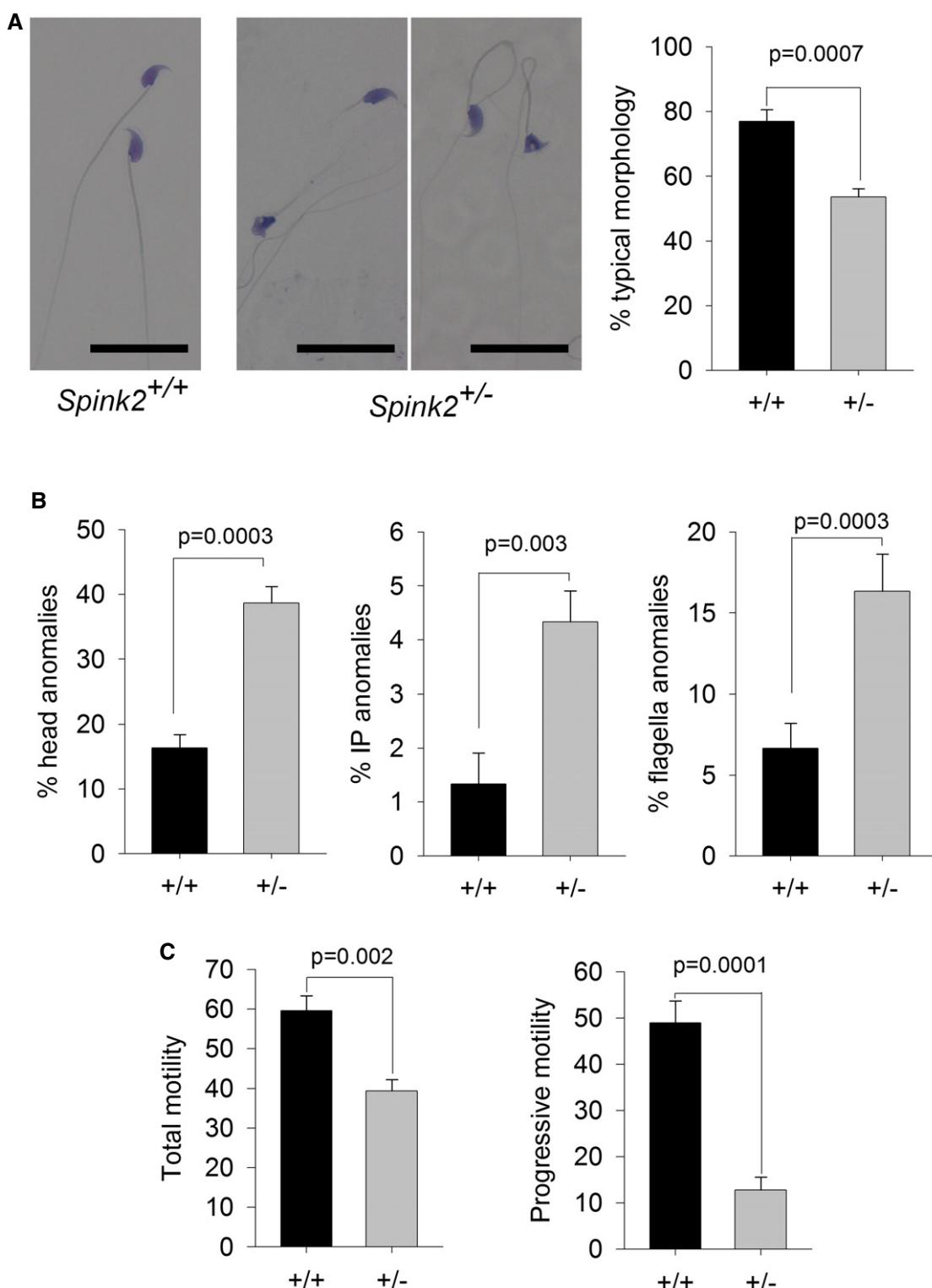


Figure 6. Sperm from $Spink2^{+/-}$ heterozygous mice exhibit morphological defects and low motility.

- A Light microscopy analysis of sperm from $Spink2^{+/-}$ heterozygous mice reveals the presence of numerous non-typical forms of sperm. Scale bars, 25 μ m. Graph on the right shows the mean \pm SD percentage of defective sperm in WT ($n = 3$) and $Spink2^{+/-}$ mice ($n = 3$).
- B Anomalies were observed in the head and the mid- and principle pieces in WT and $Spink2^{+/-}$ mice ($n = 3$).
- C Total and progressive sperm motility were strongly decreased in $Spink2^{+/-}$ heterozygous mice ($n = 5$) in comparison with WT sperm ($n = 5$).

Data information: n represents the number of biological replicates, and for each replicate, more than 100 sperm were assessed per condition. Data are presented as mean \pm SD. Statistical differences were assessed using t-test, P-values as indicated.

and was also described in newborn mice when *Spink3* (orthologue of *Spink1*) is mutated (Ohmuraya *et al.*, 2005). Based on these results, it has been postulated that SPINK1/Spink3 could have the dual function of protease inhibitor and negative regulator of autophagy (Ohmuraya *et al.*, 2012). Our results show that the absence of SPINK2 induces a microautophagy-like pathway in germ cells thereby further supporting this hypothesis.

Oligozoospermia and azoospermia is a continuum correlated with SPINK2 haploinsufficiency

We observed that in man, the presence of a homozygous *SPINK2* mutation leads to azoospermia while a heterozygous mutation can induce oligozoospermia suggesting that *SPINK2* haploinsufficiency can result in oligozoospermia. In mice, we showed that the complete absence of the protein leads to azoospermia. We also showed that heterozygous animals have terato-astheno-zoospermia but with no obvious decrease in sperm number and no impact on fertility. A previous study carried out in a different mouse hypomorphic mutant line showed that a significant inactivation of *Spink2* (likely in excess of 90%) led to a reduction by half of sperm number within the epididymis and a five-fold increase in morphologically abnormal spermatozoa. Male mice also exhibited a reduced fertility and produced litters of reduced size with an average of 5.19 pups by litter compared to 8.56 in controls (Lee *et al.*, 2011). These results and the results presented therein thus confirm that in mice, the severity of the phenotype is dependent on *Spink2* expression levels and that there is a phenotypic continuum ranging from (i) azoospermia in the complete absence of the protein (ii) to teratozoospermia and oligozoospermia associated with subfertility when only a fraction of the protein is present and finally (iii) to astheno-teratozoospermia with no impact on fertility when half of the protein is present. These observations in mice strongly support the notion that *SPINK2* heterozygous mutations in man will impact spermatogenesis with a variable effect on fertility. We identified only one heterozygous mutation out of 611 analyzed patients indicating that *SPINK2* variants are very rare, likely because heterozygous variants underwent a strong negative selection during evolution. This hypothesis is supported by data from the ExAC database which indicate that the *SPINK2* gene has a high probability of loss of function intolerance (pLI = 0.72).

The testis is the organ which expresses the highest number of tissue-specific transcripts ($n > 500$) (Feig *et al.*, 2007; Dezso *et al.*, 2008) and altered spermatogenesis has been observed in knockout mouse models for more than 388 genes (Massart *et al.*, 2012). It is therefore expected that NOA is genetically highly heterogeneous and that few patients carry causal defects on the same gene. Due to the involvement of the corresponding proteins in multiple phases of spermatogenesis, the causes of azoospermia are numerous and involve genes controlling spermatogonial self-renewal, meiosis, and spermiogenesis. Here, we have confirmed that alterations of spermiogenesis do not only lead to teratozoospermia as described several times previously (Dieterich *et al.*, 2007; Harbuz *et al.*, 2011; Ben Khelifa *et al.*, 2014) but also to azoospermia. The vast majority of patients with an altered spermatogenesis can be treated with IVF or by the direct injection of a sperm into the oocyte (ICSI). Most patients with NOA however cannot benefit from ICSI-IVF treatments. Identifying the genetic defects responsible for NOA and characterizing their molecular pathogeny will provide a basis for the

development of therapeutic solutions tailored to the patient. In this particular case, we have shown that SPINK2 deficiency can induce azoospermia and demonstrated that unrestricted acrosomal protease activity induces the arrest of spermiogenesis. Moreover, we provided evidence that this process activates a microautophagy-like pathway. As we have shown that the pool of undifferentiated spermatogonia is not affected, we can envisage a method of treatment targeting protease activity using a protease inhibitor, as is done for chronic pancreatitis caused by *SPINK1* deficiency (Kambhampati *et al.*, 2014).

Materials and Methods

Patients and biological samples

Human sperm were obtained from patients consulting for diagnosis or assisted reproductive techniques at the fertility center of the Grenoble University Hospital. All patients signed an informed consent for use of part of their samples in research programs respecting the WMA declaration of Helsinki. The samples were then stored in the CRB Germethéque (certification under ISO-9001 and NF-S 96-900) following a standardized procedure. Consent for CRB storage was approved by the CPP Sud-Ouest of Toulouse (coordination of the multisite CRB Germethéque). The storage and transfer authorization number for the CRB Germethéque is AC2009-886. The scientific and ethical board of the CRB Germethéque approved the transfer of the semen samples for this study. Additional DNA samples from patients with azoospermia and oligozoospermia were obtained from the CHU of Grenoble, Saint Etienne, and Marseille. All patients gave their informed consent for the anonymous use of their leftover samples. Brothers Br1 and Br2 are French citizens from a traveling group originating from Romania but whose recent ancestors lived in Spain and the south of France. Subject P105 is also a French citizen with eastern ascendants (from Russia).

Exome sequencing and bioinformatic analysis

Genomic DNA was isolated from saliva using Oragene saliva DNA collection kit (DNA Genotek Inc., Ottawa, Canada). Exome capture was performed using NimbleGen SeqCap EZ Kit version 2 (Roche). Sequencing was conducted on an Illumina HiSeq 2000 instrument with paired-end 76-nt reads. Sequence reads were aligned to the reference genome (hg19) using MAGIC (SEQC/MAQC-III Consortium, 2014). Duplicate reads and reads that mapped to multiple locations in the exome were excluded from further analysis. Positions with sequence coverage below 10 on either the forward or reverse strand were excluded. Single nucleotide variations (SNV) and small insertions/deletions (indels) were identified and quality-filtered using in-house scripts. The most promising candidate variants were identified using an in-house bioinformatics pipeline. Variants with a minor allele frequency > 5% in the NHLBI ESP6500 or in 1000 Genomes Project phase 1 datasets, or > 1% in ExAC, were discarded. We also compared these variants to an in-house database of 56 control exomes. All variants present in a homozygous state in this database were excluded. We used Variant Effect Predictor (VEP) to predict the impact of the selected variants. We only retained variants impacting splice donor/acceptor sites or

causing frameshift, in-frame insertions/deletions, stop gain, stop loss, or missense variants except those scored as “tolerated” by SIFT (sift.jcvi.org) and as “benign” by PolyPhen-2 (genetics.bwh.harvard.edu/pph2). All steps from sequence mapping to variant selection were performed using the ExSQLibur pipeline (<https://github.com/tkaraouzene/ExSQLibur>). Our datasets were obtained from subjects who have consented to the use of their individual genetic data for biomedical research, but not for unlimited public data release. Therefore, we submitted it to the European Genome-phenome Archive, through which researchers can apply for access of the raw data under the accession number EGAD00001003326.

Sanger sequencing

Sanger sequencing of the four *SPINK2* exons and intron borders was carried out using the primers described in Appendix Table S3. Thirty-five cycles of PCR amplification were carried out with a hybridization temperature of 60°C. Sequencing reactions were performed using BigDye Terminator v3.1 (Applied Biosystems). Sequence analyses were carried out on ABI 3130XL (Applied Biosystems). Sequences were analyzed using seqscape software (Applied Biosystems).

RT-PCR and quantitative real-time PCR

Total RNA from various tissues including testes from three WT and homozygous KO mice was extracted using the mirVana™ PARIS™ Kit (Life Technologies®) according to the manufacturer’s protocol. Human cDNAs were obtained from Life Technologies® mRNA.

Reverse transcription was carried out with 5 µl of extracted RNA (~500 ng). Hybridization of the oligo dT was performed by incubating for 5 min at 65°C and quenching on ice with the following mix: 5 µl RNA, 3 µl of poly-T-oligo primers (dT) 12–18 (10 mM; Pharmacia), 3 µl of the four dNTPs (0.5 mM, Roche Diagnostics) and 2.2 µl of H₂O. Reverse transcription then was carried out for 30 min at 55°C after the addition of 4 µl of 5× buffer, 0.5 µl RNase inhibitor, and 0.5 µl Transcripter reverse transcriptase (Roche Diagnostics). One microliter of the obtained cDNA mix was used for the subsequent PCR. Primers are described in Appendix Table S4.

A specific region of the transcript was amplified using a StepOne-Plus™ Real-Time PCR System (Life Technologies®) with Power SYBR® Green PCR Master Mix (Life Technologies®) according to the manufacturer’s protocol. PCR without template was used as a negative control to verify experimental results. The sequence for oligonucleotide primers used and their product sizes are summarized in Appendix Table S5.

After amplification, the specificity of the PCR was determined by both melt-curve analysis and gel electrophoresis to verify that only a single product of the correct size was present. Quantification of the fold change in gene expression was determined by the relative quantification method ($2^{-\Delta\Delta C_T}$) using the beta-actin gene as a reference. Data are shown as the average fold increase ± standard error of the mean.

Primary antibodies

SPINK2 rabbit polyclonal antibody was from Sigma-Aldrich (HPA026813) and used at 1/1,000 for Western blot analysis. Sperm

protein Sp56 and Golgi matrix protein GM130 (610822) mouse monoclonal antibodies were from QED Bioscience Inc. (used at 1/800 and 1/200, respectively). Promyelocytic leukemia zinc finger protein PLZF rabbit polyclonal antibody (Sc-22839) was from Santa Cruz Biotechnology. Dpy19l2 antibodies were produced in rabbit as polyclonal antibodies raised against RSKLREGSSDRPQSSC and CTGQARRRWSAATMEP peptides corresponding to amino acids 6–21 and 21–36 of the N-terminus of mouse Dpy19l2 (Pierre et al, 2012). DDK antibody was from OriGene (TA50011) or Sigma-Aldrich (FLAG® M2F1804) and used at 1/10,000 for Western blot analysis. Acrosin antibody was previously described (Gallo et al, 1991) and is a gift from Denise Escalier.

Western blot analysis

HEK293 cells were lysed in 25 mM Tris pH 7.4, 5 mM EDTA, 1% Triton X-100, and complete protease inhibitor cocktail (Roche) and were then centrifuged. After centrifugation at 20,000 g for 15 min at 4°C, the soluble supernatant was conserved and subjected to SDS-PAGE. The protein concentration from supernatants was quantified by the bicinchoninic acid assay (BCA assay) using bovine serum albumin as a standard. Sample concentrations were adjusted and mixed with 1× high-SDS sample buffer (4% SDS, 62 mM Tris-HCl pH 6.8, 0.1% bromophenol blue, 15% glycerol, 5% β-mercaptoethanol) and separated using 4–20% SDS mini-PROTEAN® TGX Stain-Free™ Precast Gels (Bio-Rad) or 10 and 20% polyacrylamide-SDS gels and transferred into PVDF membranes (Millipore, 0.2 µm) using Trans-Blot® Turbo™ Blotting System and Midi Transfer Packs (Bio-Rad). The membranes were blocked in 5% non-fat dry milk in PBS/0.1% Tween and incubated for 1 h at room temperature with the primary antibody, followed by 45-min incubation with a species-matched horseradish peroxidase-labeled secondary antibody (1/10,000) (Jackson ImmunoResearch). Immunoreactivity was detected using chemiluminescence detection kit reagents (Luminata; Millipore) and a ChemiDoc Station (Bio-Rad).

Real-time cell analysis

The growth, proliferation, and adhesion kinetics of HEK293 cells were determined using RTCA technology (ACEA Biosciences, San Diego, CA, USA). Fifty microliters of DMEM supplemented with 10% HI-FBS and 50 µg/ml gentamicin (cell culture medium) was loaded in each well of the E-plate 96 (gold-microelectrode array integrated E-plate; ACEA Biosciences). E-plate 96 was then connected to the system to obtain background impedance readings. Around 1.5×10^4 cells in 50 µl were added to the wells containing 50 µl of culture medium. The E-plates were placed on the RTCA SP Station located in a 37°C, 5% CO₂ tissue culture incubator for continuous impedance recording. The cell index values measured by continuous impedance recordings every 5 min are proportional to the number of adherent cells. After 16–17 h, cells were transfected as described below, and for each of the conditions, four replicates were done. The assay was conducted for 40 h.

Mice

All animal procedures were run according to the French guidelines on the use of animals in scientific investigations with the approval of

the local ethical committee (Grenoble-Institut des Neurosciences—ethical committee, study agreement number 004). Mice were euthanized by cervical dislocation.

The *Spink2tm1.1* (KOMP)Vlcg mouse strain used for this research project was created from ES cell clone *Spink2_AG5_M7*, generated by Regeneron Pharmaceuticals, Inc. and made into live mice by the KOMP Repository (www.komp.org) and the Mouse Biology Program (www.mousebiology.org) at the University of California Davis. The methods used to create the VelociGene targeted alleles have been published (Valenzuela *et al.*, 2003). They were then reared by the Mouse Clinical Institute—MCI—located in Strasbourg as part of the “knockout mouse project”. The colony used in this study was initiated from two couples consisting of heterozygous females and males. Mice were housed with unlimited access to food and water and were sacrificed after 8 weeks of age (the age of sexual maturity).

Genotyping

DNA for genotyping was isolated from tail biopsies. Tail biopsies (2 mm in length) were digested in 200 µl of DirectPCR Lysis Reagent (Tail) (Viagen Biotech Inc, CA, USA) and 0.2 mg of proteinase K for 12–15 h at 55°C followed by 1 h at 85°C for proteinase K inactivation. The DNA was directly used for PCRs. Multiplex PCR was done for 35 cycles, with an annealing temperature of 58°C, and an elongation time of 60 s at 72°C. PCR products were separated by 2% agarose gel electrophoresis. Genotypes were determined according to the migration pattern. Primers are described in Appendix Table S6.

Phenotypic analysis of mutant mice

To test fertility, pubescent *Spink2*^{-/-} males (8-week-old) were mated with WT females.

To determine sperm concentration, sperm samples were collected from the cauda epididymis and vas deferens of 8-week-old males, and sperm number was determined using a hemocytometer under a light microscope.

Sperm motility analysis

Experiments were performed on a CASA CEROS v.12 (Hamilton Thorne Biosciences, Beverly, MA, USA) using Leja double-chamber slides (Leja Products B.V., the Netherlands) for standard count with 100 µm depth. After epididymal extraction, sperm cells were allowed to swim for 10 min at 37°C and then were immediately analyzed. At least 150 cells were analyzed per sample with the following parameters: acquisition rate: 60 Hz; number of frames: 45; minimum contrast: 50; minimum cell size: 5; low static-size gate: 0.3; high static-size gate: 1.95; low static-intensity gate: 0.5; high static-intensity gate: 1.3; minimum elongation gate: 0; maximum elongation gate: 87; and magnification factor: 0.7. The motility parameters measured were curvilinear velocity (VCL), straight-line velocity (VSL), average path velocity (VAP), and amplitude of lateral head displacement (ALH). Motile sperm were defined by VAP > 1 and progressive sperm were defined by VAP > 30 and VSL/VAP > 0.7.

Histological analysis

To analyze testicular integrity, testes from adult *Spink2*^{+/+} and *Spink2*^{-/-} mice were fixed by immersion in 4% paraformaldehyde (PFA) for 14 h, embedded in paraffin, and sectioned (4 µm). For histological analysis, after being deparaffinized slides were stained with hematoxylin and eosin or by the PAS technique. The colored sections were digitized at ×40 magnification through an Axioscope microscope (Zeiss, Germany) equipped with a motorized X-Y-sensitive stage. For sperm morphology analysis, sperm were washed twice in PBS and then displayed over slides, dried at room temperature, and then fixed in 75% ethanol for Harris–Shorr staining. At least 100 sperm cells were analyzed per sample.

Testicular germ cell dissociation

C57BL/6 male or *Spink2* KO mice (8-week-old) were euthanized by cervical dislocation. The testes were surgically removed and placed in PBS (at room temperature). The tunica albuginea was removed from the testes with sterile forceps and discarded. Then, the testes were incubated in 1 mg/ml of collagenase solution in EKRB cell buffer containing (in mM) 2 CaCl₂, 12.1 glucose, 10 HEPES, 5 KCl, 1 MgCl₂, 6 Na-lactate, 150 NaCl, 1 NaH₂PO₄, and 12 NaHCO₃ pH 7, and agitated horizontally at a maximum of 120 rpm for 30 min at 25°C. The dispersed seminiferous tubules were then washed with PBS and cut thinly. Cells were dissociated by gentle pipetting filtered through a 100-µm filter and then pelleted by centrifugation at 500 g for 7 min. Cells were resuspended in 1 ml PBS, fixed with 4% PFA solution, washed with PBS, and finally layered onto polylysine-coated slides.

Immunohistochemistry

Mice were anesthetized by intraperitoneal injection of a ketamine/xylazine cocktail (87.5 mg/kg ketamine and 12.5 mg/kg xylazine) and sacrificed through intracardiac perfusion of PFA (4%). The testes and epididymides were removed and fixed for a further 8 h before paraffin embedding and sectioning. Mature sperm cells were obtained for analysis through mechanical dilaceration of the epididymis. Sperm cells were fixed in 4% PFA for 1 min and washed in PBS before being spotted onto poly-L-lysine-pre-coated slides. Spermatogenic cells of the round-spermatid stage were purified by unit gravity sedimentation from a spermatogenic cell suspension obtained from sexually mature males as described in Yassine *et al.* (2015).

For immunofluorescence experiments, heat-induced antigen retrieval was performed by boiling slides immersed in either 0.01 M sodium citrate buffer–0.05% Tween-20, pH 6.0, or 10 mM Tris base–1 mM EDTA solution–0.05% Tween-20, pH 9.0, for 15–25 min. Sections were blocked in 2% goat serum–0.1% Triton X-100 for 1 h at RT and incubated with primary antibodies overnight at 4°C. The slides were then washed and incubated with secondary antibody (DyLight 549-conjugated goat anti-mouse IgG or DyLight 488-conjugated goat anti-rabbit IgG, Jackson ImmunoResearch) and Hoechst 33342 for 2 h at RT, rinsed, and mounted with Dako mounting medium (Life Technology). Images were taken by confocal microscopy (Zeiss LSM 710) and processed using Zen 2009 software.

Electron microscopy (EM)

Adult male mice were anesthetized and fixed by intracardiac injection with 2% glutaraldehyde and 2.5% PFA in 0.1 M cacodylate, pH 7.2. For morphological analysis, samples were fixed with 2.5% glutaraldehyde in 0.1 M cacodylate buffer pH 7.4 over 24 h at room temperature. Samples were then washed with buffer and post-fixed with 1% osmium tetroxide and 0.1 M cacodylate pH 7.2 for 1 h at 4°C. After extensive washing with water, cells were further stained with 1% uranyl acetate pH 4 in water for 1 h at 4°C before being dehydrated through graded alcohol (30%–60%–90%–100%–100%) and infiltrate with a mix of 1/1 epon/alcohol 100% for 1 h and several baths of fresh epon (Flukka) during 3 h. Finally, samples were embedded in a capsule full of resin that was left to polymerize over 72 h at 60°C. Ultrathin sections of the samples were cut with an ultramicrotome (Leica), and the sections were post-stained with 5% uranyl acetate and 0.4% lead citrate before being observed with an electron microscope at 80 kV (JEOL 1200EX). Images were acquired with a digital camera (Veleta; SIS, Olympus), and morphometric analysis was performed with iTEM software (Olympus).

Cell culture and transfection

Mycoplasma-free HEK293 cells were a gift from A. Andrieux from Grenoble Neuroscience Institute and grown in Dulbecco's modified Eagle's medium supplemented with 10% FBS (Invitrogen, France) and 50 µg/ml gentamicin (Sigma) in a 37°C, 5% CO₂ cell culture incubator and transiently transfected with Cter-DDK-tagged human acrosin (RC214256; OriGene, Rockville, MD, USA) and/or human SPINK2 (RC205388; OriGene) and/or human c.1A>T mutated SPINK2-containing pCMV6 plasmids, using JetPRIME Transfection Reagent (Polyplus, France) according to the manufacturer's instructions. For immunochemistry experiments, transfected cells were fixed with 4% PFA 2 days after transfection.

DNA strand breaks

Sections were permeabilized using a 0.1% (v/v) Triton X-100 and 0.1% (w/v) sodium citrate in 1× PBS for 2 min and labeled by terminal deoxynucleotidyl transferase-mediated deoxy-UTP nick-end labeling (TUNEL) according to the Roche protocol of the *In Situ* Cell Detection Kit (Roche Diagnostics, Mannheim, Germany). Nuclei were counterstained in a 0.5 µg/ml Hoechst solution for 3 min, washed in PBS for 3 min, and mounted with DAKO mounting medium.

Statistical analyses

n represents the number of biological replicates. For sperm analyses, for each replicate, more than 100 sperm were assessed per condition. Statistical analyses were performed with SigmaPlot 10 and GraphPad Prism 7. *t*-Tests were used to compare WT and KO samples. Data represent mean ± SEM or SD, as indicated. Statistical tests with a two-tailed *P*-value ≤ 0.05 were considered significant.

Expanded View for this article is available online.

The paper explained

Problem

Infertility concerns one in seven couples and is usually addressed by performing *in vitro* fertilization (IVF) often by injecting spermatozoa directly into the oocytes by intracytoplasmic sperm injection (ICSI). Some men have a non-obstructive azoospermia (NOA), caused by a deficient spermatogenesis, and have no spermatozoa in the ejaculate. In some cases, a testicular biopsy can be performed in hope of finding some mature spermatozoa that will be used for ICSI, but most men with NOA will not be able to have biological children. It is believed that most cases of NOA are caused by a genetic factor, but a diagnosis is obtained for only approximately 20% of patients.

Results

We performed exome sequencing on two brothers with NOA and identified a homozygous mutation in the SPINK2 gene coding for a serine protease inhibitor believed to target the acrosin, the main protease of the acrosome, a large vesicle located to the anterior part of the spermatozoa and containing an enzyme mix necessary to perforate the zona pellucida of the oocyte to achieve fertilization. Mouse study allowed to observe that homozygous KO male also had NOA, confirming the human diagnostic. Germ cells could go through meiosis but were blocked at the round-spermatid stage. We further observed that in the round spermatids, in the absence of SPINK2, the acrosin could autoactivate during its transit through the endoplasmic reticulum and the Golgi apparatus leading to a disorganization of the Golgi and its inability to form the acrosome and a block at the round-spermatid stage. We further demonstrate that the presence of a heterozygous SPINK2 mutation was also deleterious leading to the production of sperm with variable levels of anomalies.

Impact

We identified a new gene leading to male infertility permitting to improve the diagnostic efficiency for NOA patients. We demonstrate that whole-exome sequencing is an efficient technique to identify new infertility genes and to realize a diagnostic for affected men. We showed that the control of proteases by antiproteases, and in particular by SPINK2, is critical during spermiogenesis and demonstrate that the SPINK gene family is involved not only in pancreatitis or skin disease but also in male infertility.

Acknowledgements

We thank the GIN electron microscopy platform and Anne Bertrand, and the IAB microscopy platform and Alexei Grichine and Jacques Mazzega for their technical help. We thank Myriam Dridi for her work on HEK cells and antibody validation, Jean Pascal Hograindeur for his help for CASA experiments and Denise Escalier for her generous gift of human anti-acrosin antibody. This work was mainly supported by the French research agency (ANR) within the 2009 Genopat program for the ICG2I project "Identification and characterization of genetic causes of male infertility" to PR and CA. Support was also obtained from the Fondation Maladies Rares (FMR) for the project R16070CC, "Identification of genetic causes of human NOA".

Author contributions

PFR and CA designed the study, supervised all laboratory work, and wrote the manuscript. They have full access to all of the data in the study and take responsibility for the integrity of the data and its accuracy. All authors read, corrected, and made a significant contribution to the manuscript. Z-EK, TK, AA-Y, CB, MG, NT-M, and CC produced and analyzed the genetic data, and Z-EK, MC-K, AA-Y, and ASV performed immunohistochemistry (IF) experiments. SPB, JE and EL performed Western blot experiments and real-time cell

analyses. Z-EK and GM performed sperm analyses; KP-G and ASV performed the electron microscopy; and BD, IA-S, MM, CM-C, SN, VS, MB, FB, JF, and SH provided clinical samples and data and supplied biological materials.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Arboleda CE, Gerton GL (1987) Studies of three major proteases associated with guinea pig sperm acrosomes. *J Exp Zool* 244: 277–287
- Ayhan O, Balkan M, Guven A, Hazan R, Atar M, Tok A, Tolun A (2014) Truncating mutations in TAF4B and ZMYND15 causing recessive azoospermia. *J Med Genet* 51: 239–244
- Baba T, Michikawa Y, Kawakura K, Arai Y (1989) Activation of boar proacrosin is effected by processing at both N- and C-terminal portions of the zymogen molecule. *FEBS Lett* 244: 132–136
- Ben Khelifa M, Coutton C, Zouari R, Karaouzene T, Rendu J, Bidart M, Yassine S, Pierre V, Delaroche J, Hennebicq S et al (2014) Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. *Am J Hum Genet* 94: 95–104
- Brown CR, Harrison RA (1978) The activation of proacrosin in spermatozoa from ram bull and boar. *Biochim Biophys Acta* 526: 202–217
- Cesari A, Sanchez JJ, Biancotti JC, Vazquez-Levin MH, Kaiser G, Palma GA, Alberio R, Vincenti AE, Fornes MW (2004) Immunolocalization of bovine sperm protease BSp120 by light and electron microscopy during capacitation and the acrosome reaction: its role in *in vitro* fertilization. *Mol Reprod Dev* 69: 411–418
- Chen JM, Mercier B, Audrezet MP, Ferec C (2000) Mutational analysis of the human pancreatic secretory trypsin inhibitor (PSTI) gene in hereditary and sporadic chronic pancreatitis. *J Med Genet* 37: 67–69
- Chera S, Buzgariu W, Ghila L, Galliot B (2009) Autophagy in Hydra: a response to starvation and stress in early animal evolution. *Biochim Biophys Acta* 1793: 1432–1443
- Dezso Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, Bugrim A, Rakhmatulin E, Brennan RJ, Guryanova A et al (2008) A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol* 6: 49
- Dieterich K, Soto RR, Faure AK, Hennebicq S, Ben Amar B, Zahi M, Perrin J, Martinez D, Sele B, Jouk PS et al (2007) Homozygous mutation of AURKC yields large-headed polyploid spermatozoa and causes male infertility. *Nat Genet* 39: 661–665
- Feig C, Kirchhoff C, Ivell R, Naether O, Schulze W, Spiess AN (2007) A new paradigm for profiling testicular gene expression during normal and disturbed human spermatogenesis. *Mol Hum Reprod* 13: 33–43
- Furio L, Hovnanian A (2014) Netherton syndrome: defective kallikrein inhibition in the skin leads to skin inflammation and allergy. *Biol Chem* 395: 945–958
- Gallo JM, Escalier D, Grellier P, Precigout E, Albert M, David G, Schrevel J (1991) Characterization of a monoclonal antibody to human proacrosin and its use in acrosomal status evaluation. *J Histochem Cytochem* 39: 273–282
- Harbuz R, Zouari R, Pierre V, Ben Khelifa M, Kharouf M, Coutton C, Merdassi G, Abada F, Escoffier J, Nikas Y et al (2011) A recurrent deletion of DPY19L2 causes infertility in man by blocking sperm head elongation and acrosome formation. *Am J Hum Genet* 88: 351–361
- Huang-Yang YH, Meizel S (1975) Purification of rabbit testis proacrosin and studies of its active form. *Biol Reprod* 12: 232–238
- Kambhampati S, Park W, Habtezion A (2014) Pharmacologic therapy for acute pancreatitis. *World J Gastroenterol* 20: 16868–16880
- Kennedy WP, Swift AM, Parrish RF, Polakoski KL (1982) Proacrosin conversion inhibitor. Purification and initial characterization of a boar sperm protein which prevents the conversion of proacrosin into acrosin. *J Biol Chem* 257: 3095–3099
- Kierszenbaum AL, Rivkin E, Tres LL (2003) Acroplaxome, an F-actin-keratin-containing plate, anchors the acrosome to the nucleus during shaping of the spermatid head. *Mol Biol Cell* 14: 4628–4640
- Kim KS, Cha MC, Gerton GL (2001) Mouse sperm protein sp56 is a component of the acrosomal matrix. *Biol Reprod* 64: 36–43
- Kohno N, Yamagata K, Yamada S, Kashiwabara S, Sakai Y, Baba T (1998) Two novel testicular serine proteases, TESP1 and TESP2, are present in the mouse sperm acrosome. *Biochem Biophys Res Commun* 245: 658–665
- Lee B, Park I, Jin S, Choi H, Kwon JT, Kim J, Jeong J, Cho BN, Eddy EM, Cho C (2011) Impaired spermatogenesis and fertility in mice carrying a mutation in the Spink2 gene expressed predominantly in testes. *J Biol Chem* 286: 29108–29117
- Li WW, Li J, Bao JK (2012) Microautophagy: lesser-known self-eating. *Cell Mol Life Sci* 69: 1125–1136
- Liu DY, Baker HW (1993) Inhibition of acrosin activity with a trypsin inhibitor blocks human sperm penetration of the zona pellucida. *Biol Reprod* 48: 340–348
- Maor-Sagie E, Cinnamon Y, Yaakov B, Shaag A, Goldsmidt H, Zenvirt S, Laufer N, Richler C, Frumkin A (2015) deleterious mutation in SYCE1 is associated with non-obstructive azoospermia. *J Assist Reprod Genet* 32: 887–891
- Massart A, Lissens W, Tournaye H, Stouffs K (2012) Genetic causes of spermatogenic failure. *Asian J Androl* 14: 40–48
- Meizel S, Deamer DW (1978) The pH of the hamster sperm acrosome. *J Histochem Cytochem* 26: 98–105
- O'Brien DA, Welch JE, Goulding EH, Taylor AA Jr, Baba T, Hecht NB, Eddy EM (1996) Boar proacrosin expressed in spermatids of transgenic mice does not reach the acrosome and disrupts spermatogenesis. *Mol Reprod Dev* 43: 236–247
- Ohmura M, Hirota M, Araki M, Mizushima N, Matsui M, Mizumoto T, Haruna K, Kume S, Takeya M, Ogawa M et al (2005) Autophagic cell death of pancreatic acinar cells in serine protease inhibitor Kazal type 3-deficient mice. *Gastroenterology* 129: 696–705
- Ohmura M, Sugano A, Hirota M, Takaoka Y, Yamamura K (2012) Role of intrapancreatic SPINK1/Spink3 expression in the development of pancreatitis. *Front Physiol* 3: 126
- Okutman O, Muller J, Baert Y, Serdarogullari M, Gultomruk M, Piton A, Rombaut C, Benkhalfi M, Teletin M, Skory V et al (2015) Exome sequencing reveals a nonsense mutation in TEX15 causing spermatogenic failure in a Turkish family. *Hum Mol Genet* 24: 5581–5588
- Pierre V, Martinez G, Coutton C, Delaroche J, Yassine S, Novella C, Pernet-Gallay K, Hennebicq S, Ray PF, Arnoult C (2012) Absence of Dpy19l2, a new inner nuclear membrane protein, causes globozoospermia in mice by preventing the anchoring of the acrosome to the nucleus. *Development* 139: 2955–2965
- Rawlings ND, Tolle DP, Barrett AJ (2004) Evolutionary families of peptidase inhibitors. *Biochem J* 378: 705–716
- Redelfs L, Fischer J, Weber C, Wu Z, Meyer-Hoffert U (2016) The serine protease inhibitor of Kazal-type 9 (SPINK9) is expressed in lichen simplex chronicus, actinic keratosis and squamous cell carcinoma. *Arch Dermatol Res* 308: 133–137
- Rivinoja A, Pujol FM, Hassinen A, Kellokumpu S (2012) Golgi pH, its regulation and roles in human disease. *Ann Med* 44: 542–554

- SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 32: 903–914
- Taylor SC, Posch A (2014) The design of a quantitative western blot experiment. *Biomed Res Int* 2014: 361590
- Tuttelmann F, Werny F, Cooper TG, Kliesch S, Simoni M, Nieschlag E (2011) Clinical experience with azoospermia: aetiology and chances for spermatozoa detection upon biopsy. *Int J Androl* 34: 291–298
- Valenzuela DM, Murphy AJ, Frendewey D, Gale NW, Economides AN, Auerbach W, Poueymirou WT, Adams NC, Rojas J, Yasenckak J et al (2003) High-throughput engineering of the mouse genome coupled with high-resolution expression analysis. *Nat Biotechnol* 21: 652–659
- Witt H, Luck W, Hennies HC, Classen M, Kage A, Lass U, Landt O, Becker M (2000) Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. *Nat Genet* 25: 213–216
- Yan W, Si Y, Slaymaker S, Li J, Zheng H, Young DL, Aslanian A, Saunders L, Verdin E, Charo IF (2010) Zmynd15 encodes a histone deacetylase-dependent transcriptional repressor essential for spermiogenesis and male fertility. *J Biol Chem* 285: 31418–31426
- Yang F, Silber S, Leu NA, Oates RD, Marszalek JD, Skaletsky H, Brown LG, Rozen S, Page DC, Wang PJ (2015) TEX11 is mutated in infertile men with azoospermia and regulates genome-wide recombination rates in mouse. *EMBO Mol Med* 7: 1198–1210
- Yassine S, Escoffier J, Nahed RA, Pierre V, Karaouzene T, Ray PF, Arnoult C (2015) Dynamics of Sun5 localization during spermatogenesis in wild type and Dpy19l2 knock-out mice indicates that Sun5 is not involved in acrosome attachment to the nuclear envelope. *PLoS One* 10: e0118698
- Yatsenko AN, Georgiadis AP, Ropke A, Berman AJ, Jaffe T, Olszewska M, Westernstroer B, Sanfilippo J, Kurpisz M, Rajkovic A et al (2015) X-linked TEX11 mutations, meiotic arrest, and azoospermia in infertile men. *N Engl J Med* 372: 2097–2107
- Zahn A, Furlong LI, Biancotti JC, Ghiringhelli PD, Marijn-Briggiler CI, Vazquez-Levin MH (2002) Evaluation of the proacrosin/acrosin system and its mechanism of activation in human sperm extracts. *J Reprod Immunol* 54: 43–63
- Zhang T, Murphy MW, Gearhart MD, Bardwell VJ, Zarkower D (2014) The mammalian Doublesex homolog DMRT6 coordinates the transition between mitotic and meiotic developmental programs during spermatogenesis. *Development* 141: 3662–3671



License: This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Principaux résultats

L'analyse et le filtrage des données de séquençage des deux frères par notre pipeline a permis de retenir seulement 2 variants répondant à tous les critères établis (**Table : 2.4**). Le premier de ces variants chevauchait la séquence codante du gène *GUF1* et entraînait la substitution d'une sérine par une isoleucine. Le second impactait le gène *SPINK2* et se situait trois nucléotides en amont du deuxième exon pouvant ainsi entraîner des erreurs d'épissage pouvant par la suite soit engendrer un décalage du cadre de lecture menant à la formation d'un codon stop prématûr ou bien sauter complètement la transcription de cet exon créant un codon stop au début de l'exon 3.

Parmi ces deux gènes, seul *SPINK2* présentait une forte expression testiculaire dans les données Ensembl (**Figure : 2.13**) et nous avons confirmé cette surexpression par RT-PCR dans cette étude. De plus, un KO partiel de *Spink2* chez la souris avait déjà été décrit comme induisant des défauts partiels de la spermatogenèse [202]. Ces arguments ont ainsi fait de *SPINK2* le candidat évident pour expliquer le phénotype de ces deux frères. Après avoir confirmé en séquençage Sanger la présence de la mutation de ce gène à l'état homozygote pour les deux frères et hétérozygotes pour les parents, nous avons, afin de continuer nos investigations, développé un modèle murin KO *Spink2*^{-/-} confirmant une azoospermie complète pour les souris mâles causée par un arrêt de la spermatogenèse au stage des spermatides rondes. De plus, malgré une fertilité normale, nous avons pu noter un taux élevé d'anomalies morphologiques du spermatozoïde ainsi qu'une motilité spermatique réduite chez les souris mâles hétérozygotes *Spink2*^{+/+}. Les femelles, elles, ne présentaient aucun phénotype apparent. L'étude de la localisation de la protéine Spink2 chez la souris et SPINK2 chez l'humain a révélé que ces deux protéines localisaient dans la vésicule acrosomale depuis le début de la biogénèse de l'acrosome jusqu'au spermatozoïde mature. Nous avons, par ailleurs, démontré que Spink2, une antiprotéase à sérine, permettait de neutraliser l'acrosine (Acr) pendant son transit par l'appareil de golgi, et qu'en son absence, Acr déstructurait le golgi, empêchait la formation acrosomale et entraînait le blocage au stade de spermatide ronde.

Suite à cela, afin d'évaluer l'importance des variants du gène *SPINK2* dans l'infertilité humaine, nous avons effectué le séquençage Sanger de 611 patients parmi lesquels 210 étaient azoospèrme, 393 oligozoospèrme et 8 dont la cause n'était pas spécifiée. Parmi ces patients, seul 1 s'est révélé porter un variant non répertorié dans ExAC sur le gène *SPINK2*. Ce patient présentant un phénotype d'oligozoospermie porte à l'état hétérozygote un variant altérant le codon start du gène *SPINK2*. Ces résultats indiquent donc que chez l'homme, comme chez la souris, la présence de mutations homozygotes sur le gène *SPINK2* induit un phénotype d'azoospermie ou d'oligozoospermie sévère tandis que la présence d'une mutation hétérozygote pouvait entraîner un phénotype d'oligozoospermie avec un taux élevé d'anomalies morphologiques du spermatozoïde. Le fait que même les hétérozygotes puissent subir une pression de sélection négative pourrait expliquer la rareté observée des mutations *SPINK2*.

Table 2.4 – Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs de la famille AZ

Gene	Impact		Frequency		
	HGVSc, HGVSp	Consequence	ExAC	ESP	1KG
GUF1	c.443A>T ; p.Ser148Ile	missense	0.00207	0.0028	9e-04
SPINK2	c.56-3C>G ; .	splice region	.	.	.

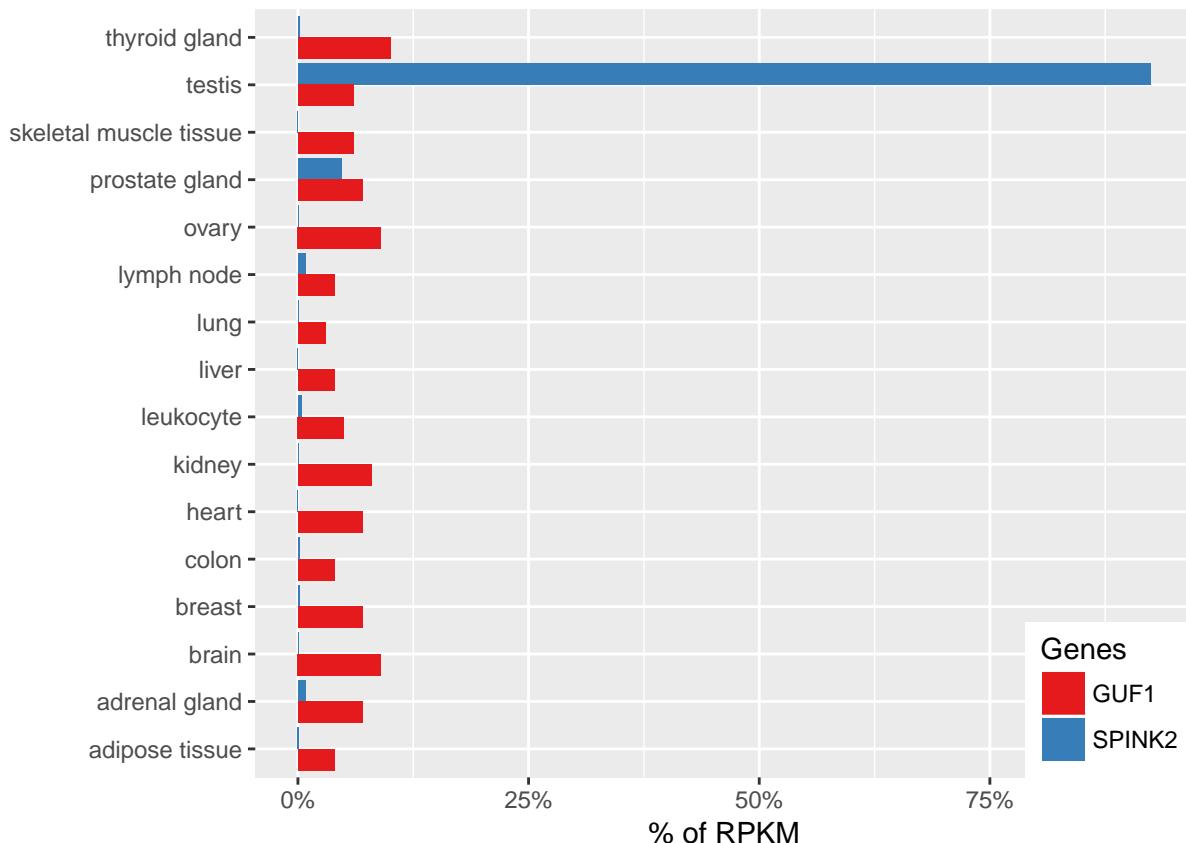


Figure 2.13 – Expression tissulaire des gènes SPINK2 et GUF1 : Données provenant du projet de transcriptome Illumina BodyMap. Contrairement au gène *GUF1* (en rouge) qui a une expression relativement ubiquitaire, *SPINK2* (en bleu) a une expression quasi spécifique au testicule.

2.2.3 Article n°2

Homozygous mutation of PLCZ1 leads to defective human oocyte activation and infertility that is not rescued by the WW-binding protein PAWP

Jessica Escoffier J*, Lee HC*, Yassine S*, Zouari R, Martinez G, Karaouzène T, Couston C, Kherraf ZE, Halouani L, Triki C, Nef S, Thierry-Mieg N, Savinov SN, Fissore R, Ray PF, Arnoult C

* Co-premiers auteurs

Human Molecular Genetics, Décembre 2015

Contexte et objectifs

L'activation ovocytaire regroupe une série de processus intervenant lors de la fécondation de l'ovocyte par le spermatozoïde. En 1990, plusieurs études ont démontré que chez les mammifères ces processus reposent principalement sur le relargage par le spermatozoïde de “facteurs spermatiques” qui déclenchent un signal constitué d'oscillations Ca^{2+} . Plus tard, la protéine *PLCZ1* fut identifiée comme la molécule responsable de l'induction de ces oscillations calciques. Cependant, l'incapacité à produire des modèles animaux *PLCZ1* KO capables de produire des spermatozoïdes matures a empêché d'attribuer l'exclusivité de ce rôle à *PLCZ1*, laissant ouverte la possibilité que l'activation ovocytaire puisse être tributaire d'autres facteurs spermatiques. C'est ainsi qu'en 2014 fut proposée la protéine PAWP comme facteur spermatique alternatif ou complémentaire à *PLCZ1* [203, 204].

Les travaux ci-dessous décrivent les analyses effectuées sur deux frères issus d'une union consanguine ayant tous deux été dans l'incapacité de concevoir un enfant spontanément. Des fécondations in vitro avec injection directe d'un spermatozoïde (ICSI) ont par la suite été réalisées et n'ont pas permis la fécondation ovocytaires. Comme dans l'étude précédente, en raison de l'historique de consanguinité de la famille des deux frères et du non-apparentement de leurs femmes respectives nous avons exclu l'hypothèse d'une cause féminine et nous avons recherché un variant homozygote commun aux deux frères par séquençage WES. Comme précédemment, j'ai été en charge de l'ensemble des analyses des données issues du séquençage des deux frères.

ORIGINAL ARTICLE

Homozygous mutation of PLCZ1 leads to defective human oocyte activation and infertility that is not rescued by the WW-binding protein PAWP

Jessica Escoffier^{1,†}, Hoi Chang Lee^{2,†}, Sandra Yassine^{4,5,†}, Raoudha Zouari⁶, Guillaume Martinez^{4,5}, Thomas Karaouzène^{4,5}, Charles Coutton^{4,7}, Zine-eddine Kherraf^{4,5}, Lazhar Halouani⁶, Chema Triki⁸, Serge Nef¹, Nicolas Thierry-Mieg^{4,9}, Sergey N. Savinov³, Rafael Fissore^{2,‡}, Pierre F. Ray^{4,5,10,‡} and Christophe Arnoult^{4,5,‡,*}

¹Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland,

²Department of Veterinary and Animal Sciences and ³Department of Biochemistry and Molecular Biology,

University of Massachusetts, Amherst, MA 01003, USA, ⁴Université Grenoble Alpes, Grenoble, F-38000, Grenoble,

France, ⁵Institut Albert Bonniot, INSERM U823, La Tronche F-38700, France, ⁶Polyclinique les Jasmins, Centre

d'Aide Médicale à la Procréation, Centre Urbain Nord, 1003 Tunis, Tunisia, ⁷CHU de Grenoble, UF de Génétique

Chromosomique, Grenoble F-38000, France, ⁸Clinique Hannibal, Centre d'AMP, les berges du lac, 1053 Tunis,

Tunisia, ⁹Laboratoire TIMC-IMAG, UMR CNRS 5525, Grenoble F-38000, France and ¹⁰CHU de Grenoble, UF de

Biochimie et Génétique Moléculaire, Grenoble F-38000, France

*To whom correspondence should be addressed at: Faculté de Médecine et de Pharmacie, Equipe 'Génétique, Epigénétique et Thérapies de l'Infertilité', Bâtiment Jean Roget – 3 étage, Pièce 311, Place du Cdt NAL – Domaine de la Merci, 38700 La tronche, France. Tel: +33 476637408; Email: christophe.arnoult@ujf-grenoble.fr

Abstract

In mammals, sperm–oocyte fusion initiates Ca^{2+} oscillations leading to a series of events called oocyte activation, which is the first stage of embryo development. Ca^{2+} signaling is elicited by the delivery of an oocyte-activating factor by the sperm. A sperm-specific phospholipase C (PLCZ1) has emerged as the likely candidate to induce oocyte activation. Recently, PAWP, a sperm-born tryptophan domain-binding protein coded by WBP2NL, was proposed to serve the same purpose. Here, we studied two infertile brothers exhibiting normal sperm morphology but complete fertilization failure after intracytoplasmic sperm injection. Whole exomic sequencing evidenced a missense homozygous mutation in PLCZ1, c.1465A>T; p.Ile489Phe, converting Ile 489 into Phe. We showed the mutation is deleterious, leading to the absence of the protein in sperm, mislocalization of the protein when injected in mouse GV and MII oocytes, highly abnormal Ca^{2+} transients and early embryonic arrest. Altogether these alterations are consistent with our patients' sperm inability to induce oocyte activation and initiate embryo development. In contrast, no deleterious variants were identified in WBP2NL and PAWP presented normal expression and localization. Overall we demonstrate in humans, the absence of PLCZ1 alone is sufficient to prevent oocyte activation irrespective of the presence of

†J.E., H.C.L. and S.Y. contributed equally.

‡R.F., P.F.R and C.A. shared leadership.

Received: November 6, 2015. Revised: December 6, 2015. Accepted: December 17, 2015

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

PAWP. Additionally, it is the first mutation located in the C2 domain of PLCZ1, a domain involved in targeting proteins to cell membranes. This opens the door to structure–function studies to identify the conserved amino acids of the C2 domain that regulate the targeting of PLCZ1 and its selectivity for its lipid substrate(s).

Introduction

Increases in the intracellular concentration of free calcium (Ca^{2+}) was demonstrated to be a sufficient and necessary stimulus to trigger oocyte activation and embryo development in all species studied to date (1). This finding along with subsequent confirmatory studies (2) stimulated interest to elucidate the signaling cascade responsible for Ca^{2+} release at fertilization. In 1990, studies showed that a soluble component of mammalian sperm extracts, aptly named the sperm factor, was sufficient to induce oocyte activation and replicate the periodical Ca^{2+} responses, also known as Ca^{2+} oscillations, which are a hallmark of mammalian fertilization (3,4). PLCZ1 was later identified as the candidate molecule to be the active factor in sperm responsible for the oscillations (5,6). Research from several laboratories confirmed the unique properties of PLCZ1 to induce Ca^{2+} oscillations in oocytes as well as the association of its absence with infertility (7–12). However, the inability thus far to obtain a PLCZ1 KO animal model capable of producing mature sperm (13) has prevented assigning to this molecule the exclusive role for oocyte activation, while leaving open the possibility that other sperm factors may be required (14). Toward that end, PAWP has been proposed as an alternative or complementary pathway for oocyte activation based on findings that injection of recombinant PAWP induced oscillations comparable to those of fertilization (15,16). The relationship between the proposed function of PAWP and its structure is not yet understood. PAWP displays sequence homology to WW domain-binding protein 2 (WBP2) in its N terminal end and a variable number of PPXY motifs (one in human, six in mouse) in its C-terminal end, a motif known to interact with WW domain. Moreover, the C terminal end contains an unidentified repeated motif (YGXPPXG) (17). It is presently unknown, however, how these motifs may engage the oocyte's signaling machinery to induce Ca^{2+} oscillations. In this vein, attempts to replicate those studies in mouse oocytes failed (18,19), which raised concerns regarding the importance of PAWP in fertilization. However, its action in human oocyte activation was not formally ruled out, especially as numerous studies have shown that testis and reproductive tissues evolve faster than other tissues leading to noticeable differences between species (20).

Here, we studied two infertile brothers showing complete fertilization failure after intracytoplasmic sperm injection (ICSI). Whole-exome sequencing enabled us to find a missense homozygous mutation in PLCZ1, c.1465A>T; p.Ile489Phe, converting Ile 489 into Phe. Structure–function models revealed that the mutation causes a conformational change that might affect the enzyme's ability to bind to its substrate(s). Using western blotting (WB), immunofluorescence (IF), live fluorescence and Ca^{2+} imaging, we show that the mutation is deleterious, leading to mislocalization of the protein, lower Ca^{2+} signaling and lower rates of oocyte activation and embryo development. In contrast, no mutations were identified in WBP2NL and PAWP showed normal expression and localization. Overall we demonstrate the absence of PLCZ1 alone in humans is sufficient to prevent oocyte activation irrespective of PAWP. Moreover, it is the first mutation located in the C2 domain of the enzyme, an important domain targeting proteins to lipidic membranes, opening the door for better

structure–function analysis of PLCZ1 and of other proteins carrying this domain.

Results

Patients' description: medical records and spermatocytogram

Two Tunisians brothers and their respective wives sought medical advice from infertility clinic in Tunis between 2011 and 2014 after unsuccessful attempts for a full year to conceive a pregnancy. The brothers were born from first cousin parents and have one fertile brother with children conceived spontaneously and two fertile sisters. The morphology of patients' sperm was assessed with Shorr staining (Fig. 1A) and sperm parameters were well above the low reference limits set by the WHO guidelines (21), although Patient 1 (P1) had higher than average number of sperm with acrosome defects (Table 1). To investigate this in more detail, we examined the presence and morphology of the acrosome by IF using an anti-acrosin antibody. We found that the acrosome presented a normal shape in ~50% of the brothers' sperm (Fig. 1D and E). While not examined, we estimate that the absence of staining on the other half is due to a premature acrosome reaction rather than to an abnormal acrosome biogenesis, as the shape of the sperm heads was normal and not globozoospermic (Fig. 1). We also assessed DNA quality using three different methods: chromomycin A3, aniline blue (AB) and terminal deoxynucleotidyl transferase-mediated deoxyuridine triphosphate-nick-end labeling (TUNEL), which allowed evaluation of DNA protamination, histone content and DNA fragmentation, respectively. The percentages of positive sperm were higher in patients than in fertile controls, but all values were ~20% (Fig. 2) and remained below abnormal thresholds defined by several studies (22–24). Remarkably, although these sperm parameters were compatible with spontaneous conception, for both brothers the clinical outcomes following assisted reproduction technologies procedures was oocyte activation failure (OAF); P1 had two unsuccessful attempts of artificial insemination and then sperm from both patients were used for ICSI. In total, 3 ICSI cycles were carried out and 20 MII oocytes were injected, but none showed signs of oocyte activation (Table 2).

Whole-exome sequencing identified a homozygous missense mutation in PLCZ1

Since both brothers were married to unrelated women, we excluded the possibility of a female factor and focused our research on the brothers. Given the family history of consanguinity, we postulated that this infertility was caused by a homozygous mutation. We, therefore, proceeded to whole-exome sequencing to identify a possible genetic defect(s) that could explain the observed OAF. After exclusion of frequent variants, only four homozygous variants were identified in both brothers (Table 3). Three variants, located in EPS8 (also known as DFNB102), RP11-1021N1.1 and LKAAEAR1 (also known as C20orf201), had no expected deleterious effect. The fourth variant was a missense mutation on PLCZ1, c.1465A>T located in exon 13 (NM_033123.3), changing an Ile at position 489 into a Phe (Ile489Phe) (Fig. 3A). No other

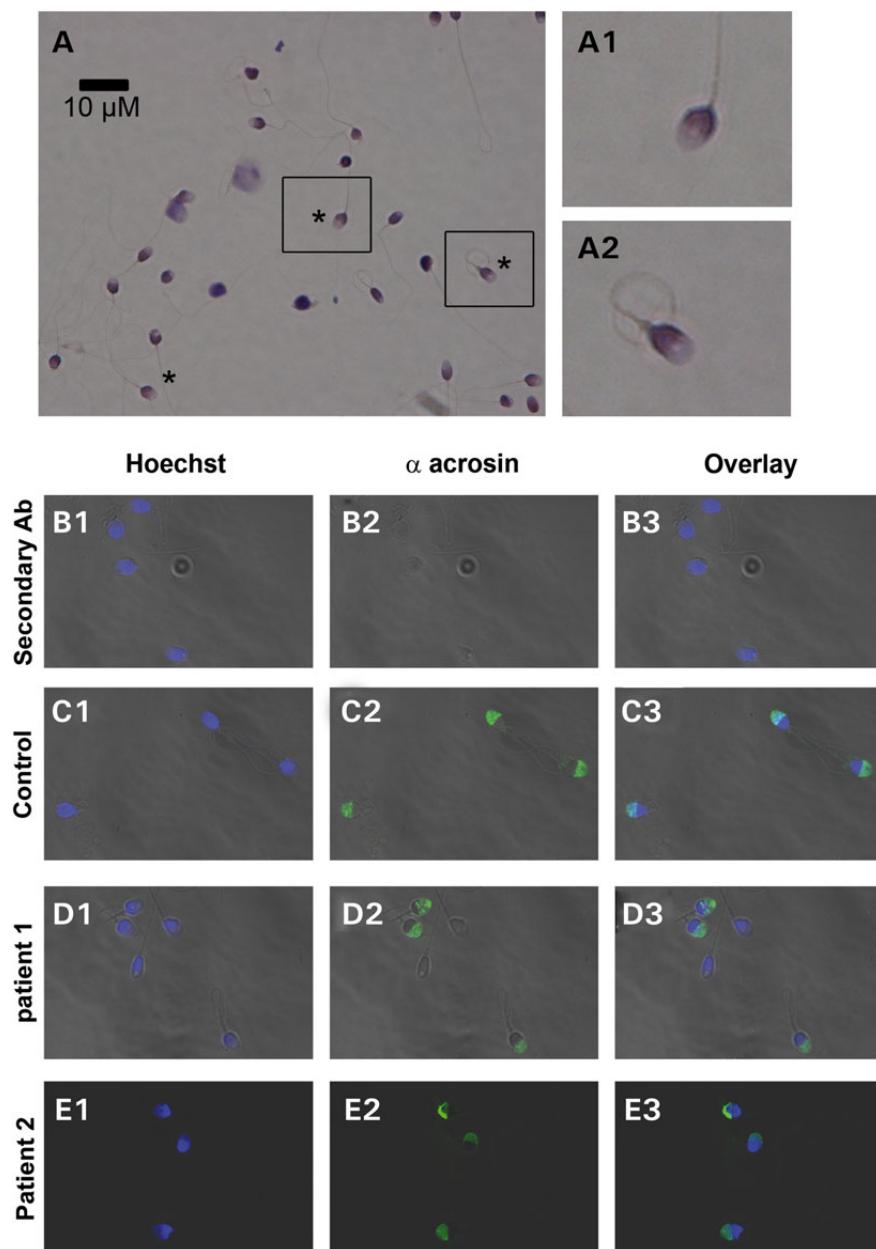


Figure 1. Morphology of the patients' sperm. (A) Sperm samples were spread over a slide and dried at RT, fixed in ether/ethanol 1:1 for Harris-Schorr staining. Sperm were then scored according to WHO's laboratory manual for examination and processing of human semen (5th edition). Asterisk indicates fully normal spermatozoa. A1–A2 correspond to enlargement of the black squares in (A), in order to show the morphology of normal sperm. (B–E) Morphology of acrosome stained with anti-acrosin antibody (B) Negative control experiment with the addition of secondary antibody without primary antibody. (C–E) Acrosin staining (green staining, C2–E2) of sperm from subjects with normal fertility (C) and from OAF patients (D and E) and counterstained with Hoechst (blue, B1–E1) to evidence the nucleus/acrosome ratio. The third column corresponds to overlay (B3–E3).

variants were identified in the PLCZ1-coding sequence, UTR regions or close intronic regions. Given that PLCZ1 has been suggested to be necessary for oocyte activation in mammals, this mutation could underlie the patients' phenotype. Sanger sequencing confirmed the homozygous mutation for both infertile brothers and showed that the third fertile brother was heterozygous (Fig. 3B). The c.1465A>T variant was absent from over 60 000 individuals described in the ExAC database (exac.broadinstitute.org), which confirms it is not a polymorphism and that missense variations occurring at this localization would likely cause

negative selection throughout evolution. Moreover, we found Ile489 to be well conserved throughout evolution (Fig. 3C), suggesting that this mutation could be deleterious.

Deleterious effects of the identified mutation

In order to assess the impact of the mutation, PLCZ1 expression and localization were first studied on sperm from both patients (Fig. 4A). We used an antibody targeting human PLCZ1 (anti-hPLCZ1). This antibody has been published and used in three

previous publications and the respective western blots showed remarkably specificity (7,11,25). Whereas the sperm of fertile patients showed a band of strong staining on the post-acrosomal area and more diffuse staining over the acrosome (Fig. 4A1, inset), there was no staining detectable in the patients' sperm, except for a few of them showing a faint punctate staining over the acrosome (Fig. 4A2, inset), suggesting that the mutant PLCZ1 was absent or present at very low concentrations. The absence of PLCZ1 in the patients' sperm was confirmed by WB, where a single band at the expected MW of PLCZ1 was noticeable in the control lane, whereas there was no reactivity in the patients' lanes (Fig. 4B), although similar sperm number were loaded per lane (Fig. 4C). These results suggest that during spermatogenesis Ile489Phe PLCZ1 displays defects in stability, trafficking and/or anchoring, which prevent its presence in mature sperm. To examine these possibilities, cRNAs of human WT and Ile489Phe Venus-tagged PLCZ1 were injected into mouse GV oocytes, which were maintained arrested at the GV stage with IBMX; this was done to ascertain the distribution of the enzyme independently of the stage of the cell cycle. It is noteworthy that is presently unknown how PLCZ1 distribution is regulated in oocytes, although unlike other PLCs (26), it does not appear to localize to the plasma membrane (27), where most of phosphatidylinositol 4,5-bisphosphate (PIP₂), the enzyme's substrate, is found. In agreement with this remark, WT PLCZ1 exhibited a homogeneous distribution in the ooplasm of GV oocytes that partly overlapped with distribution of the endoplasmic reticulum (ER), which was marked by ER-DsRed (Fig. 5A). In contrast, Ile489Phe PLCZ1 displayed an uneven distribution characterized by large patches near the nucleus and decreased

peripheral localization (Fig. 5B). Furthermore, Ile489Phe PLCZ1 distribution did not overlap with the ER. These results suggest that the mutation does not affect the stability of PLCZ1, although its trafficking and/or anchoring properties were clearly altered. To further test the latter, we took advantage of the fact that following fertilization, *Plcz1* is translocated into the pronuclei (PN) (28,29). The sequestration of *Plcz1* into the PNs is thought to contribute to the cessation of Ca²⁺ oscillations, which in the mouse closely corresponds with their formation. WT and Ile527Phe m*Plcz1* cRNAs [the equivalent mutation of Ile489Phe human PLCZ1 (hPLCZ1)] were injected into MII oocytes, which were activated by the oscillations initiated by the translated proteins. Following PN formation, as expected, WT m*Plcz1* accumulated into the PN. In contrast, following injection of Ile527Phe m*Plcz1* cRNA, PN formation was delayed by ~3 h, presumably due to the lower frequency of oscillations, and the mutant protein failed to localize to the PN (Fig. 5C). These results confirm that the mutation modifies PLCZ1's anchoring and/or trafficking properties in oocytes and zygotes.

We next assessed the enzymatic activity of the mutated human (h) and mouse (m) PLCZ1s, by examining Ca²⁺ responses elicited by the injection of their respective cRNAs into mouse MII oocytes. WT hPLCZ1 cRNA (0.001 µg/µl) initiated high-frequency oscillations in all injected oocytes (Fig. 6A) whereas the mutant hPLCZ1 cRNA failed to initiate oscillations in 14/31 oocytes (46%) or induced responses with a low frequency in 17/31 oocytes (54%) (Fig. 6B). Furthermore, the enzymatic activity of the mutant versus WT hPLCZ1 was reduced irrespective of the concentrations of cRNA injected (Fig. 6C). These results suggest that possible trace amounts of hPLCZ1 are likely not sufficient to activate oocytes. To confirm this hypothesis, both WT and mutant cRNAs were injected into mouse oocytes, and the development of parthenote embryos was followed by evaluating the rates of PN formation, cleavage to the two-cell stage and blastocysts. Following injection of WT hPLCZ1 64.6% of the oocytes showed signs of activation (2PN) and 35% reached the blastocyst stage. In contrast, Ile489Phe hPLCZ1 showed a greatly reduced ability to induce oocyte activation and allowed the fertilization of only 13.9% of the oocytes and none developed to the blastocyst stage (Fig. 6D). Similar results were obtained with the mutant m*Plcz1*, as Ca²⁺ responses triggered by the mIle527Phe mutant were weaker than those observed with WT *Plcz1*, showing longer lag time and reduced frequency (Fig. 7A, B). Consistent with this, the timing of PN formation was delayed and the percentages of zygotes reaching the 2PN stage at 6 and 10 h post-insemination

Table 1. Semen parameters and spermatocytogram of P1 and P2

Semen parameters	P1	P2
Semen volume (ml)	3	5.8
Sperm concentration (10 ⁶ /ml)	150	101
Total motility 1 h (%)	50	40
Vitality (%)	63	70
Normal spermatozoa (%)	20	7
Anomalies of the flagella (%)	10	20
Abnormal acrosome (%)	50	78
Other anomalies of the head (%)	10	20
Multiple anomalies index (%)	1.3	1.4

Values are the average of two separate analyses.

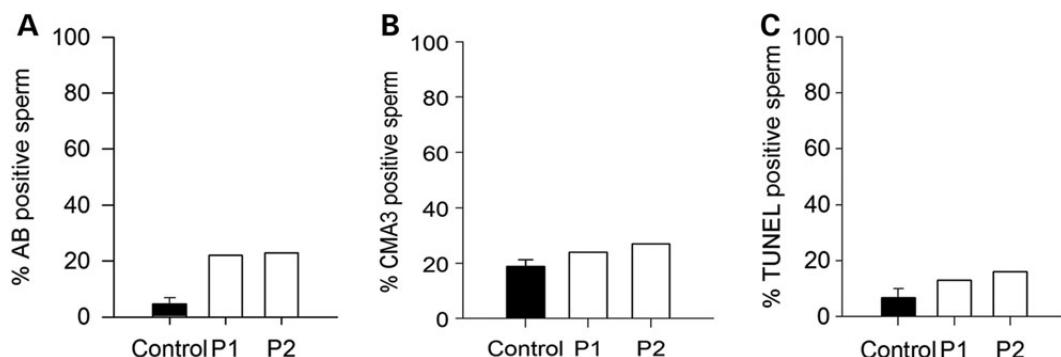


Figure 2. Assessment of nucleus compaction and DNA breaks of sperm from OAF patients. (A) Histones content of sperm was assessed using the AB test. The histogram shows the percentage of stained sperm in samples from control ($n = 5 \pm SD$) and patients ($n = 1$). (B) Protamination of sperm was evaluated using the chromomycin A3 test, and results are displayed in the histogram, which shows the percentage of stained sperm in control ($n = 5 \pm SD$) and patients' ($n = 1$) samples. (C) The histogram of DNA fragmentation analysis with TUNEL assay showing the level of TUNEL-positive sperm in control ($n = 5 \pm SD$) and patients' ($n = 1$) samples.

Table 2. ICSI outcomes following stimulation cycles with sperm from P1 and P2

Patient and procedure	Years	No. of follicles (n)	No. of abnormal oocytes (GV, M1, atretic) (n)	No. of mature oocytes injected (n)	No. of 2PN oocytes (n)
P1 ICSI—Jasmin clinic	2011	9	4	5	0
P1 ICSI—Jasmin clinic	2012	9	2	7	0
P2 ICSI—Jasmin clinic	2012	14	6	8	0

Number (n), P1 and P2 are brothers.

Table 3. List of common homozygous variants present in P1 and P2

Gene	Variant coordinates	Transcript	cDNA variation	Amino acid variation	Prediction
PLCZ1	chr12:18841149, T>A	NM_033123.3	c.1465A>T	p.Ile489Phe	Damaging
EPS8	chr12:15776164, A>C	NM_004447.5	c.2283A>C	p.Asp761Glu	Benign
RP11-1021N1.1	chr16:15528278, T>C	ENST00000568222 ^a	c.286T>C	p.Tyr96His	Benign
LKAAEAR1	chr20:62715318, T>G	NM_001007125.1	c.255T>G	p.Glu85Asp	Benign

Coordinates of all variations are based on the UCSC GRCh37/hg19 assembly.

^aNo RefSeq transcript accession number currently available.

was decreased (Fig. 7C) when Ile527Phe mPlcz1 cRNA was injected compared with WT Plcz1 cRNA injection; the altered Ca^{2+} signaling also prevented most zygotes from cleaving past the two-cell stage (Fig. 7D).

The Ile489Phe mutation alters the conformation of the C2 domain

To assess the possible impact of the mutation on the structure of hPLCZ1, a 3D structure of hPLCZ1 was modeled from the crystallographic structure of rPLC81 (42% identity and 59% homology). The model shows that Ile489 is at the interface of the EF hand and C2 domains, and it appears as being too far from the catalytic domain to directly affect its function (Fig. 8A). Using molecular dynamics simulations of the hPLCZ1 and hPLCZ1-Ile489Phe models, we observed that the larger size of Phe over Ile (203 versus 169 Å³), which requires more accommodating space (Supplementary Material, Fig. S1), results first in a displacement of its intra-domain neighbors (Y582, F601, Y603 and R487) and second in establishing a unique inter-domain hydrophobic contact with I76 from the proximal EF-hand 2 helix (Fig. 8B and C). The first notable outcome of these perturbations is that causes a significant shift of the EF2 domain toward the C2 domain (12.2 versus 10.7 Å for $\text{Ca}_{\text{I76}}-\text{Ca}_{\text{I489}}$ and $\text{Ca}_{\text{I76}}-\text{Ca}_{\text{F489}}$ distances, respectively), which is reinforced by the newly formed H-bond between the newly displaced Y582 and Y80 of EF2 (Fig. 8C). Significantly, this tighter inter-domain arrangement returns back to the original looser status when Phe489 is mutated back to Ile489 in the course of a simulation (Supplementary Material, Fig. S2). The second notable outcome is the formation in the C2 domain of an aromatic-rich concave-like sub-site capable of associating with lipophilic molecules or protein surfaces (Fig. 8D). Because EF hands have been shown to be important for both PLCZ1 anchoring on phospholipids and its nuclear translocation (30,31), this unique new inter-domains interaction in the mutant enzyme could support the observed deleterious effects of the mutation.

Normal PAWP expression in patients with Ile489Phe PLCZ1 mutation

We next examined the sequence of WBP2NL, which encodes for PAWP. Analysis of exome data from P1 and P2 did not reveal

any sequence variation in the WBP2NL-coding sequence, UTR regions or close intronic regions indicating that P1 and P2 should produce a fully functional PAWP protein. Because exome sequencing only allows covering 80–90% of all targeted sequences, WBP2NL coverage was verified: all exons and exon borders for both patients was at least 40×, which unambiguously confirmed the absence of deleterious variants in WBP2NL in the patients (Supplementary Material, Fig. S3). Consistent with these results, we confirmed by WB normal expression of PAWP (Fig. 9A) using extracts of sperm from patients and a fertile control prepared with similar number of sperm (Fig. 9B). We also examined PAWP localization by IF and in agreement with previous reports, PAWP reactivity appeared as a compact band around the equatorial/post-acrosomal area in the sperm of a control fertile human (Fig. 9C1) as well as in the sperm of both patients (Fig. 9C2). Altogether, these results show that PAWP is unable to support activation of human oocytes when PLCZ1 is absent and/or non-functional.

To extend these results, we examined the sperm from *Dpy19l2* KO males. *Dpy19l2* homozygous KO male mice are infertile, they have round-headed acrosomeless spermatozoa, which fail to induce oocyte activation following ICSI due to loss of *Plcz1* (11,32). Here, we show by WB that PAWP expression is normal in these round-headed sperm (Supplementary Material, Fig. S4), demonstrating that its presence is not sufficient to trigger oocyte activation when ICSI is performed. Therefore, it appears the presence of PAWP is incapable of rescuing the lack of activation caused by the absence of PLCZ1 expression in human or mouse sperm.

Discussion

Infertility and Ile489Phe PLCZ1

Herein, we have used whole-exome sequencing to identify a homozygous missense mutation in PLCZ1 in two infertile brothers presenting OAF. The mutation led to an almost complete disappearance of the protein in the patients' sperm and based on the targeting and/or anchoring defects observed following injection of cRNAs into oocytes, we estimate it hampers the retention of PLCZ1 in mature sperm. During spermatogenesis, as in all cells, proteins are synthesized in the reticulum and targeted to their final location according to their function. However, the

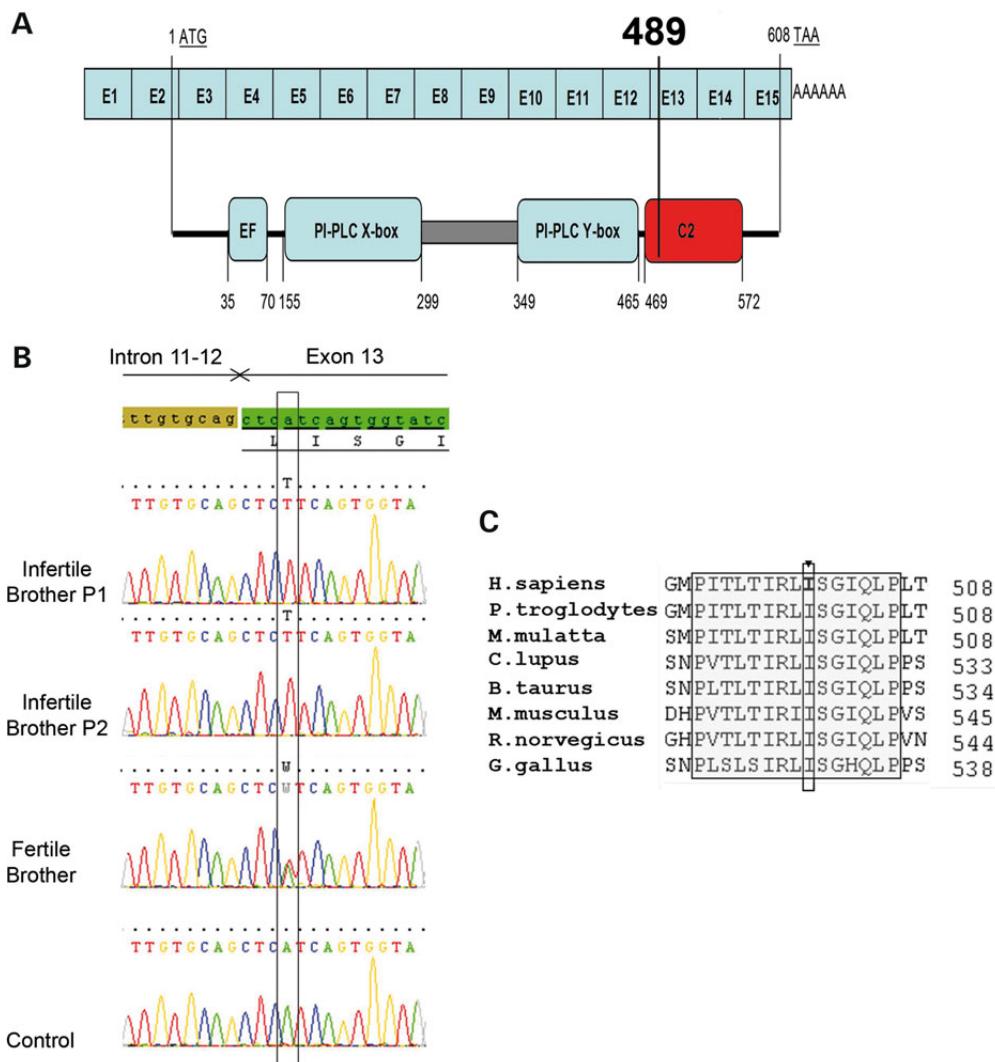


Figure 3. The Ile489Phe mutation is located in the C2 domain of PLCZ1. (A) Schematic representation of the exonic structure of human PLCZ1 cDNA sequence and corresponding functional domains of PLCZ1 (<http://www.uniprot.org/uniprot/Q86YW0>). The first coding exon is exon 2 (exon sizes are not to scale). The mutation c.1465A>T; p.Ile489Phe (NM_033123.3) is located in exon 13 and changes Isoleucine 489 located in the C2 domain into a phenylalanine. (B) The presence of the identified variation c.1465A>T; p.Ile489Phe (NM_033123.3) was verified by Sanger sequencing of PLCZ1 exon 13. Electropherogram of PLCZ1 exon 13 showing the mutated sequence and sequence obtained from a control individual. The two infertile brothers carried a homozygous missense mutation (p.Ile489Phe) in PLCZ1 exon 13 whereas the fertile brother harbors the mutation in a heterozygous state. (C) The mutation is located in a cluster of 15 highly conserved amino acids.

endoplasmic reticulum (ER) and all components necessary for proteins synthesis are eventually discarded at the end of the spermatozoon differentiation, leading to the release of a giant anucleate vesicle known as the residual body (33). We have previously shown that PLCZ1 is specifically located in the perinuclear theca in the vicinity of the inner acrosomal membrane of mature human sperm (11). The absence of PLCZ1 in the patients' sperm suggests that the mutation may prevent PLCZ1 from reaching and attaching to the perinuclear theca/inner acrosomal membrane, which renders it susceptible to disposal through the residual body. It is worth noting that PLCZ1 is also absent in globozoospermic sperm, which display defects in the perinuclear theca region and lack the acrosome vesicle (11). Moreover, we showed that the mutant protein had a strongly impaired ability to produce inositol 1,4,5-trisphosphate (IP_3), as witnessed by the failure to induce Ca^{2+} oscillations and to sustain normal embryonic development, contrary to the WT protein. The QAF of

these patients is thus due to the dramatic decrease of both PLCZ1 concentration and IP₃ production.

Ile489Phe PLCZ1 and function of the C2 domain of PLCZ1

PLCs belong to a large family of enzymes able to bind to lipids in membranes where they hydrolyze-specific phospholipids, mostly PIP2. Several classical molecular domains are found in PLCs, including two catalytic domains, called X and Y, and working in tandem, lipid-binding domains such as PH and C2 and EF-hand domains, which are Ca^{2+} -dependent domains (34). PLCZ1 is the shortest PLC and contains only four domains: four EF hands, the XY catalytic tandem and a C2 domain (Fig. 3) and the lack of a PH domain favor its more widespread distribution in the ooplasm. The C2 domain, which is the site of the homozygous mutation in our patients, is a lipid-binding domain shared by more than 100 human proteins. Originally, C2 domains were

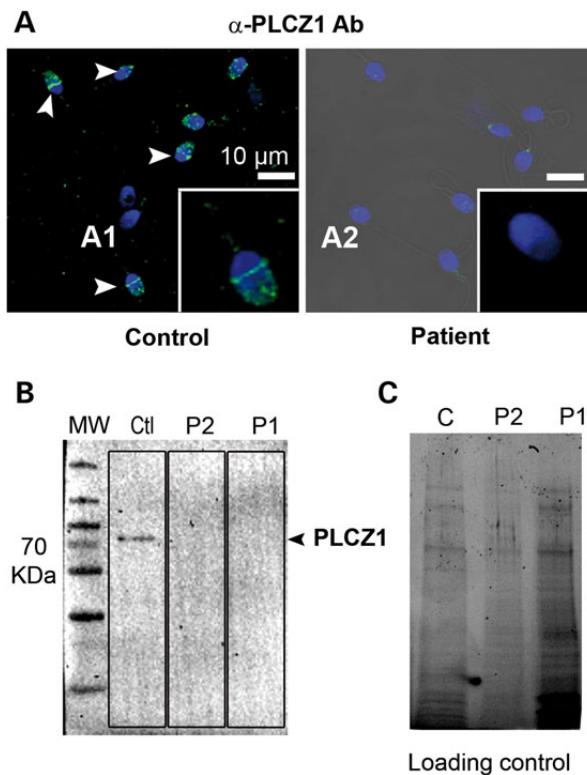


Figure 4. PLCZ1 expression is undetectable by IF and WB in sperm of patients exhibiting OAF after ICSI. (A1) Overlay of Hoechst staining (blue) and PLCZ1 staining (green) of sperm from a fertile control patient showing strong immunoreactivity in the post-acrosomal area (white arrow heads and inset) and to a lesser extent in the acrosome. (A2) Representative staining of sperm from patients with OAF showing the absence of PLCZ1 staining in the post-acrosomal area and very faint staining over the acrosome. More than 100 sperm were observed per patient. (B) WB using protein extracts of sperm from a fertile control (Ctl), or Patient 1 (P1) and Patient 2 (P2) and an anti-PLCZ1 antibody that fails to detect reactivity in the patients' lanes. (C) Protein gel representing loading control for Figure 1B showing that all lanes were similarly loaded. Protein loads were controlled with TGX stain free™ precast gels.

shown to bind membranes in a Ca^{2+} -dependent manner (35,36). Crystallographic studies have shown that C2 domains have a common fold of conserved eight-stranded antiparallel β -sandwich (37,38), which, according to our modeling studies, are also present in PLCZ1 (Fig. 8). There are two areas important for lipid binding in this domain, which are the highly variable loops between β -strands and a cationic patch in the concave face of the β -sandwich that corresponds to $\beta 3$ - and $\beta 4$ -strands (39). However, the sequence conservation among C2 domains is low and a significant number of C2 domains with little to no Ca^{2+} affinity have been identified, which is the case for the C2 domain of PLCZ1 (39). Moreover, some C2 domains have low-membrane affinity and are involved in protein–protein interaction. Although it has been shown that PLCZ1 deleted of the full C2 domain does not produce Ca^{2+} oscillations, demonstrating thus that the C2 domain of PLCZ1 is necessary for the overall enzymatic activity (30,40), its molecular function remains uncharacterized so far. The Ile489Phe mutation, which is located in the $\beta 1$ -strand of the C2 domain, opens the door for structure–function studies to define how the conserved amino acids on the C2 domain are involved in protein targeting and selectivity for lipid substrate(s).

Our results, following injection of PLCZ1 cRNAs into oocytes, have showed that the WT protein was distributed evenly in the

ooplasm and its area of distribution largely overlapped that of the ER, whereas the Ile489Phe PLCZ1 was unevenly distributed and accumulated around the nucleus. It is worth noting that the subcellular localization of C2-domain containing proteins is controlled by the lipid selectivity of the C2 domain and it is also known that lipid selectivity is highly variable among C2 domains (39,41). For instance, PLC δ is located in the phosphatidylserine (PS) rich PM because its C2 domain selectively binds PS (26), whereas cPLA 2α translocates to the perinuclear ER, which is rich in phosphatidylcholine, the target of its C2 domain (41). We can, therefore, speculate that the C2 domain of Ile489Phe PLCZ1 displays changed lipid affinity, which alters the intracellular distribution of the enzyme, and this may conspire to its retention in mature sperm and its reduced catalytic activity in oocytes. Altogether, our results suggest for the first time that the C2 domain of PLCZ1 participates in the targeting of the enzyme to lipid-containing membranes and our results demonstrate a pivotal role of a stretch of highly conserved residues in the C2 domain surrounding Ile489. It is also the first description of a functional role for residues located in the $\beta 1$ -strand of a C2 domain (42).

Our molecular modeling studies of WT and mutant PLCZ1 offer insights into the atomic-level perturbations caused by the Ile489Phe mutation. Dynamic simulations showed the Ile489Phe mutation is likely to induce a shift of the EF2-hand helix 1 toward C2, and given that EF2 hand is involved in binding to lipid membrane (30,31), this shift may reduce enzymatic activity. Further, the opening of a new aromatic-rich surface patch in the mutant protein capable of associating with additional hydrophobic counterparts may also alter the distribution and/or activity of the enzyme. Finally, the nuclear translocation of the mIle527Phe Plcz1 is hampered, suggesting that binding to nuclear import proteins necessary for PN translocation is defective (43), which may suggest a role of C2 domain in protein–protein interactions. Therefore, by changing the structural properties of the C2 domain and its interaction with the EF-hand domains, Ile489Phe-mutation is likely to modify the interactions of PLCZ1 with lipid membranes and possibly proteins, directly affecting its targeting and/or anchoring properties and enzymatic activity, all of which undermine the fertility of patients bearing this mutation.

PLCZ1 and oocyte activation

The function of PLCZ1 as the sperm factor has recently been challenged, and another protein, PAWP, has been proposed as an alternative candidate (15,16). Controversial results, however, have been published (18,19) and it is presently unclear how this protein is involved in human fertilization. The study of genetic diseases often provides the opportunity to better understand protein functions and our understanding of reproduction, including gametogenesis and fertilization, has benefited from the genetic characterization of several phenotypes of male and female infertilities (44). In human, a link between OAF and PLCZ1 was first reported in patients displaying abnormal expression and localization of PLCZ1 in sperm (7). However, these patients exhibited severe teratozoospermia, which raised concerns about miss-expression and/or malfunction of several proteins. Heytens et al. (8) provided the first genetic evidence linking PLCZ1 to infertility, as they found a heterozygous mutation at position 398 of the Y catalytic domain of PLCZ1 that reduced its ability to induce Ca^{2+} oscillations. Nevertheless, it was unclear how this heterozygous mutation could cause infertility, although, subsequently, another heterozygous mutation was found in the same patient at position 233 of the X catalytic domain, resulting in compound

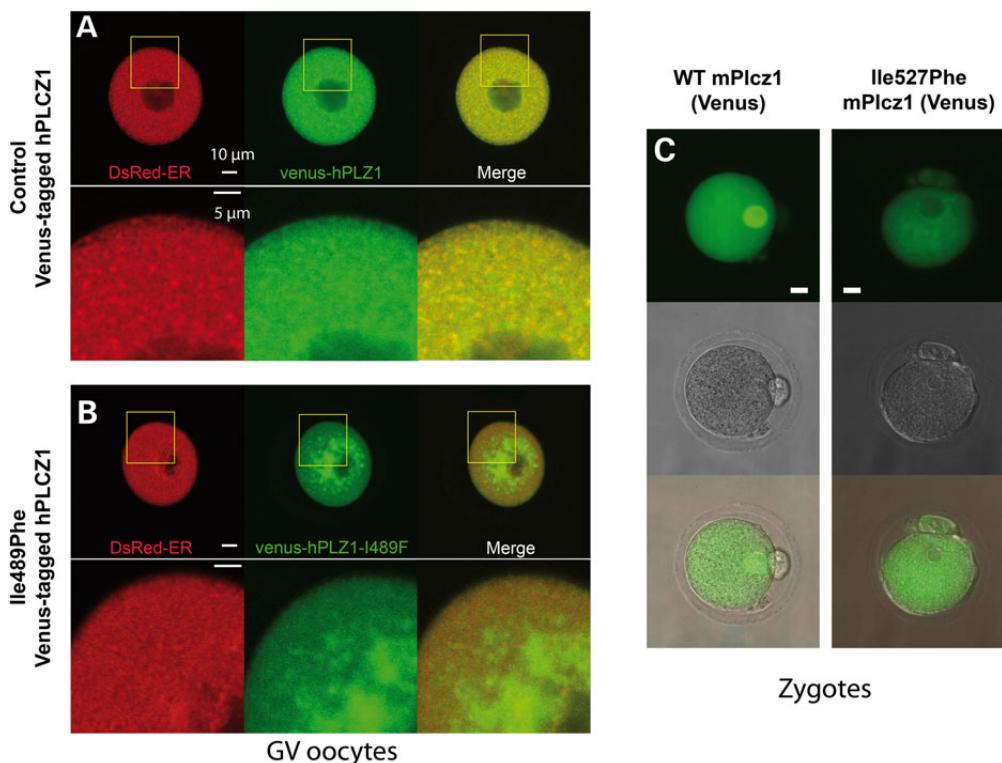


Figure 5. Mislocalization of mutant PLCZ1 following expression in GV and MII oocytes. (A) Venus-tagged human PLCZ1 cRNA (green) was co-injected in mouse GV stage oocytes with ER-DsRed cRNA, as a marker of the ER (red); the GV arrest was maintained by incubation in 100 μM IBMX. Both PLCZ1 and ER-DsRed exhibited a widespread punctiform staining throughout the cytoplasm and the overlay shows mostly overlapping distribution of the signals. (B) Co-injection of Ile489Phe PLCZ1 and ER-DsRed resulted in different distribution, with most of the PLCZ1 staining (green) concentrated around the nucleus (GV), whereas ER-DsRed remained uniformly distributed in the cytoplasm; there was minimal overlap of the signals at the oocyte periphery. (C) cRNAs encoding for WT mPlcz1 and the equivalent mutant in mice, Ile527Phe, were injected in MII oocytes and their ability to translocate into the PN compared. Whereas WT Plcz1 (left images) concentrated on the PN, Ile527Phe Plcz1 did not (right images).

heterozygosity, which reinforced the link between PLCZ1, OAF and infertility (45). Here, we show the first identified homozygous mutation of PLCZ1 leading to OAF and infertility. Importantly, it is the first exomic analysis of patients presenting OAF. We identified only four homozygous variants shared by both brothers: three of them were not expected to have any deleterious effects, which is in contrast to the missense mutation found in PLCZ1. Importantly, this mutation was not associated with teratozoospermia and the numbers of normal sperm were well above the lower accepted reference values (21). Moreover, the sperm of both brothers showed only slightly reduced DNA quality, ruling this out as the cause of OAF (46,47). Altogether, these results indicate that PLCZ1 plays a direct and primary role in the activation of mammalian oocytes. It is worth noting that a Plcz1 KO animal model does not presently exist, due to an early spermatogenesis arrest in this model (13), and this absence has been used to raise doubts about the role of PLCZ1 in oocyte activation. Here, we present the first functional knock-down of PLCZ1 in human sperm without effects on spermatogenesis and the results demonstrate a required role for oocyte activation and fertility in this species. Finally, unlike PLCZ1, the expression and localization of PAWP were unchanged, dismissing its contribution to the phenotype of our patients. This conclusion was reinforced by our results showing that sperm from the *Dpy19l2* KO mouse model despite exhibiting normal PAWP expression cannot induce oocyte activation after ICSI, as they lack Plcz1 (11,32). These results are also consistent with the report showing that PAWP null mice are fertile, their sperm show no morphological defects and that they can trigger oocyte activation (48). Our results thus do not support the notion

that PAWP triggers Ca²⁺ release and oocyte activation (15), and instead confirm the importance of PLCZ1 for this function in human.

In summary, whole-exome sequencing of two infertile brothers identified for the first time a homozygous mutation in the C2 domain of PLCZ1 responsible for ICSI failure and infertility. We also show PAWP expression was unaffected in these patients indicating it is unable to induce Ca²⁺ responses or oocyte activation. Therefore, given the required role of PLCZ1 for fertility and our findings demonstrating the significance of specific residues in the C2 domain for the enzyme's function, future studies should establish the molecular target(s) in the ooplasm and the host organelle that guide the enzyme to sites of accessible and abundant substrate, which is required to support the long-lasting oscillations that underlie egg activation in mammals.

Materials and Methods

Biological samples

Sperm were obtained following informed consent from patients consulting with the Department of fertility at Grenoble (France) or with the Clinique des Jasmins (Tunis, Tunisia) after approval by the Ethics committee of the university. In addition, all patients gave informed consent for preservation of unused sperm in the Germatheque biobank and subsequent use for studies on human fertility in accordance with the Helsinki Declaration on human experimentation. *Dpy19l2* KO mice were obtained from the Mutant Mouse Regional Resource Center, University of California, Davis, CA, USA.

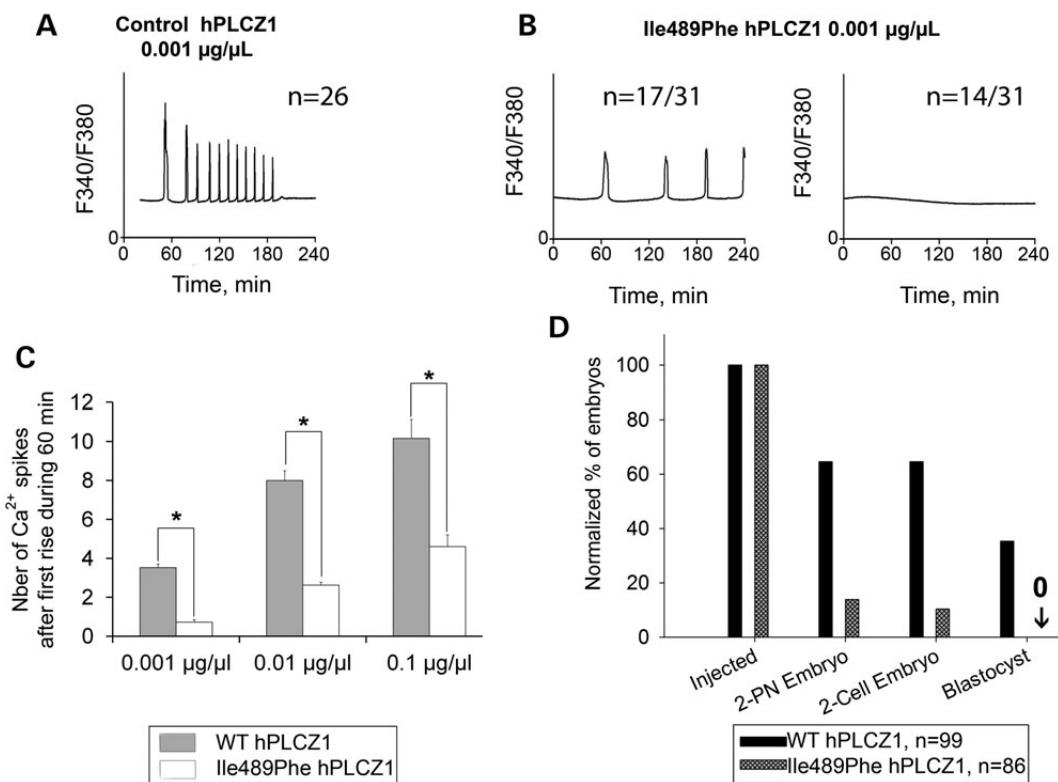


Figure 6. Ile489Phe PLCZ1 induces fewer Ca²⁺ oscillations compromising embryo development. (A) Injection of hPLCZ1 cRNA (0.001 μg/μL) in MII mouse oocytes triggered high-frequency Ca²⁺ oscillations, and between 6 and 10 rises per hour that end by 120–180 min after injection. (B) In contrast, injection of cRNA of Ile489Phe hPLCZ1 (0.001 μg/μL) triggered oscillations that were either delayed and with a low frequency (17/31) or failed to initiate them (14/31). (C) Histogram showing means + SD of the frequency of Ca²⁺ rises in function of control or Ile489Phe cRNA PLCZ1 concentrations injected into oocytes; asterisks indicate significant differences determined by t-test, $P < 0.01$. (D) Rates of 2PN, two-cell embryos and blastocysts obtained after injection of WT or Ile489Phe cRNA PLCZ1 (0.01 μg/μL) in mouse MII oocytes; n corresponds to the number of injected oocytes.

Exome sequencing and bioinformatics analysis

Genomic DNA was isolated from saliva using Oragen DNA extraction kit (DNAgenotech®, Ottawa, Canada). Coding regions and intron/exon boundaries were enriched using the ‘all Exon V5 kit’ (Agilent Technologies, Wokingham, UK). DNA sequencing was undertaken at the Genoscope, Evry, France, on the HiSeq 2000 from Illumina®. Sequence reads were aligned to the reference genome (hg19) using MAGIC (49). Duplicate reads and reads that mapped to multiple locations in the exome were excluded from further analysis. Positions whose sequence coverage was <10 on either the forward or reverse strand were excluded. Single-nucleotide variations and small insertions/deletions (indels) were identified and quality-filtered using in-house scripts. The most promising candidate variants were identified using an in-house bioinformatics pipeline. Variants with a minor allele frequency >5% in the NHLBI ESP6500 or in 1000 Genomes Project phase 1 data sets, or >1% in ExAC, were discarded. We also compared these variants with an in-house database of 56 control exomes obtained from subjects from the same geographic origin as our two patients (North Africa). All variants present in homozygous state in this database were excluded. We used variant effect predictor to predict the impact of the selected variants. We only retained variants impacting splice donor/acceptor or causing frameshift, inframe insertions/deletions, stop gain, stop loss or missense variants except those scored as ‘tolerated’ by SIFT (sift.jcvi.org) and as ‘benign’ by Polyphen-2 (genetics.bwh.harvard.edu/pph2).

Sanger sequencing

The presence of the identified variation was verified by Sanger sequencing of PLCZ1 exon 13. Primers were as followed: PLCZ1_14F: TCAATTTGTGGGAGCTGA and PLCZ1_14R: GGACATAATG-GAAAACCCTTG. Thirty-five cycles of polymerase chain reaction amplification were carried out with an hybridization temperature of 60°C. Sequencing reactions were carried out with BigDye Terminator v3.1 (Applied Biosystems). Sequence analysis were carried out on ABI 3130XL (Applied Biosystems).

Primary antibodies

Anti-human acrosin and PAWP antibodies were from Sigma-Aldrich and Proteintech, respectively; anti-PLCZ1 antibodies were raised against a 15-mer peptide sequence (305KETHERKGSDKRGDN319) of the human PLCZ1 (7), affinity purified and stored at a concentration of 1 μg/μL and used at 1/100 for IF and 1/1000 for WB.

Western blotting

WB was carried out as described by our laboratory (11). Briefly, sperm were first washed in phosphate-buffered saline (PBS) and resuspended in Laemmli sample buffer and boiled. Protein extracts equivalent to $1-2 \times 10^6$ sperm were loaded per lane into 4–20% sodium dodecyl sulfate–polyacrylamide gel electrophoresis, and resolved proteins were transferred onto PVDF membranes. The membranes were blocked and incubated

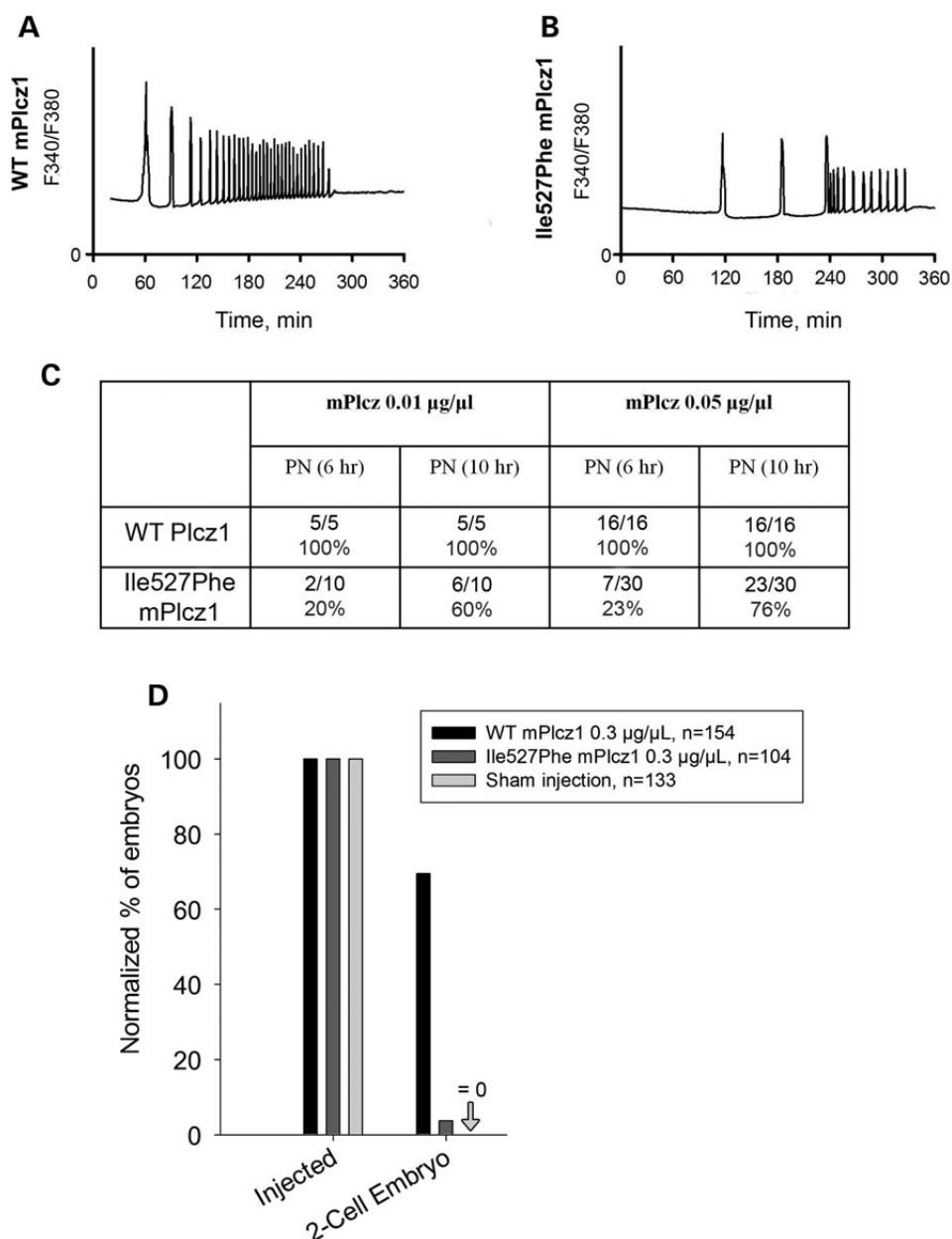


Figure 7. Ca^{2+} responses induced by Ile527Phe mouse Plcz1 are altered and embryo development is compromised. (A) Profile of Ca^{2+} responses induced by injection of WT mPlcz1 cRNA (0.01 $\mu\text{g}/\mu\text{l}$) into mouse oocytes. (B) Ca^{2+} profiles recorded after injection of Ile527Phe mPlcz1 cRNA (the mouse equivalent of Ile489Phe hPLCZ1) show delayed and reduced response. (C) 2PN formation rates are reduced and delayed when both Ile527Phe mPlcz1 0.01 and 0.05 $\mu\text{g}/\mu\text{l}$ concentrations are used. (D) Injection of Ile527Phe mPlcz1 cRNAs does not support embryo development, as the rate of two-cell embryos formation was strongly reduced or abolished; n corresponds to the number of injected MII oocytes.

overnight with anti-PLCZ1 Ab (1/1000) and next with horseradish peroxidase labeled secondary Ab (1 h). Immunoreactivity was detected using chemiluminescence detection kit reagents and a Chimidoc™ Station (Bio-Rad).

Immunofluorescence

IF was carried out as described by our laboratory (11). Sperm were fixed with 4% paraformaldehyde, washed in PBS, and spotted onto poly-L-lysine precoated slides. Sperm were then permeabilized with 0.1% (v/v) Triton X-100. Slides were then blocked in 5% normal goat serum–Dulbecco's PBS (DPBS) (GIBCO, Invitrogen) and incubated overnight at 4°C with primary antibodies. Washes

were performed with 0.1% (v/v) Tween 20–DPBS, followed by 1 h incubation with secondary Ab (1/400). Samples were counterstained with Hoechst 33342 and mounted with DAKO mounting media (Life Technology). Fluorescence images were captured with a confocal microscope (Zeiss LSM 710).

Generation of constructs and preparation of cRNA

Human and mouse PLCZ1 constructs were kind gifts from Dr K. Fukami (Tokyo University of Pharmacy and Life Science, Japan) and Dr K. Jones (University of Southampton, UK), respectively. WT h and mPLCZ1-venus sequences were subcloned into a pcDNA6/Myc-His B (Invitrogen) between EcoRI and XbaI

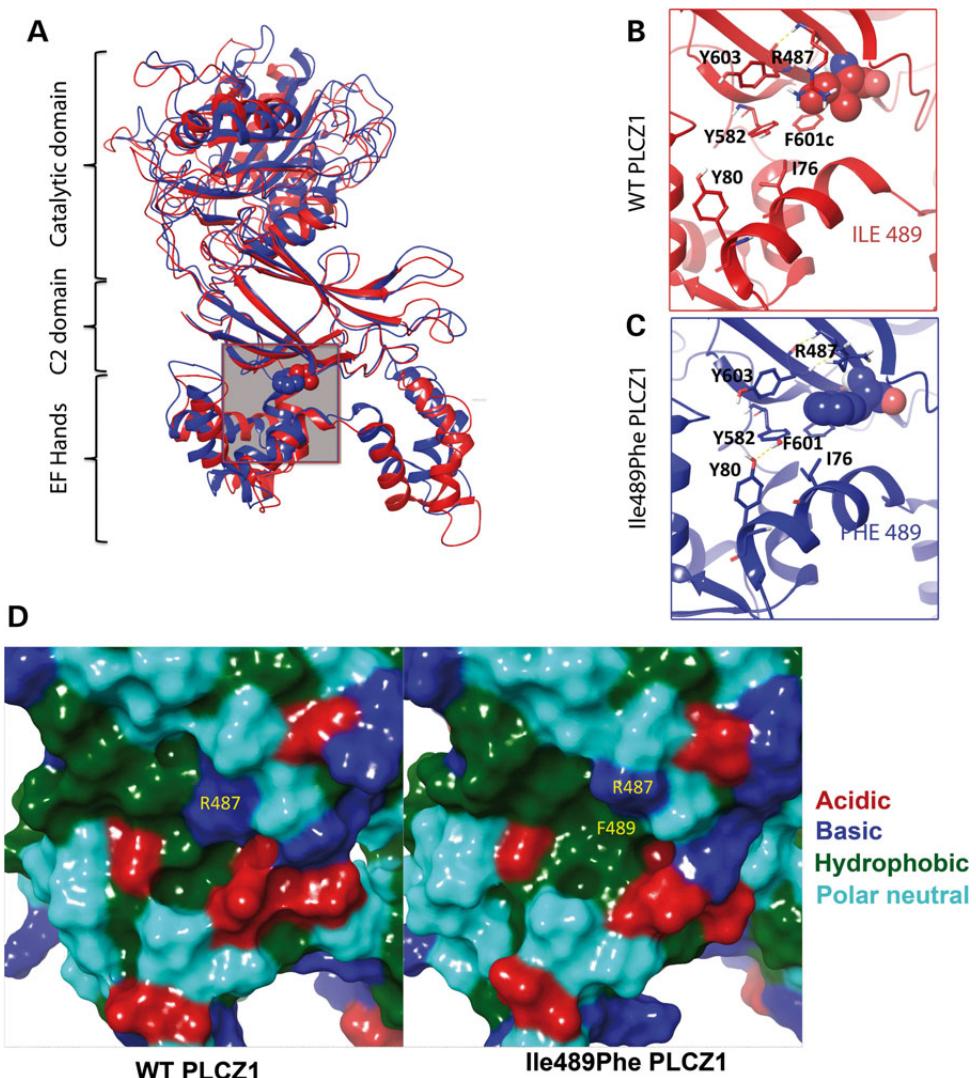


Figure 8. Molecular dynamic simulation of the effect of the Ile489Phe mutation on the conformation of PLCZ1. (A) The site of the Ile489Phe mutation is located at the interface of C2 and EF-hand domains of PLCZ1 (gray box) (B and C) Enlargement of the red/blue boxes for WT and mutant PLCZ1. As suggested by molecule dynamics simulation, the mutation results in a displacement of the surrounding C2 residues (Y582, F601, Y603 and R487) by the larger side chain of Phe. This reorganization, in turn, leads to novel contacts between C2 and EF2 hand via persistent hydrophobic (Phe489–Ile76) and H-bonding (Tyr582–Tyr80) side-chain interactions, which shifts the entire EF-hand domain toward C2 by an ~1.5 Å by the end of the 1.2-ns simulation. (D) The molecular surfaces of WT and mutant PLCZ1 in the vicinity of the 489 residue, demonstrating the formation of a novel hydrophobic sub-site at the C2/EF-hand junction. The basic, acidic hydrophobic and polar neutral residues shown in blue, red, green and cyan, respectively.

restrictions sites. hPLCZ1-Ile489Phe and mPlcz1 Ile527Phe were generated by substituting Ile to Phe using the Gibson Assembly Cloning Kit (New England Bio Labs), as previously reported (50). All constructs were finally sequenced. pDsRed2-ER was kindly provided by Dr M Trebak (Penn State Hershey College of Medicine, PA, USA). The ER-targeting sequence of calreticulin, DsRed2 and the KDEL ER retention sequence were ligated to pcDNA6/Myc-His B. Plasmids were linearized outside of the coding region with PmeI and in vitro transcribed using mMESSAGE/mMACHINE T7 Kit (Ambion). Poly-A tail was added to the mRNAs using a Tailing Kit (Ambion).

Confocal microscopy of fluorescent-PLCZ1

Live-cell images of oocytes and zygotes expressing fluorescently tagged proteins were captured with a confocal laser-scanning

microscope (LSM 510 META, Carl Zeiss) using a 63×1.4 numerical aperture oil-immersion objective lens. Images were reconstructed using the LSM software (Carl Zeiss). Oocytes were maintained in HCZB medium and those expressing Venus-hPLCZ1 and DsRed-ER proteins were imaged at the GV stage, whereas expression of Venus-mPlcz1 and DsRed-ER proteins was imaged in PN stage zygotes.

Ca²⁺ monitoring

Ca²⁺ monitoring was carried out as described by our laboratory (7). Briefly, mouse oocytes were loaded with fura-2-AM (molecular probes) prior to injecting the cRNAs, after which they were transferred into a monitoring dish containing 50 µl drops of TL-HEPES medium under mineral oil. Excitation wavelengths of 340/380 nm were alternated using a filter wheel (Ludl Electronic

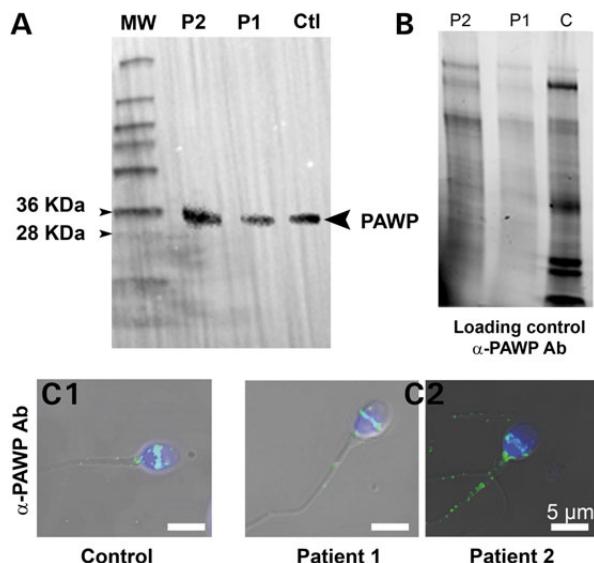


Figure 9. Normal expression and localization of PAWP in patients with OAF. (A) WB using protein extracts from sperm of a fertile control (Ctl), P1 and P2 and an anti-PAWP antibody. Immunoreactivity is observed in the lanes of control and patients' sperm extracts. (B) Protein gel representing loading control for (A) showing that all lanes were similarly loaded. Protein loads were controlled with TGX stain free™ precast gels. (C1) Sperm from a fertile control were stained with anti-PAWP antibody and counterstained with Hoechst. Overlay shows that the antibody stains strongly the post-acrosomal area. (C2) Similar staining of sperm from P1 and P2, showing the same reactivity to that of the control fertile sperm.

Products Ltd). Fluorescence ratios were taken every 20 s. After passing through a 510 nm barrier filter, the emitted light was collected by a cooled Photometrics SenSys CCD camera (Roper Scientific). The SimplePCI imaging software was used to run all the hardware and capture images (Hamamatsu). $[Ca^{2+}]_i$ values are reported as the ratio of 340/380 nm fluorescence in the whole oocyte.

Parthenogenetic oocyte activation

Oocytes were activated by injection of PLCZ1 cRNA into the ooplasm (concentration as indicated). After cRNA microinjection, oocytes were cultured in KSOM/EAA, supplemented with 5 μ g/ml cytochalasin B to diploidize the parthenotes. The injection volume was \sim 5–10 pl, which is \sim 1–3% of the total oocyte volume. PN formation was checked 6 h after injection and development was followed up to the blastocyst stage.

Homology modeling

A homology model for hPLCZ1 was generated with the Prime homology modeling workflow (version 3.8, Schrödinger, LLC) using the structure of rPLC81, in complex with inositol-1,4,5-trisphosphate and associated Ca^{2+} (PDB code: 1DJX) as the template. Briefly, all sequences (1DJX, rPLCD1 and hPLCZ1) were initially aligned by the BLAST homology search. First, a contiguous template structure was generated by replacing the missing 445–486 loop in the rPLCD1 structure with a peptide bond between the proximal (per 1DJX) G444 and K487 and minimization of the resulting loop using the OPLS2005 force field within Prime. Next, a similar operation resulted in the exclusion of the basic loop from the catalytic domain of hPLCZ1 (304–344) by linking

residues 303 and 345 to recreate a contiguous mode. A single-chain, liganded model, comprising the residues 64–303 and 345–608, was then built by the energy-based algorithm in Prime to construct and refine non-identical residues and loops with deletions and insertions.

Molecular dynamics simulations

Both the resulting complex and pre-minimized I489F mutant were subjected to an unconstrained 1.2 ns, molecular dynamics procedure within Desmond (version 4.0, D. E. Shaw Research & Schrödinger, LLC). An NPT ensemble was built at 300 K and 1 atm with the neutralized (by 10 Cl^- ions) system and further Na^+/Cl^- ions to simulate 150 mM concentration in an explicit SPC solvent model. The interactions that both I489 and the corresponding phenylalanine were participating throughout the simulations were identified using the simulated interaction diagram routine within Desmond. Topological changes were detected using simulated event analysis program by monitoring the $C\alpha_{176}-C\alpha_{1489}$ and $C\alpha_{176}-C\alpha_{449}$ distances and root mean square fluctuations as a function of simulation time. The mutant structure from the final frame was modified back to contain I489, and a simulation under the identical conditions was conducted to ascertain the persistence of the mutant's conformation.

Supplementary Material

Supplementary Material is available at HMG online.

Authors' Contributions

R.Z., L.H. and C.T. identified patients, J.E., S.Y., H.C.L. and G.M. performed research; C.C., T.K., Z.K., N.T.-M. contributed to genetic analyses; S.N.S. performed PLCZ1 structure analyses; R.F., P.F.R. and C.A. designed research, analyzed data and wrote the paper.

Acknowledgements

We thank S. Crouzy (CEA Grenoble) for valuable discussions.

Conflict of Interest statement. None declared.

Funding

This work was supported by following grants: ANR Genopat 2009, project ICG2I to P.R. and C.A. and NIH R01 HD051872 to R.A.F.

References

- Steinhardt, R.A., Epel, D., Carroll, E.J. Jr. and Yanagimachi, R. (1974) Is calcium ionophore a universal activator for unfertilised eggs? *Nature*, **252**, 41–43.
- Ridgway, E.B., Gilkey, J.C. and Jaffe, L.F. (1977) Free calcium increases explosively in activating medaka eggs. *Proc. Natl Acad. Sci. USA*, **74**, 623–627.
- Swann, K. (1990) A cytosolic sperm factor stimulates repetitive calcium increases and mimics fertilization in hamster eggs. *Development*, **110**, 1295–1302.
- Stice, S.L. and Robl, J.M. (1990) Activation of mammalian oocytes by a factor obtained from rabbit sperm. *Mol. Reprod. Dev.*, **25**, 272–280.
- Parrington, J., Jones, K.T., Lai, A. and Swann, K. (1999) The soluble sperm factor that causes Ca^{2+} release from sea-urchin

- (*Lytechinus pictus*) egg homogenates also triggers Ca^{2+} oscillations after injection into mouse eggs. *Biochem. J.*, **341**, 1–4.
6. Jones, K.T., Cruttwell, C., Parrington, J. and Swann, K. (1998) A mammalian sperm cytosolic phospholipase C activity generates inositol triphosphate and causes Ca^{2+} release in sea urchin egg homogenates. *FEBS Lett.*, **437**, 297–300.
 7. Yoon, S.Y., Jellerette, T., Salicioni, A.M., Lee, H.C., Yoo, M.S., Coward, K., Parrington, J., Grow, D., Cibelli, J.B., Visconti, P.E. et al. (2008) Human sperm devoid of PLC zeta 1 fail to induce Ca^{2+} release and are unable to initiate the first step of embryo development. *J. Clin. Invest.*, **118**, 3671–3681.
 8. Heytens, E., Parrington, J., Coward, K., Young, C., Lambrecht, S., Yoon, S.Y., Fissore, R.A., Hamer, R., Deane, C.M., Ruas, M. et al. (2009) Reduced amounts and abnormal forms of phospholipase C zeta (PLCzeta) in spermatozoa from infertile men. *Hum. Reprod.*, **24**, 2417–2428.
 9. Taylor, S.L., Yoon, S.Y., Morshedi, M.S., Lacey, D.R., Jellerette, T., Fissore, R.A. and Oehninger, S. (2010) Complete globozoospermia associated with PLCzeta deficiency treated with calcium ionophore and ICSI results in pregnancy. *Reprod. Biomed. Online*, **20**, 559–564.
 10. Amdani, S.N., Jones, C. and Coward, K. (2013) Phospholipase C zeta (PLCzeta): oocyte activation and clinical links to male factor infertility. *Adv. Biol. Regul.*, **53**, 292–308.
 11. Escouffier, J., Yassine, S., Lee, H.C., Martinez, G., Delaroche, J., Coutton, C., Karaouzene, T., Zouari, R., Metzler-Guillemain, C., Pernet-Gallay, K. et al. (2015) Subcellular localization of phospholipase Czeta in human sperm and its absence in DPY19L2-deficient sperm are consistent with its role in oocyte activation. *Mol. Hum. Reprod.*, **21**, 157–168.
 12. Nomikos, M., Kashir, J., Swann, K. and Lai, F.A. (2013) Sperm PLCzeta: from structure to Ca^{2+} oscillations, egg activation and therapeutic potential. *FEBS Lett.*, **587**, 3609–3616.
 13. Ito, M., Nagaoka, K., Kuroda, K., Kawano, N., Yoshida, K., Harada, Y., Shikano, T., Miyado, M., Oda, S., Toshimori, K. et al. (2010) Arrest of spermatogenesis at round spermatids in PLCz1 deficient mice. Abstracts of the 11th International Symposium on Spermatology; June 26–29, 2010; Okinawa, Japan.
 14. Aarabi, M., Yu, Y., Xu, W., Tse, M.Y., Pang, S.C., Yi, Y.J., Sutovsky, P. and Oko, R. (2012) The testicular and epididymal expression profile of PLCzeta in mouse and human does not support its role as a sperm-borne oocyte activating factor. *PLoS One*, **7**, e33496.
 15. Aarabi, M., Balakier, H., Bashar, S., Moskovtsev, S.I., Sutovsky, P., Librach, C.L. and Oko, R. (2014) Sperm-derived WW domain-binding protein, PAWP, elicits calcium oscillations and oocyte activation in humans and mice. *FASEB J.*, **28**, 4434–4440.
 16. Aarabi, M., Balakier, H., Bashar, S., Moskovtsev, S.I., Sutovsky, P., Librach, C.L. and Oko, R. (2014) Sperm content of postacrosomal WW binding protein is related to fertilization outcomes in patients undergoing assisted reproductive technology. *Fertil. Steril.*, **102**, 440–447.
 17. Wu, A.T., Sutovsky, P., Manandhar, G., Xu, W., Katayama, M., Day, B.N., Park, K.W., Yi, Y.J., Xi, Y.W., Prather, R.S. and Oko, R. (2007) PAWP, a sperm-specific WW domain-binding protein, promotes meiotic resumption and pronuclear development during fertilization. *J. Biol. Chem.*, **282**, 12164–12175.
 18. Nomikos, M., Sanders, J.R., Theodoridou, M., Kashir, J., Matthews, E., Nouresis, G., Lai, F.A. and Swann, K. (2014) Sperm-specific post-acrosomal WW-domain binding protein (PAWP) does not cause Ca^{2+} release in mouse oocytes. *Mol. Hum. Reprod.*, **20**, 938–947.
 19. Nomikos, M., Sanders, J.R., Kashir, J., Sanusi, R., Buntwal, L., Love, D., Ashley, P., Sanders, D., Knaggs, P., Bunkheila, A. et al. (2015) Functional disparity between human PAWP and PLCzeta in the generation of Ca^{2+} oscillations for oocyte activation. *Mol. Hum. Reprod.*, **21**, 702–710.
 20. Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M. et al. (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.
 21. World Health Organization (2010) WHO laboratory manual for the Examination and processing of human semen. 5th edn. WHO Press, Geneva, Switzerland.
 22. Sharma, R.K., Sabanegh, E., Mahfouz, R., Gupta, S., Thiagarajan, A. and Agarwal, A. (2010) TUNEL as a test for sperm DNA damage in the evaluation of male infertility. *Urology*, **76**, 1380–1386.
 23. Ribas-Maynou, J., Garcia-Peiro, A., Fernandez-Encinas, A., Abad, C., Amengual, M.J., Prada, E., Navarro, J. and Benet, J. (2013) Comprehensive analysis of sperm DNA fragmentation by five different assays: TUNEL assay, SCSA, SCD test and alkaline and neutral Comet assay. *Andrology*, **1**, 715–722.
 24. Erenpreiss, J., Bars, J., Lipatnikova, V., Erenpreisa, J. and Zalkalns, J. (2001) Comparative study of cytochemical tests for sperm chromatin integrity. *J. Androl.*, **22**, 45–53.
 25. Lee, H.C., Arny, M., Grow, D., Dumesic, D., Fissore, R.A. and Jellerette-Nolan, T. (2014) Protein phospholipase C Zeta1 expression in patients with failed ICSI but with normal sperm parameters. *J. Assist. Reprod. Genet.*, **31**, 749–756.
 26. Ananthanarayanan, B., Das, S., Rhee, S.G., Murray, D. and Cho, W. (2002) Membrane targeting of C2 domains of phospholipase C-delta isoforms. *J. Biol. Chem.*, **277**, 3568–3575.
 27. Yu, Y., Nomikos, M., Theodoridou, M., Nouresis, G., Lai, F.A. and Swann, K. (2012) PLCzeta causes Ca^{2+} oscillations in mouse eggs by targeting intracellular and not plasma membrane PI(4,5)P2. *Mol. Biol. Cell*, **23**, 371–380.
 28. Yoda, A., Oda, S., Shikano, T., Kouchi, Z., Awaji, T., Shirakawa, H., Kinoshita, K. and Miyazaki, S. (2004) Ca^{2+} oscillation-inducing phospholipase C zeta expressed in mouse eggs is accumulated to the pronucleus during egg activation. *Dev. Biol.*, **268**, 245–257.
 29. Larman, M.G., Saunders, C.M., Carroll, J., Lai, F.A. and Swann, K. (2004) Cell cycle-dependent Ca^{2+} oscillations in mouse embryos are regulated by nuclear targeting of PLCzeta. *J. Cell Sci.*, **117**, 2513–2521.
 30. Kuroda, K., Ito, M., Shikano, T., Awaji, T., Yoda, A., Takeuchi, H., Kinoshita, K. and Miyazaki, S. (2006) The role of X/Y linker region and N-terminal EF-hand domain in nuclear translocation and Ca^{2+} oscillation-inducing activities of phospholipase Czeta, a mammalian egg-activating factor. *J. Biol. Chem.*, **281**, 27794–27805.
 31. Nomikos, M., Sanders, J.R., Parthimos, D., Buntwal, L., Calver, B.L., Stamatiadis, P., Smith, A., Clue, M., Sideratou, Z., Swann, K. and Lai, F.A. (2015) Essential role of the EF-hand domain in targeting sperm phospholipase Czeta to membrane PIP2. *J. Biol. Chem.*, **290**, 29519–29530.
 32. Yassine, S., Escouffier, J., Martinez, G., Coutton, C., Karaouzene, T., Zouari, R., Ravantat, J.L., Metzler-Guillemain, C., Fissore, R., Hennebicq, S. et al. (2015) Dpy19l2-deficient globozoospermic sperm display altered genome packaging and DNA damage that compromises the initiation of embryo development. *Mol. Hum. Reprod.*, **21**, 169–185.
 33. Breucker, H., Schafer, E. and Holstein, A.F. (1985) Morphogenesis and fate of the residual body in human spermiogenesis. *Cell Tissue Res.*, **240**, 303–309.

34. Saunders, C.M., Larman, M.G., Parrington, J., Cox, L.J., Royse, J., Blayney, L.M., Swann, K. and Lai, F.A. (2002) PLC zeta: a sperm-specific trigger of Ca(2+) oscillations in eggs and embryo development. *Development*, **129**, 3533–3544.
35. Bommert, K., Charlton, M.P., DeBello, W.M., Chin, G.J., Betz, H. and Augustine, G.J. (1993) Inhibition of neurotransmitter release by C2-domain peptides implicates synaptotagmin in exocytosis. *Nature*, **363**, 163–165.
36. Clark, J.D., Lin, L.L., Kriz, R.W., Ramesha, C.S., Sultzman, L.A., Lin, A.Y., Milona, N. and Knopf, J.L. (1991) A novel arachidonic acid-selective cytosolic PLA2 contains a Ca(2+)-dependent translocation domain with homology to PKC and GAP. *Cell*, **65**, 1043–1051.
37. Sutton, R.B., Davletov, B.A., Berghuis, A.M., Sudhof, T.C. and Sprang, S.R. (1995) Structure of the first C2 domain of synaptotagmin I: a novel Ca²⁺/phospholipid-binding fold. *Cell*, **80**, 929–938.
38. Essen, L.O., Perisic, O., Cheung, R., Katan, M. and Williams, R. L. (1996) Crystal structure of a mammalian phosphoinositide-specific phospholipase C delta. *Nature*, **380**, 595–602.
39. Cho, W. and Stahelin, R.V. (2006) Membrane binding and subcellular targeting of C2 domains. *Biochim. Biophys. Acta*, **1761**, 838–849.
40. Kouchi, Z., Shikano, T., Nakamura, Y., Shirakawa, H., Fukami, K. and Miyazaki, S. (2005) The role of EF-hand domains and C2 domain in regulation of enzymatic activity of phospholipase Czeta. *J. Biol. Chem.*, **280**, 21015–21021.
41. Stahelin, R.V., Rafter, J.D., Das, S. and Cho, W. (2003) The molecular basis of differential subcellular localization of C2 domains of protein kinase C-alpha and group IVa cytosolic phospholipase A2. *J. Biol. Chem.*, **278**, 12452–12460.
42. Corbalan-Garcia, S. and Gomez-Fernandez, J.C. (2014) Signaling through C2 domains: more than one lipid target. *Biochim. Biophys. Acta*, **1838**, 1536–1547.
43. Piazzì, M., Blalock, W.L., Bavelloni, A., Faenza, I., D'Angelo, A., Maraldi, N.M. and Cocco, L. (2013) Phosphoinositide-specific phospholipase C beta 1b (PI-PLC β 1b) interactome: affinity purification-mass spectrometry analysis of PI-PLC β 1b with nuclear protein. *Mol. Cell Proteomics*, **12**, 2220–2235.
44. Coutton, C., Escoffier, J., Martinez, G., Arnoult, C. and Ray, P.F. (2015) Teratozoospermia: spotlight on the main genetic actors in the human. *Hum. Reprod. Update*, **21**, 455–485.
45. Kashir, J., Konstantinidis, M., Jones, C., Lemmon, B., Chang, L. H., Hamer, R., Heindryckx, B., Deane, C.M., De Sutter, P., Fissore, R.A. et al. (2012) A maternally inherited autosomal point mutation in human phospholipase C zeta (PLC ζ) leads to male infertility. *Hum. Reprod.*, **27**, 222–231.
46. Marianov, I., Brancorsini, S., Catena, R., Gansmuller, A., Kotaja, N., Parvinen, M., Sassone-Corsi, P. and Davidson, I. (2005) Polar nuclear localization of H1T2, a histone H1 variant, required for spermatid elongation and DNA condensation during spermiogenesis. *Proc. Natl Acad. Sci. USA*, **102**, 2808–2813.
47. Zhao, M., Shirley, C.R., Hayashi, S., Marcon, L., Mohapatra, B., Suganuma, R., Behringer, R.R., Boissonneault, G., Yanagimachi, R. and Meistrich, M.L. (2004) Transition nuclear proteins are required for normal chromatin condensation and functional sperm development. *Genesis*, **38**, 200–213.
48. Satoh, Y., Nozawa, K. and Ikawa, M. (2015) Sperm postacrosomal WW domain-binding protein is not required for mouse egg activation. *Biol. Reprod.*, **93**, 94–100.
49. SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903–914.
50. Gibson, D.G., Young, L., Chuang, R.Y., Venter, J.C., Hutchison, C.A. III and Smith, H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.

Principaux résultats

Suite à l'analyse bio-informatique de ces deux frères, un seul variant subsistait après l'application de l'ensemble des filtres. Celui-ci était recensé uniquement dans la base de données ExAC avec une fréquence de 8.24e-06 et entraînait un faux-sens prédict comme *deleterious* par SIFT et *possibly damaging* par PolyPhen sur la séquence du gène *PLCZ1*. La forte expression testiculaire de ce gène (**Figure** : 2.14) couplée à l'implication déjà connue de celui-ci dans l'activation ovocytaire, ont fait de ce variant le candidat évident pour expliquer le phénotype de ces deux frères. De plus, aucun variant n'a été retrouvé sur la séquence du gène *WBP2NL* codant pour la protéine PAWP (l'autre gène candidat à la fonction d'activateur ovocytaire) bien que l'intégralité de la séquence codante de *WBP2NL* ait une couverture $\geq 40x$ (les zones moins couvertes du début de l'exon 1 et de la fin de l'exon 6 correspondant aux régions UTR) (**Figure** : 2.15). Ces résultats suggérant une parfaite fonctionnalité de la protéine PAWP ont pu être confirmés par *Western blot*, de même, la bonne localisation de la protéine PAWP a pu être observée chez les deux patients par Immunofluorescence alors que *PLCZ1* était absent du sperme de nos patients.

Cette étude confirme le rôle primordial de *PLCZ1* dans l'activation ovocytaire et démontre que la présence de PAWP seul ne permet pas cette activation.

Table 2.5 – Liste des variants ayant passé l'ensemble des filtres pour les deux frères de la famille FF

Gene	HGVSc, HGVSp	Impact			PolyPhen	Frequency
		Consequence	SIFT			
PLCZ1	c.1465G>T ; p.Ile489Phe	missense	deleterious	possib damaging		8.24e-06

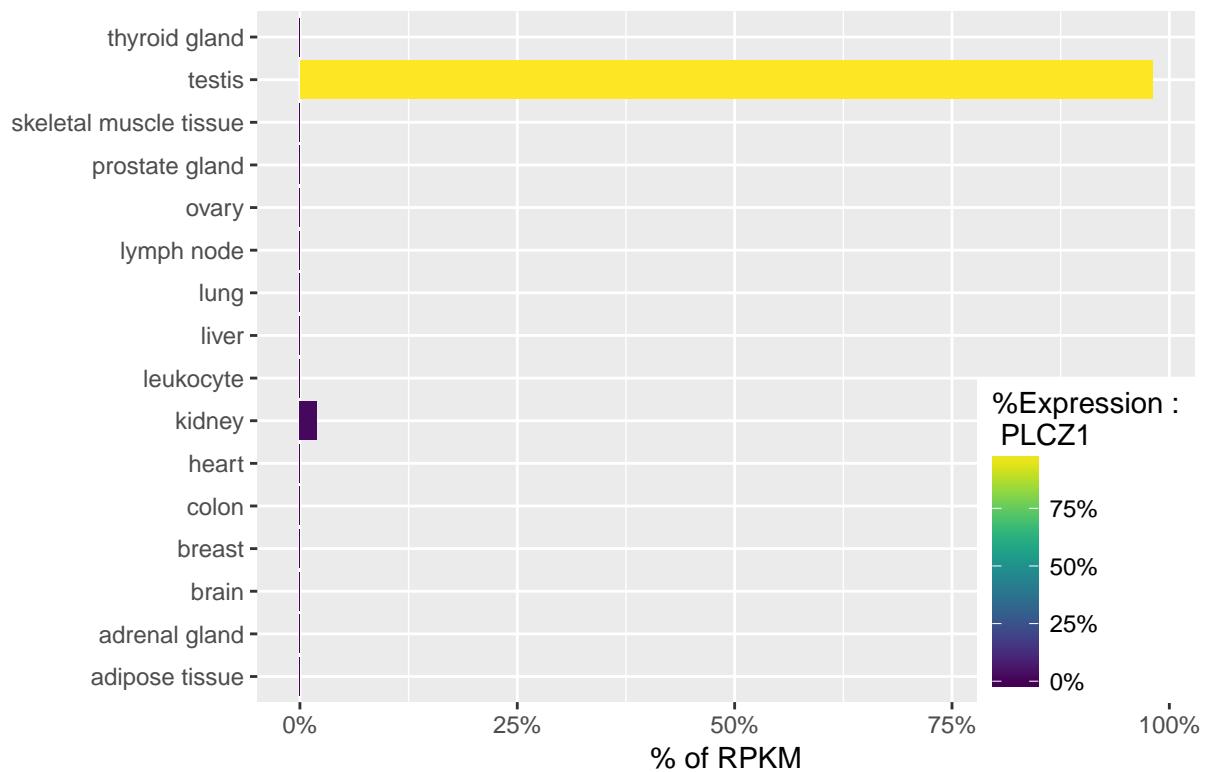


Figure 2.14 – Expression tissulaire du gène PLCZ1 : Données provenant du projet de transcriptome Illumina BodyMap

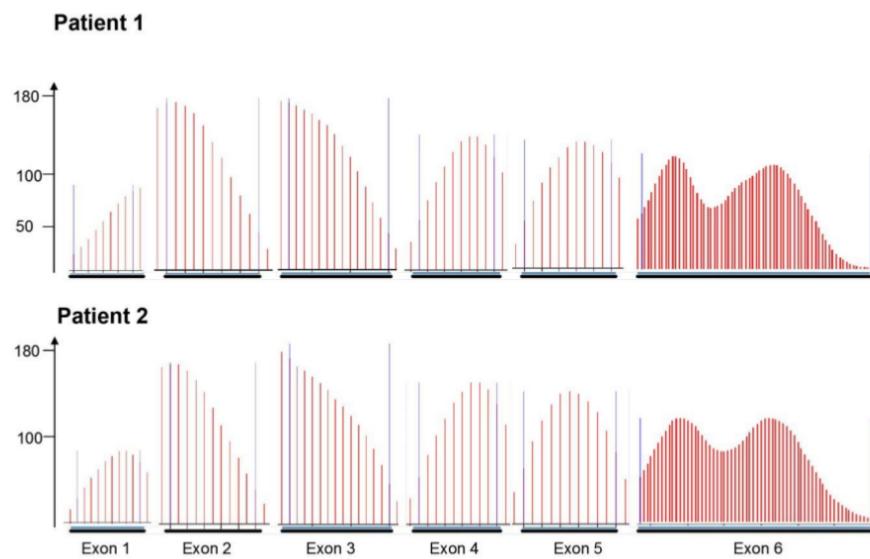


Figure 2.15 – Couverture des six exons de WBP2NL pour les deux frères de la famille FF : Les barres rouges représentent la couverture moyenne de 10 nucléotides, les bleues représentent les bornes de chaque exon.

2.2.4 Article n°3

Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations

Amiri-Yekta A*, Coutton C*, Kherraf ZE, **Karaouzène T**, Le Tanno P, Sanati MH, Sabbaghian M, Almadani N, Sadighi Gilani MA, Seyedeh Hanieh Hosseini, Bahrami S, Daneshipour A, Bini M, Arnoult C, Colombo R, Gourabi H, Ray PF

* Co-premiers auteurs

Human Reproduction, Octobre 2016

Contexte et objectifs

Dans une étude précédente non détaillée dans ce manuscrit (en annexe), notre équipe a pu identifier *DNAH1* comme le premier gène codant pour une dynéine axonémale responsable uniquement d'infertilité masculine. Dans cette première étude, 5 de nos 18 patients non-apparentés, soit environ 28% d'entre eux, étaient porteur d'une mutation homozygote sur le gène *DNAH1* responsable de leur phénotype MMAF. Ces résultats ont ainsi démontré l'importance de l'implication de ce gène dans ce phénotype.

Dans cette nouvelle étude, nous nous concentrons sur l'analyse des données génétiques de 5 familles iraniennes et d'une famille italienne soit un total de 12 individus parmi lesquels 10 ont été séquencés en WES. Cependant, pour des raisons techniques, seules les données de 9 d'entre eux ont été analysées sur notre pipeline décrit précédemment.

Dans ce contexte j'ai effectué l'ensemble des analyses bio-informatiques de ces patients.

Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new *DNAH1* mutations

Amir Amiri-Yekta^{1,2,†}, Charles Coutton^{2,3,†},
Zine-Eddine Kherraf^{2,4}, Thomas Karaouzène^{2,4},
Pauline Le Tanno^{2,3}, Mohammad Hossein Sanati^{1,5},
Marjan Sabbaghian⁶, Navid Almadani¹, Mohammad Ali Sadighi Gilani⁶,
Seyedeh Hanieh Hosseini⁷, Salahadin Bahrami¹, Abbas Daneshipour¹,
Maurizio Bini⁷, Christophe Arnoult², Roberto Colombo^{8,9},
Hamid Gourabi^{1,†#}, and Pierre F. Ray^{2,4,**}

¹Department of Genetics, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, PO Box 16635-148, Tehran, Iran ²Genetic Epigenetic and Therapies of Infertility, Institute for Advanced Biosciences, INSERM 1209, CNRS UMR 5309, Université Grenoble Alpes, Grenoble F38000, France ³CHU de Grenoble, UF de Génétique Chromosomique, Grenoble F-38000, France ⁴CHU de Grenoble, UF de Biochimie Génétique et Moléculaire, Grenoble F-38000, France ⁵Department of Medical Genetics, Institute of Medical Biotechnology, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran ⁶Department of Andrology, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, PO Box 16635-148, Tehran, Iran ⁷Center for the Study and Treatment of Fertility Disorders, Niguarda Ca' Granda Metropolitan Hospital, Milan, Italy ⁸Faculty of Medicine, Institute of Clinical Biochemistry, Catholic University, Rome, Italy ⁹Center for the Study of Rare Hereditary Diseases, Niguarda Ca' Granda Metropolitan Hospital, Milan, Italy

*Correspondence address. UF de Biochimie et Génétique Moléculaire, CHU de Grenoble, 38043 Grenoble cedex 9, France. Tel: +33-4-76-76-55-73; E-mail: pray@chu-grenoble.fr

Submitted on July 15, 2016; resubmitted on September 9, 2016; accepted on September 15, 2016

STUDY QUESTION: Can whole-exome sequencing (WES) of patients with multiple morphological abnormalities of the sperm flagella (MMAF) identify causal mutations in new genes or mutations in the previously identified dynein axonemal heavy chain 1 (*DNAH1*) gene?

SUMMARY ANSWER: WES for six families with men affected by MMAF syndrome allowed the identification of *DNAH1* mutations in four affected men distributed in two out of the six families but no new candidate genes were identified.

WHAT IS KNOWN ALREADY: Mutations in *DNAH1*, an axonemal inner dynein arm heavy chain gene, have been shown to be responsible for male infertility due to a characteristic form of asthenozoospermia called MMAF, defined by the presence in the ejaculate of spermatozoa with a mosaic of flagellar abnormalities including absent, coiled, bent, angulated, irregular and short flagella.

STUDY DESIGN, SIZE, DURATION: This was a retrospective genetics study of patients presenting a MMAF phenotype. Patients were recruited in Iran and Italy between 2008 and 2015.

PARTICIPANTS/MATERIALS, SETTING, METHODS: WES was performed for a total of 10 subjects. All identified variants were confirmed by Sanger sequencing. Two additional affected family members were analyzed by direct Sanger sequencing. To establish the

†These authors are equally contributing first authors.

#These authors have co-leadership.

prevalence of the *DNAH1* mutation identified in an Iranian family, we carried out targeted sequencing on 38 additional MMAF patients of the same geographical origin. RT-PCR and immunochemistry were performed on sperm samples to assess the effect of the identified mutation on RNA and protein.

MAIN RESULTS AND THE ROLE OF CHANCE: WES in six families identified a causal mutations in two families. Two additional affected family members were confirmed to hold the same homozygous mutation as their sibling. In total, *DNAH1* mutations were identified in 5 out of 12 analyzed subjects (41.7%). If we only include index cases, we detected two mutated subjects out of six (33%) tested MMAF individuals. Furthermore we sequenced one *DNAH1* exon found to be mutated (c.8626-1G > A) in an Iranian family in an additional 38 MMAF patients from Iran. One of these patients carried the variant confirming that this variant is relatively frequent in the Iranian population. The effect of the c.8626-1G > A variant was confirmed by RT-PCR and immunochemistry as no RNA or protein could be observed in sperm from the affected men.

LARGE SCALE DATA: N/A.

LIMITATIONS, REASONS FOR CAUTION: WES allows the amplification of 80–90% of all coding exons. It is possible that some *DNAH1* exons may not have been sequenced and that we may have missed some additional mutations. Also, WES cannot identify deep intronic mutations and it is not efficient for detection of large genomic events (deletions, insertions, inversions). We did not identify any causal mutations in *DNAH1* or in other candidate genes in four out of the six tested families. This indicates that the technique and/or the analysis of our data can be improved to increase the diagnosis efficiency.

WIDER IMPLICATIONS OF THE FINDINGS: Our findings confirm that *DNAH1* is one of the main genes involved in MMAF syndrome. It is a large gene with 78 exons making it challenging and expensive to sequence using the traditional Sanger sequencing methods. We show that WES sequencing is good alternative to Sanger sequencing to reach a genetic diagnosis in patients with severe male infertility phenotypes.

STUDY FUNDING/COMPETING INTEREST(S): This work was supported by following grants: the 'MAS-Flagella' project financed by the French ANR and the DGOS for the program PRTS 2014 and the 'Whole genome sequencing of patients with Flagellar Growth Defects (FGD)' project financed by the Fondation Maladies Rares for the program Séquençage à haut débit 2012. The authors have no conflict of interest.

Key words: male infertility / genetic diagnosis / gene mutations / exome sequencing / teratozoospermia / flagellum / MMAF / *DNAH1*

Introduction

Men with primary ciliary dyskinesia (PCD), a pathology grouping different phenotypic entities due to molecular defects altering cilia function and affecting mainly the pulmonary function, are often infertile with asthenozoospermia and abnormal flagellar morphology (Storm van's and Omran, 2005). Some patients however present with similar sperm abnormalities with no other associated syndromes, suggesting analogous molecular mechanisms affecting only the sperm flagellum. As this phenotype is restricted to fertility, it could be considered as a new clinical entity and such sperm defects have indeed been described on many occasions since 1984 (Escalier and David, 1984). Patients consistently have had astheno-teratozoospermia characterized by a mosaic of flagellar abnormalities including absent, coiled, bent, angulated, irregular or short flagella mainly due to numerous ultrastructural defects of the axoneme. This phenotype has been reported as dysplasia of the fibrous sheath, short/stump tails or non-specific flagellar anomalies (Chemes et al., 1987; Stalf et al., 1995; Chemes and Rawe, 2003). Until recently, genetic causes of flagellar abnormalities have remained largely unexplained. In 2005, deletions in AKAP3 and AKAP4, two genes encoding proteins of the fibrous sheath, were reported in one patient (Baccetti et al., 2005); these findings however remain to be confirmed. In 2014, our team carried out a large genetics study and we proposed to use the term multiple morphological abnormalities of the sperm flagella (MMAF) which seemed to clearly define the phenotype of the included patients (Ben Khelifa et al., 2014). In this study, homozygosity mapping using a single nucleotide polymorphism (SNP)

array allowed the identification of pathogenic mutations in the dynein axonemal heavy chain 1 (*DNAH1*) gene in 28% of the analyzed patients. The identification of *DNAH1* mutations indicated that *DNAH1* is a major gene involved in the MMAF phenotype and is expected to account for up to one-third of cases (Ben Khelifa et al., 2014). This indicates that MMAF is genetically heterogeneous and that other genes are likely involved in this syndrome (Coutton et al., 2015).

It is currently estimated that 1500–2000 genes are involved in the control of spermatogenesis, including 300–600 specifically expressed in male germ cells (Matzuk and Lamb, 2008). The abundance of potential candidate genes makes the identification of pathogenic mutations difficult and complex. Gene identification is however the key to improving our understanding of the pathophysiology of infertility and could open new perspectives for the diagnosis and treatment of infertile patients. In recent years, different promising genomic approaches have catalyzed the identification of new genes involved in male infertility (El et al., 2012). Whole-exome sequencing (WES), the sequencing of the coding sequence located in the exons of the translated genes, now appears as the best strategy to detect disease-causing variations in individuals affected with Mendelian disorders (Bamshad et al., 2011). Analysis of WES data however remains challenging as 20 000–30 000 variants differing from the genomic reference sequence are usually found in any given individual (Gilissen et al., 2012). It is then extremely arduous to identify the causal variant(s) from this large number of variants of usually unknown significance. The number of potentially pathogenic variants can be reduced by analyzing cohorts of affected individuals and looking for variants or defects in the same gene present in several affected

individuals. Alternatively, consanguineous kindreds can be analyzed and this strategy can be particularly successful for genetically heterogeneous pathology such as infertility (Boycott *et al.*, 2013). Indeed, WES of small families has been successfully used for the detection of causal genes in phenotypes such as non-obstructive azoospermia (Ayhan *et al.*, 2014) or sperm fertilization defects (Escoffier *et al.*, 2015). In this study, we wanted to evaluate the efficiency of family-based WES to identify new *DNAH1* mutations or new genetic causes of MMAF syndrome.

Materials and Methods

Patient and control individuals

Five Iranian families (1–5) and one Italian family (6) were included in this study (Fig. 1). Among these 6 families, we included 12 subjects (P1–P12) presenting with asthenozoospermia due to a combination of 5 morphological defects of the sperm flagella including: absent, short, bent, coiled flagella and of irregular width (Table 1) without any of the additional symptoms associated with PCD. About 10 individuals originated from Middle East (Iranians) and were treated in Tehran at the Royan Institute, (Reproductive Biomedicine Research Center) for primary infertility from

2008 to 2015. Two brothers of European origin (Southern Italia) consulted for primary infertility at the Center for the Study of Rare Inherited Diseases in Milan, Italia. About 10 of the 12 subjects were born from related parents, usually first cousins. WES was performed for a total of 10 patients and *DNAH1* targeted Sanger sequencing was performed for two additional brothers (P7 and P12) who had not been included in time for WES (Fig. 1). Except for Families 5 and 6, at least two infertile brothers from each family were included and were analyzed by WES (Fig. 1). All subjects had normal somatic karyotypes.

Sperm analysis was carried out in the source laboratories during the course of the routine biological examination of the patient, according to World Health Organization (WHO) guidelines (World Health Organization, 2010). The morphology of patients' sperm was assessed with Papanicolaou staining (Fig. 2). Small variations in protocol might occur between the different laboratories. Subjects were recruited on the basis of the identification of >5% of at least four of the aforementioned flagellar morphological abnormalities (absent, short, coiled, bent and irregular flagella) (Table 1). Unfortunately, sperm parameters from patient P2 were not available. All subjects presented with severe asthenozoospermia: seven patients had no motility (<5%), two had sperm motility < 15% and two (P11 and P12) had approximately 30% motility. These latter two patients were considered to have a milder form of MMAF syndrome.

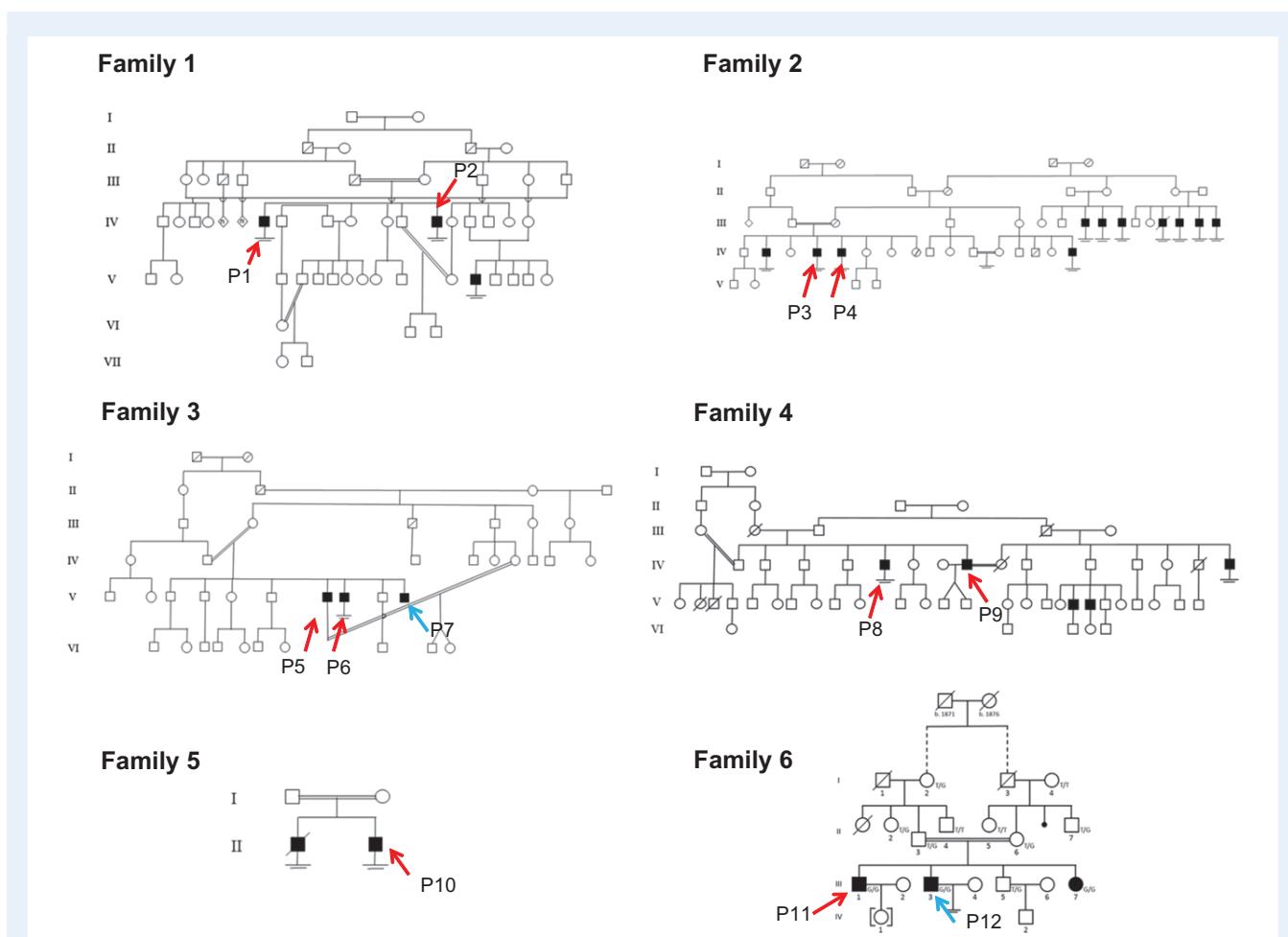


Figure 1 Pedigrees of five Iranian (1–5) and one Italian family (6). Black squares indicate infertile subjects in the family. Red arrows indicate patients for whom WES was performed. Blue arrow indicates patients (P7 and P12) who did not have WES but subsequent exon targeted *DNAH1* Sanger sequencing.

Table I Semen parameters of the 12 patients among whom five subjects (P5, P6, P7, P11 and P12) carried *DNAH1* homozygous variants. Semen parameters from the additional sporadic patient SP4 with a *DNAH1* mutation are also reported. *DNAH1* mutations were identified by WES in patients P5, P6 and P11 and by Sanger sequencing in patients P7, P12 and SP4. Values are the average of two separate analyses.

Semen parameters	Family 1		Family 2		Family 3		Family 4		Family 5		Family 6	
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
DNA/HI mutation	None	None	None	None	c.8626-1G > A	c.8626-1G > A	c.8626-1G > A	c.8626-1G > A	None	None	c.3860-T > G	c.3860-T > G
Sperm volume (ml)	4	NA	3.2	3.5	4	4	4	1	0.9	1	3.5	3
Sperm concentration ($10^6/ml$)	20	NA	6	16	29	18	37	52	50	21	24	31
Motility (A + B) h (%)	0	NA	1.5	14.3	5	0	0	0	0	0	27	33
Vitality (%)	88	NA	NA	NA	97	50	82	92	0	89	74	85
Normal spermatozoa (%)	0	NA	0	1	0	0	0	0	0	3	8	11
Anomalies of the head (%)	1	NA	6	34	12	33	22	2	5	14	24	18
Absent flagella (%)	0	NA	5	1	0	0	0	0	2	15	5	8
ShortFlagella (%)	90	NA	53	6	85	65	80	70	70	35	41	48
Coiled Flagella (%)	+	NA	+	6	8	1	0	5	1	5	6	12

NA = not available. + : anomalies reported ($>5\%$) but not accurately quantified.

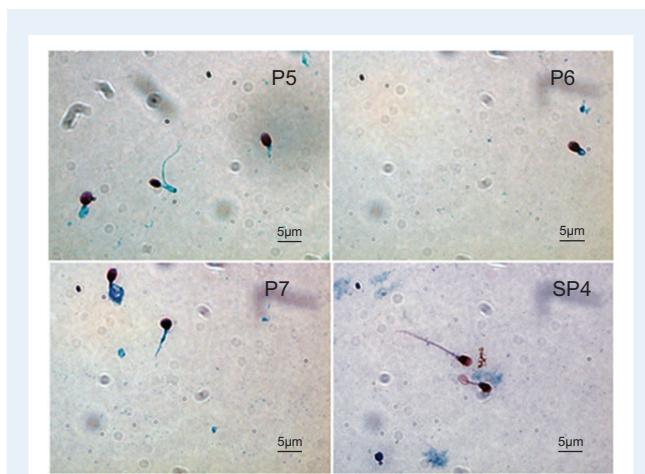


Figure 2 Light microscopy analysis of spermatozoa from three c.8626-1G > A homozygous Iranian brothers (P5, P6, P7) and one sporadic Iranian patient (SP4). Sperm samples were spread over a slide and dried at room temperature for Papanicolaou staining. These images are given as examples of typical spermatozoa observed in mutated patients.

In order to evaluate the incidence of the detected mutation in the general Iranian population, we recruited 38 additional patients (SP1 to SP38) presenting with MMAF phenotype with at less 80% of sperm with flagellar abnormalities ([Supplementary Table S1](#)).

Saliva and/or peripheral blood was obtained for all participants. Sperm samples were obtained, following informed consent, from patients P5, P6, P7 and SP4. All subjects answered a health questionnaire focused on PCD manifestations during their medical consultation for infertility. Informed consent was obtained from all the subjects participating in the study according to the local Institutional Review Board protocols and the principles of the Declaration of Helsinki. In addition, the study was approved by local ethic committees.

Molecular analysis

DNA extraction

DNA was extracted from blood or saliva. Blood DNA extraction was carried out from 5 to 10 ml of frozen EDTA blood using the quick guanidium chloride extraction procedure. Saliva was collected with Oragene DNA Self-Collection Kits (DNA genotek Inc., Canada) and DNA extraction was performed following the manufacturer's recommendations.

Exome sequencing and bioinformatics analysis

WES was carried out on DNA extracted from the 10 studied subjects. Coding regions and intron/exon boundaries were enriched using the 'all Exon V5 kit' (Agilent Technologies, Wokingham, UK). DNA sequencing was undertaken at the Plateforme Biopuces et Séquençage IGBMC, Illkirch, France, on the HiSeq 2000 from Illumina®. All steps from sequence mapping to variant selection were performed using the ExSQLibur pipeline (<https://github.com/tkaraouzene/ExSQLibur>). Sequence reads were aligned to the reference genome (hg19) using MAGIC (SEQC/MAQC-III Consortium, 2014). Duplicate reads and reads that mapped to multiple locations in the exome were excluded from further analysis. Positions whose sequence coverage was below 10 on either the forward or reverse strand were excluded. Single nucleotide variations (SNV) and small insertions/deletions (indels) were identified and quality-filtered using in-house

scripts. The most promising candidate variants were identified using an in-house bioinformatics pipeline. Variants with a minor allele frequency greater than 5% in the NHLBI ESP6500 or in 1000 Genomes Project Phase I data sets, or greater than 1% in ExAC were discarded. We also compared these variants to an in-house database of 56 control exomes. All variants present in a homozygous state in this database were excluded. We used Variant Effect Predictor to predict the impact of the selected variants. We only retained variants impacting splice donor/acceptor sites or causing frame-shift, inframe insertions/deletions, stop gain, stop loss or missense mutations, except those scored as 'tolerated' by SIFT (sift.jcvi.org) and as 'benign' by Polyphen-2 (genetics.bwh.harvard.edu/pph2).

Sanger sequencing

The two variations identified in *DNAH1* using WES were verified by Sanger sequencing. The coding exon 23 and *intron-exon boundary* adjacent to *intron54/exon55* were amplified as indicated in the [Supplementary Table SII](#). Sequencing reactions were carried out with BigDye Terminator v3.1 (Applied Biosystems). Sequence analysis were carried out on ABI 3130XL (Applied Biosystems).

RNA extraction

Nucleated cells were isolated from whole blood using Ficoll® 400 (Sigma-Aldrich Corp., St. Louis, MO, USA) following the manufacturer's protocol. RNA extraction was carried out on the isolated white blood cells using Macherey Nagel NucleoSpin® RNA II columns (Macherey Nagel, Hoerdt, France) using the manufacturer's protocol.

RT-PCR

Reverse transcription was carried out in three patients P5, P6, P7 and two healthy controls (C1 and C2) with 5 µl of extracted RNA (approximately 500 ng). Hybridization of the oligo-dT was performed by incubating for 5 min at 65°C and quenching on ice with the following mix: 5 µl of RNA, 3 µl of poly T oligo primers (dT)12–18 (10 mM, Pharmacia), 3 µl of the four dNTPs (0.5 mM, Roche diagnostics) and 2.2 µl of H₂O. Reverse Transcription was then carried out for 30 min at 55°C after the addition of 4 µl of 5× buffer, 0.5 µl RNase inhibitor and 0.5 µl of Transcriptor Reverse transcriptase (Roche Diagnostics). Then 2 µl of the obtained cDNA mix was used for the subsequent PCR. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as a housekeeping gene (internal control). Primers sequences and RT-PCR conditions are indicated in the [Supplementary Table SIII](#).

Immunostaining

Sperm were fixed in PBS/4% paraformaldehyde for 1 min at room temperature. After washing in 1 ml PBS, the sperm suspension was spotted onto 0.1% poly L-lysine pre-coated slides (Thermo Scientific). After attachment, sperm were permeabilized with 0.1% (v/v) Triton X-100—DPBS (Triton X-100; Sigma-Aldrich) for 5 min at room temperature. Slides were then blocked in 5% corresponding normal serum—DPBS (normal goat or donkey serum; GIBCO, Invitrogen) and incubated overnight at 4°C with primary antibodies. Polyclonal DNAH1 antibodies were purchased from Prestige Antibodies® (Sigma-Aldrich, France) (1:50). Monoclonal mouse anti-acetylated-α-tubulin antibodies were purchased from Sigma-Aldrich (1:2500). Washes were performed with PBS containing 0.1% of Tween 20, followed by 1 h incubation at room temperature with Alexa Fluor 555-labeled goat anti-rabbit or Dylight 488-labeled goat anti-rabbit (1:400) secondary antibodies. Appropriate controls were performed, omitting the primary antibodies. Samples were counterstained with 5 mg/ml Hoechst 33342 (Sigma-Aldrich) and mounted with DAKO mounting media (Life Technology). Whole images were reconstructed and projected from Z-stack images using ZEN software.

Results

WES identified two new *DNAH1* mutations

Given the notion of consanguinity in the families we studied, we postulated that the infertility has been likely transmitted by autosomal recessive inheritance and was thus caused by a homozygous mutation. We proceeded to WES to identify a possible genetic defect that could explain the observed MMAF phenotype. After exclusion of frequent variant and applying stringent filters, a limited list of homozygous variants was identified in each proband. Apart from these three variants in *DNAH1*, no variants were present in genes described to be strongly expressed in the testis nor in any gene described to be connected with cilia, the axoneme or the flagellum. Only the *DNAH1* variants were retained as likely causal ([Supplementary Table SIV](#)). In cases where the brothers were simultaneously analyzed by WES, only common variants shared by both brothers were retained. Two different pathogenic mutations in the *DNAH1* gene were identified in Families 3 and 6.

In Family 3, the c.8626-1G > A variant was identified in the two MMAF brothers analyzed by WES. This mutation affects the final G nucleotide of *DNAH1* intron 54, one of the consensus splice acceptor bases (Fig. 3). The alternate splicing is predicted to cause a frameshift in the new transcript and to induce a premature stop codon. The c.8626-1G > A variant was absent from over 60 000 individuals described in the ExAC database which confirms that it is not a polymorphism and that a splicing mutation occurring at this localization would be negatively selected throughout evolution. No other variants were identified in the *DNAH1* coding sequence and UTR regions. Sanger sequencing confirmed the splicing mutation for both infertile brothers (Fig. 3) and showed that the third infertile brother has the same mutation while their father and a non-affected brother were heterozygous (data not shown).

In Family 6, a second mutation was found in *DNAH1*. The identified mutation is c.3860 T > G (p.Val1287Gly). This was confirmed in the same subject by bidirectional Sanger sequencing of *DNAH1* exon 23. No other unreported *DNAH1* variant was identified by exome sequencing. No rare variants were present in other genes reported to be associated with male infertility. The presence of the c.3860 T > G mutation was tested in the other family members available for genotyping by restriction analysis and confirmed by Sanger sequencing. In this large Italian family, the other infertile brother and one sister were homozygous for the *DNAH1* mutation (Fig. 3). Six other related fertile subjects of both sexes were heterozygous for the mutation. The sister, also homozygous for p.Val1287Gly mutation, was healthy by general evaluation and had not yet attempted to have children.

Targeted PCR-Sanger sequencing in 38 MMAF Iranian patients

We next assessed 38 Iranian patients (SP1 to SP38) presenting with morphological abnormality of flagella ([Supplementary Table S1](#)) for the c.8626-1G > A splicing mutation identified in the Iranian Family 3. Sequencing results showed that one of these affected men (SP4) is homozygous for this new splice site mutation indicating that this mutation segregates in the Iranian population.

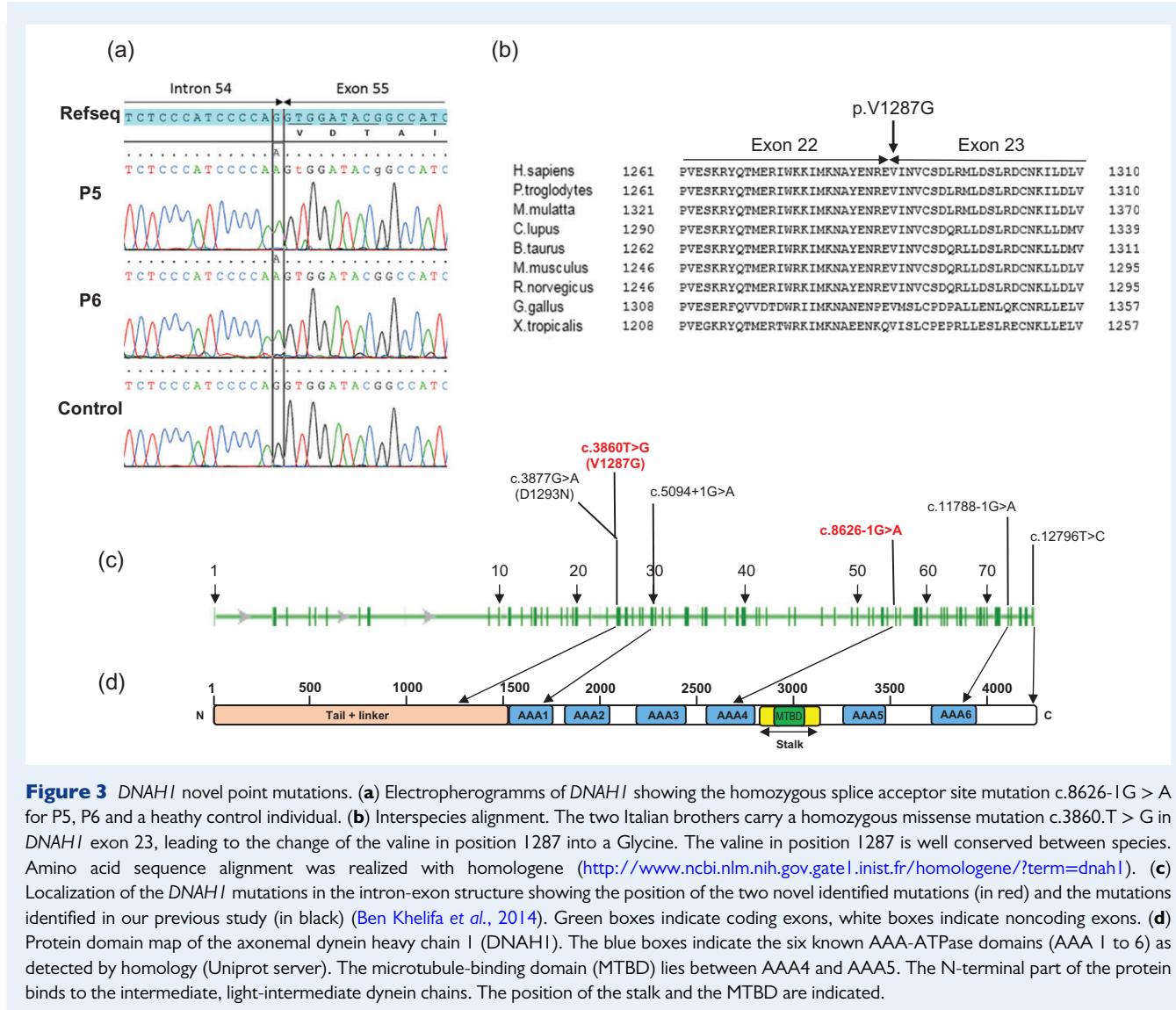


Figure 3 DNAH1 novel point mutations. **(a)** Electropherograms of DNAH1 showing the homozygous splice acceptor site mutation c.8626-1G > A for P5, P6 and a healthy control individual. **(b)** Interspecies alignment. The two Italian brothers carry a homozygous missense mutation c.3860.T > G in DNAH1 exon 23, leading to the change of the valine in position 1287 into a Glycine. The valine in position 1287 is well conserved between species. Amino acid sequence alignment was realized with homologene (<http://www.ncbi.nlm.nih.gov/gatel.inist.fr/homologene/?term=dnah1>). **(c)** Localization of the DNAH1 mutations in the intron-exon structure showing the position of the two novel identified mutations (in red) and the mutations identified in our previous study (in black) (Ben Khelifa et al., 2014). Green boxes indicate coding exons, white boxes indicate noncoding exons. **(d)** Protein domain map of the axonemal dynein heavy chain I (DNAH1). The blue boxes indicate the six known AAA-ATPase domains (AAA 1 to 6) as detected by homology (Uniprot server). The microtubule-binding domain (MTBD) lies between AAA4 and AAA5. The N-terminal part of the protein binds to the intermediate, light-intermediate dynein chains. The position of the stalk and the MTBD are indicated.

Detrimental effects of the two identified mutations

To assess the functional impact of the DNAH1 splice acceptor site mutation c.8626-1G > A, we studied mRNA products isolated from two control and from P5, P6 and P7. RT-PCR of patients' samples yielded no product despite repeated attempts, contrary to what was observed from the controls yielding bands of the expected size (Fig. 4). RT-PCR targeting GAPDH confirmed the integrity of patients' RNA (Fig. 4). This suggests a specific degradation of the mutant DNAH1 transcripts by nonsense mediated mRNA decay (NMD). To further validate the pathogenicity of this variant, we analyzed DNAH1 localization by immunofluorescence on patients' sperm. In control individuals, DNAH1 antisera decorated the full length of the sperm flagellum (Fig. 5). In contrast, in sperm from the three brothers carrying the c.8626-1G > A mutation as well as in sperm from the sporadic case SP4, DNAH1 immunostaining was absent, indicating that the splicing defect induces the

degradation of the transcripts by NMD thus precluding protein production (Fig. 5).

Unfortunately, no mRNA analysis or immunostaining could be performed on sperm cells from the Italian patients, P11 and P12. We however found Val1287 to be very well conserved throughout evolution (Fig. 3) and this missense change is also predicted to be likely damaging by SIFT and PolyPhen-2, two prediction software for nonsynonymous SNPs. This variant was also absent from all the control sequence databases (dbSNP v137, 1000 Genomes Project, NHLBI Exome Variant Server).

Discussion

To date, SNP array-based homozygosity mapping has permitted the identification of mutations in three main genes leading to teratozoospermia (Dieterich et al., 2007; Harbuz et al., 2011; Ben Khelifa et al., 2014). This strategy is however time-consuming and requires that several patients

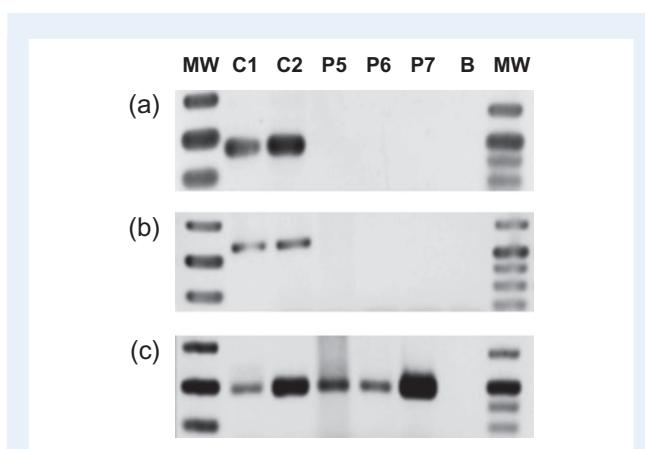


Figure 4 RT-PCR analyses of subjects P5, P6 and P7 (c.8626-1G > A homozygotes) and control individuals from the general population. Electrophoresis showing the RT-PCR amplification of (a) *DNAH1* exons 53–56 (b) *DNAH1* exons 53–57 and (c) GAPDH amplification used as housekeeping gene. Controls DNA C1 and C2 show good amplification for all tested loci whereas the three patients do not amplify *DNAH1* but only the GAPDH control. There is no amplification from the RT-negative blank control (column B).

share a common genetic cause responsible for the phenotype. Due to the high genetic heterogeneity of infertility, SNPs array has now reached its limits and has been supplanted by next generation sequencing and in particular WES. Unfortunately in this study, this strategy has not led to the identification of new reliable candidate genes in MMAF patients. Some deleterious variants have been identified but they do not concern genes that have been described to have a strong connection with spermatogenesis. What appears as a negative result can be explained by: (i) a lack of information regarding a mutated gene explaining why it was not selected; (ii) a false starting assumption based on the hypothesis that the infertility has been likely transmitted by autosomal recessive inheritance; (iii) exclusion criteria being too stringent, eliminating the pathogenic variant (e.g. silent variant modifying splicing sites), (iv) a variant undetectable by the technique used or our current bio-informatic pipeline, e.g., genomic rearrangements (large deletions and duplications), unsequenced deep intronic variants or some exonic variants (as only about 90% of coding nucleotides were covered). These results therefore highlight the fact that WES cannot be expected to provide 100% positive diagnoses. A diagnosis was however reached for 2 out of 6 of the analyzed families (33%). Two new *DNAH1* mutations were identified. These results thus reinforce the fact that *DNAH1* remains the main gene associated with MMAF and corroborate our previous study indicating that approximately 30% of subjects with MMAF are expected to harbor *DNAH1* mutations. These findings raise

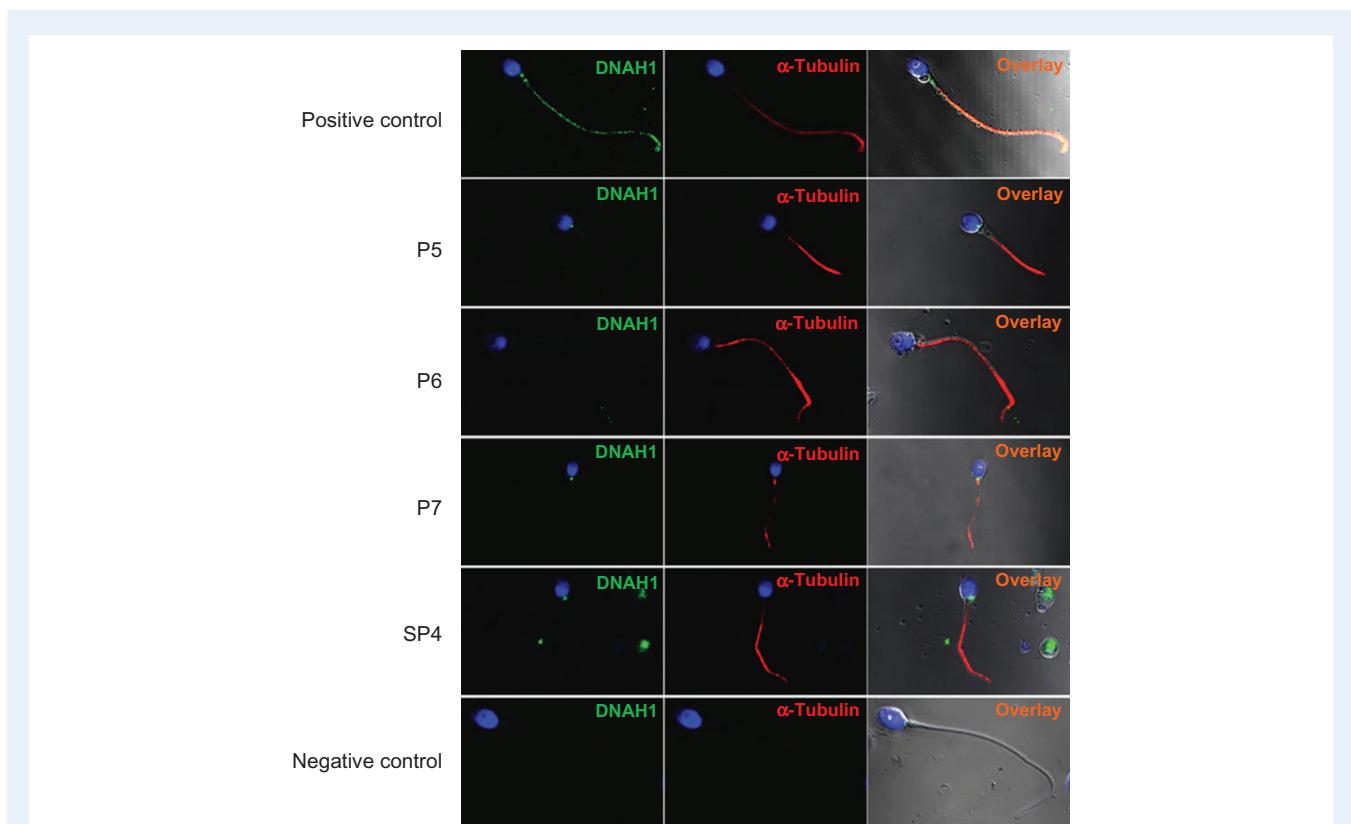


Figure 5 Immunofluorescence images of human spermatozoa from the four patients with the c.8626-1G > A mutation with *DNAH1* antibodies and α -Tubulin. *DNAH1* staining (green) was observed throughout the flagellum in positive control sperm, whereas it was absent in all sperm samples from patients with the c.8626-1G > A variant. The α -tubulin signal (red) was observed in controls and patients. Sperm were counterstained with Hoechst 33342 (blue) as nuclei marker. These images are given as examples of typical stainings observed in patients.

the question of whether *DNAH1* routine diagnosis should be proposed to all MMAF patients. *DNAH1* is a very large gene (84 Kb and 78 exons) making conventional Sanger sequencing difficult, laborious and costly. In view of our results, we recommend to perform WES in MMAF patients as a first approach. WES in MMAF patients provides a fast way to sequence all *DNAH1* exons while giving the opportunity to identify new gene defects. This strategy, followed by the assessment of ICSI results obtained with patients with different gene defects, could allow improvements in the prediction of ICSI success rates and thus to provide better counseling to MMAF patients. We recently reported that ICSI with spermatozoa from MMAF patients with *DNAH1* mutations had a high pregnancy rate following ICSI (Wambergue et al., 2016). *DNAH1* mutation positive patients identified in this study can thus be encouraged to initiate an IVF/ICSI procedure.

One of the main difficulties associated with WES is the confirmation of the deleterious effect of the identified variant. The effect of the first *DNAH1* variant (*c.8626-1G > A*) cannot be questioned as: (i) it was identified in four individuals (P5, P6, P7 and SP4); (ii) it affects a consensus splice site known to be essential to the mRNA splicing machinery; and (iii) the mRNA and the protein were shown to be absent from mutated patients. The effect of the second variant, Val1287Gly, is not as easy to predict as it concerns only one amino acid. We however have several arguments in favor of its pathogenicity: (i) the variant was described as deleterious by two prediction tools; (ii) it was not found in any database now including in excess of 60 000 individuals; and (iii) it affects a conserved residue located in the N-terminal part of the protein known to be important for the structure of the dynein arms (Habura et al., 1999) and it is positioned close to the first missense *DNAH1* mutation (Asp1293Asn) identified previously. The four patients with the severe variant present 0% of morphologically normal spermatozoa with a motility <5% in contrast with the two brothers with the Val1287Gly variant who present a milder phenotype with approximately 30% motility and 10% of morphologically normal spermatozoa (Table 1). This suggests that the Val1287Gly mutated protein is likely preserve a residual activity as was previously observed in the patient with the nearby missense mutation. This supports the hypothesis of a phenotype continuum depending on the severity of *DNAH1* mutations (Ben Khelifa et al., 2014). We can therefore expect that individuals harboring homozygous or compound heterozygous *DNAH1* mutations of moderate severity could present with intermediate asthenozoospermia and low levels of morphological anomalies.

When all patients with MMAF analyzed in our previous and present work are pooled, we have a total of 32 patients including 24 index cases. Seven unrelated individuals out of 24 carried a homozygous variant in *DNAH1* (29.2%). Although a large majority of patients are of North African origin, we have now identified *DNAH1* mutations in Middle East and European patients indicating that *DNAH1* diagnosis should not be ruled out for non-Maghrebian individuals. The *c.8626-1G > A* variant was found in 1/38 sporadic MMAF Iranian cases resulting in a prevalence estimated at 2.6% in this Iranian population. It would be interesting to compare the haplotype of these affected Iranian patients in order to highlight a possible founder effect.

As reported in our previous work, all individuals carrying mutations in *DNAH1* only presented with male infertility and did not report any other symptoms associated with PCD such as an impairment of the respiratory functions (Ben Khelifa et al., 2014). The data reported here is consistent with the hypothesis suggesting that *DNAH1* function in cilia may be compensated by other dyneins. The absence of clinical signs does

not formally exclude that *DNAH1* mutations could lead to a slight functional impairment of motile cilia function in respiratory epithelium but the impairment would not necessarily lead to lung dysfunction. Unfortunately, family members in which a *DNAH1* mutation was identified did not consent to the investigation of their respiratory function.

The abundance of potential candidate genes makes identification of pathogenic mutations difficult and complex. However, gene identification is the key to improving knowledge of the pathophysiology of MMAF and opens new perspectives for diagnosis and treatment of infertile patients. Further genetic studies are therefore warranted to identify other genes involved in MMAF to better characterize the genetic etiology of the MMAF phenotype and to improve the management of patients diagnosed with flagellar defects.

Supplementary data

Supplementary data are available at <http://humrep.oxfordjournals.org/>.

Acknowledgments

The authors thank the patients for their interest and cooperation.

Authors' roles

P.F.R., C.C., A.A.Y. and H.G. analyzed the data, wrote the manuscript and had full access to all of the data and take responsibility for the integrity of the data and its accuracy. A.A.Y., Z.E.K., P.L.T. performed the experimental analyses. T.K. performed data mining for the genetic data. M.H.S., M.S., N.A., M.A.S.G., S.H.H., S.B., A.D., C.A., R.C., M.B. and H.G. provided clinical samples and data and supplied biological materials. All authors contributed to the report.

Funding

This work was supported by following grants: the 'MAS-Flagella' project financed by the French ANR and the DGOS for the program PRTS 2014 and the 'Whole genome sequencing of patients with Flagellar Growth Defects (FGD)' project financed by the Fondation Maladies Rares for the program Séquençage à haut débit 2012.

Conflict of interest

None declared.

References

- Ayhan O, Balkan M, Guven A, Hazan R, Atar M, Tok A, Tolun A. Truncating mutations in TAF4B and ZMYND15 causing recessive azoospermia. *J Med Genet* 2014;51:239–244.
- Baccetti B, Collodel G, Estenoz M, Manca D, Moretti E, Piomboni P. Gene deletions in an infertile man with sperm fibrous sheath dysplasia. *Hum Reprod* 2005;20:2790–2794.
- Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 2011;12:745–755.
- Ben Khelifa M, Coutton C, Zouari R, Karaouzene T, Rendu J, Bidart M, Yassine S, Pierre V, Delaroche J, Hennebicq S et al. Mutations in *DNAH1*, which encodes an inner arm heavy chain dynein, lead to male

- infertility from multiple morphological abnormalities of the sperm flagella. *Am J Hum Genet* 2014; **94**:95–104.
- Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* 2013; **14**:681–691.
- Chemes EH, Rawe YV. Sperm pathology: a step beyond descriptive morphology. Origin, characterization and fertility potential of abnormal sperm phenotypes in infertile men. *Hum Reprod Update* 2003; **9**:405–428.
- Chemes HE, Brugo S, Zanchetti F, Carrere C, Lavieri JC. Dysplasia of the fibrous sheath: an ultrastructural defect of human spermatozoa associated with sperm immotility and primary sterility. *Fertil Steril* 1987; **48**:664–669.
- Coutton C, Escoffier J, Martinez G, Arnoult C, Ray PF. Teratozoospermia: spotlight on the main genetic actors in the human. *Hum Reprod Update* 2015; **21**:455–485.
- Dieterich K, Soto RR, Faure AK, Hennebicq S, Ben Amar B, Zahi M, Perrin J, Martinez D, Sele B, Jouk PS et al. Homozygous mutation of AURKC yields large-headed polyploid spermatozoa and causes male infertility. *Nat Genet* 2007; **39**:661–665.
- El IE, Muller J, Viville S. Autosomal mutations and human spermatogenic failure. *Biochim Biophys Acta* 2012; **1822**:1873–1879.
- Escalier D, David G. Pathology of the cytoskeleton of the human sperm flagellum: axonemal and peri-axonemal anomalies. *Biol Cell* 1984; **50**:37–52.
- Escoffier J, Lee HC, Yassine S, Zouari R, Martinez G, Karaouzene T, Coutton C, Kherraf ZE, Halouani L, Triki C et al. Homozygous mutation of PLCZ1 leads to defective human oocyte activation and infertility that is not rescued by the WW-binding protein PAWP. *Hum Mol Genet* 2016; **25**:878–891.
- Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 2012; **20**:490–497.
- Habura A, Tikhonenko I, Chisholm RL, Koonce MP. Interaction mapping of a dynein heavy chain. Identification of dimerization and intermediate-chain binding domains. *J Biol Chem* 1999; **274**:15447–15453.
- Harbuz R, Zouari R, Pierre V, Ben Khelifa M, Kharouf M, Coutton C, Merdassi G, Abada F, Escoffier J, Nikas Y et al. A recurrent deletion of DPY19L2 causes infertility in man by blocking sperm head elongation and acrosome formation. *Am J Hum Genet* 2011; **88**:351–361.
- Matzuk MM, Lamb DJ. The biology of infertility: research advances and clinical challenges. *Nat Med* 2008; **14**:1197–1213.
- Stalf T, Sanchez R, Kohn FM, Schalles U, Kleinstein J, Hinz V, Tielsch J, Khanaga O, Turley H, Gips H et al. Pregnancy and birth after intracytoplasmic sperm injection with spermatozoa from a patient with tail stump syndrome. *Hum Reprod* 1995; **10**:2112–2114.
- Storm van's GK, Omran H. Primary ciliary dyskinesia: clinical presentation, diagnosis and genetics. *Ann Med* 2005; **37**:439–449.
- Wambergue C, Zouari R, Fourati Ben MS, Martinez G, Devillard F, Hennebicq S, Satre V, Brouillet S, Halouani L, Marrakchi O et al. Patients with multiple morphological abnormalities of the sperm flagella due to DNAH1 mutations have a good prognosis following intracytoplasmic sperm injection. *Hum Reprod* 2016; **31**:1164–1172.
- World Health Organization. *WHO Laboratory Manual for the Examination and Processing of Human Semen*, Geneva, 5th edn. World Health Organization, 2010.

Supplementary Table S1 Semen parameters of 38 Iranian sporadic patients.

Patients	Semen volume (ml)	Sperm conc. ($10^6/\text{ml}$)	Total motility 1 h (%)	Vitality (%)	Normal spermatozoa (%)	Anomalies of the flagella (%)			Anomalies of the head (%)			Amorph (%)
						Short	Coiled	Other	Round	Pin	Giant	
SP1	4.50	0.15	0	0	98							2
SP2	5	7	0	60	0	86	2	2		1		9
SP3	0.3	6	0	60	1	92				2		5
SP4*	4.5	13	Few spz	88	0	86	1					13
SP5	2.5	30	0	70	0	98						2
SP6	4	11	Few spz	88	0	89			3	1	2	5
SP7	3	12	0	88	0	92		2				8
SP8	1	5	0	55	1	81				2	6	10
SP9	2.4	6	0	35	0	95				3	2	
SP10	2.1	3	0	82	0	84	1		4			11
SP11	5	4	Few spz	13	0	79	1			1		19
SP12	3.2	0.2	Few spz	0	86	5	2					5
SP13	3	20	0	92	0	96				2		2
SP14	2	20	0	80	0	100						
SP15	2.5	14	Few spz	88	0	86	4	2	2	2		4
SP16	3	1	0	85	0	88	2		4		2	4
SP17	5.5	7	0	88	0	92			3			5
SP18	4	3	0	88	0	90				5		5
SP19	3	16	0	85	0	90	4		2	1		3
SP20	6.5	6	0	80	0	90		3				7
SP21	3.2	5	0	95	0	100						
SP22	4.6	0.5	0	80	0	89	1		7			3
SP23	3.4	25	0	80	0	84	1	8				7
SP24	6.5	4	Few spz	70	0	92						8
SP25	4.2	2	0	35	2	83		2		2		11
SP26	4	5	Few spz	84	1	78		4		2		15
SP27	4.6	25	0	80	0	93	3		2			2
SP28	3	34	0	56	0	71	8	1	1	3		16
SP29	3.4	2.5	Few spz	48	0	79	3	2		1	1	14
SP30	3.5	30	5	75	1	60			5	2		32
SP31	2	5	0	3	0	88	1		1			10
SP32	3	0.25	0	0	86			8	4			2
SP33	5	0.5	Few spz	1		72		2		2		23
SP34	1	8	0	95	0	95						5
SP35	3	1	0	0	0	80	2		7	1		10
SP36	2.5	7	0	97	0	84	9		2	2		5
SP37	4.5	12	0	84	0	98						2
SP38	2.5	32	0	92	0	83			4	3		10

*Mutated patient.

Supplementary Table SII Primer sequences used for Sanger sequencing of *DNAH1* exons.

Primer name	Primer sequence (5'-3')	Tm
DNAH1-int54F	CACCCCAACTCTCCTTCCAT	58°C
DNAH1-int54R	TCTGGGCATCGTCAGCAATA	
DNAH1-ex23F	TGGGATGAGCCTATCTTGCT	60°C
DNAH1-ex23R	AGCCTTGTGGGCAGACAGT	

Supplementary Table SIII Primer sequences used in RT-PCR and respective melting temperatures (Tm).

Primer name	Primer sequence (5'-3')	Size of amplicons	Tm
RTex55-DNAH1F	GCTTCATATTTCTCCATCC	55 F/55R1: 584 bp	54°C
RTex55-DNAH1R1	CAATGTTGCCTGTCAAAC	55 F/55R2: 441 bp	
RTex55-DNAH1R2	ATGCACACAGCTTCTATGAC		

Supplementary Table SIV All *DNAH1* variations identified by WES for all probands analyzed from six MMAF families.

Gene	Variant coordinates	Transcript	cDNA variation	Amino acid variation	Prediction	Nationality
<i>DNAH1</i>	Chr3:52420175:G:A	ENST00000420323	c.8626-1G > A	Splice acceptor	Damaging	Iranian
<i>DNAH1</i>	Chr3:52391630:T:G	ENST00000420323	c.3860.T > G	p.Val1287Gly	Possibly_damaging	Italian

Principaux résultats

Après avoir été séquencés en WES, les données des 9 patients *Ghs119*, *Ghs130*, *Ghs131*, *Ghs56*, *Ghs58*, *Ghs59*, *Ghs60*, *Ghs62* et *Ghs63* ont été analysées au sein de notre pipeline. Compte tenu de l'historique de consanguinité de ces familles, l'ensemble des variants hétérozygotes ont été filtrés de même que l'ensemble des variants observés fréquemment dans la population générale et ceux n'ayant aucun impact sur la séquence codante. Aussi, dans les cas où les données de WES de plusieurs frères d'une même famille étaient disponibles, seuls les variants partagés par l'ensemble des frères ont été gardés. À l'issue de cette étape de filtre, seuls quelques variants subsistaient pour chacune des familles impactant entre 1 et 23 gènes différents en fonction de celles-ci (**Tables** : B.1, B.2, B.3, B.4, B.5, **Figure** : 2.16).

Parmi cette liste de gènes, *DNAH1* fut retrouvé muté chez les deux frères de la famille MMAF3 ainsi que chez les deux frères de la famille MMAF6 (non analysée par notre pipeline). De même, un variant entraînant un décalage du cadre de lecture dans la séquence du gène *SPEF2* (codant pour la protéine *Sperm flagellar 2* (SPEF2)) a été retrouvé chez le patient P10 de la famille MMAF5. Aucun autre candidat évident n'a pu être identifié pour les individus composant les 3 autres familles. Ensuite, bien que le gène *SPEF2* ait déjà été caractérisé comme ayant un rôle dans la biogénèse du flagelle spermatique [205] nous nous sommes dans un premiers temps concentrés sur la caractérisation des deux variants retrouvés sur *DNAH1*.

1. **Famille MMAF3 :** Les deux frères P5 et P6 analysés en WES étaient tous deux porteurs de la même mutation c.8626-1G > A qui fut par la suite confirmée en Sanger pour ces deux patients ainsi que pour leur troisième frère (P7) non analysé en WES. Cette mutation, absente des différentes bases de données, impacte le dernier nucléotide de l'intron 54 de *DNAH1*, c'est-à-dire, l'une des deux bases composant le site accepteur consensus d'épissage. Afin d'évaluer l'impact de cette mutation sur le transcrit de *DNAH1*, nous avons étudiés, par RT-PCR, l'ARNm provenant de ces trois frères ainsi que de deux individus contrôle. Cette étude a révélé qu'aucune amplification du transcrit de *DNAH1* n'était observée chez les trois frères contrairement aux deux contrôles tandis que l'amplification ciblant *GAPDH* était positif pour les cinq individus confirmant ainsi l'intégrité de l'ARNm de l'ensemble des individus testés. Ces résultats suggèrent donc que les transcrits produits par les trois frères mutés ont été soumis au mécanisme de dégradation spécifique par *mRNA decay*. Afin de valider la pathogénicité de ce variant, la protéine DNAH a ensuite été localisée par immunofluorescence à la fois chez les patients et les contrôles, révélant que contrairement aux contrôles, la protéine DNAH1 était absente chez les trois frères, renforçant ainsi l'hypothèse d'une dégradation spécifique des ARNm.
2. **Famille MMAF6 :** Le variant c.3860 T > G (p.Val1287Gly) induisant une mutation faux sens dans la séquence de l'exon 23 de *DNAH1* a été retrouvé dans les données WES des deux frères P11 et P12 (non analysés par notre pipeline) et a par la suite été confirmé en Sanger. Malheureusement, par manque de matériel,

aucune étude par RT-PCR ou immunofluorescence n'a pu être effectuée sur ces patients. Cependant, le fait que ce variant faux-sens soit absent des bases de données et qu'il soit prédict comme *probably damaging* par PolyPhen et *deleterious* par SIFT renforce l'hypothèse de la pathogénicité de ce variant.

Ainsi, à l'issue de l'analyse de ces 6 familles présentant un phénotype MMAF (dont seulement 5 ont été analysées sur notre pipeline), le variant responsable a pu être déterminé pour deux d'entre elles grâce à l'identification de deux nouveaux variants impactant la séquence codante du gène *DNAH1* confirmant ainsi l'importance de l'implication de ce gène dans le phénotype MMAF. De même un indel entraînant un décalage du cadre de lecture dans la séquence du gène *SPEF2* fait de ce gène, déjà connu comme ayant un rôle dans la formation des flagelles, un excellent candidat pour expliquer le phénotype du patient P10. Notre équipe travaille à l'heure actuelle à la caractérisation de ce gène. Cependant, aucun candidat évident n'a pu être identifié pour les 3 familles restantes laissant supposer que le variant responsable de leur phénotype ait été éliminé par un de nos filtres. Ainsi, la cause génétique de leur phénotype n'a soit pas été détectée soit a été éliminée par un de nos filtres. Afin d'identifier la cause génétique de leur phénotype, nous réanalysons actuellement les données de ces patients en appliquant des filtres moins stringents, en conservant les gènes sur lesquels deux variants hétérozygotes sont retrouvés, ce qui pourrait être la signature d'une hétérozygotie composite.

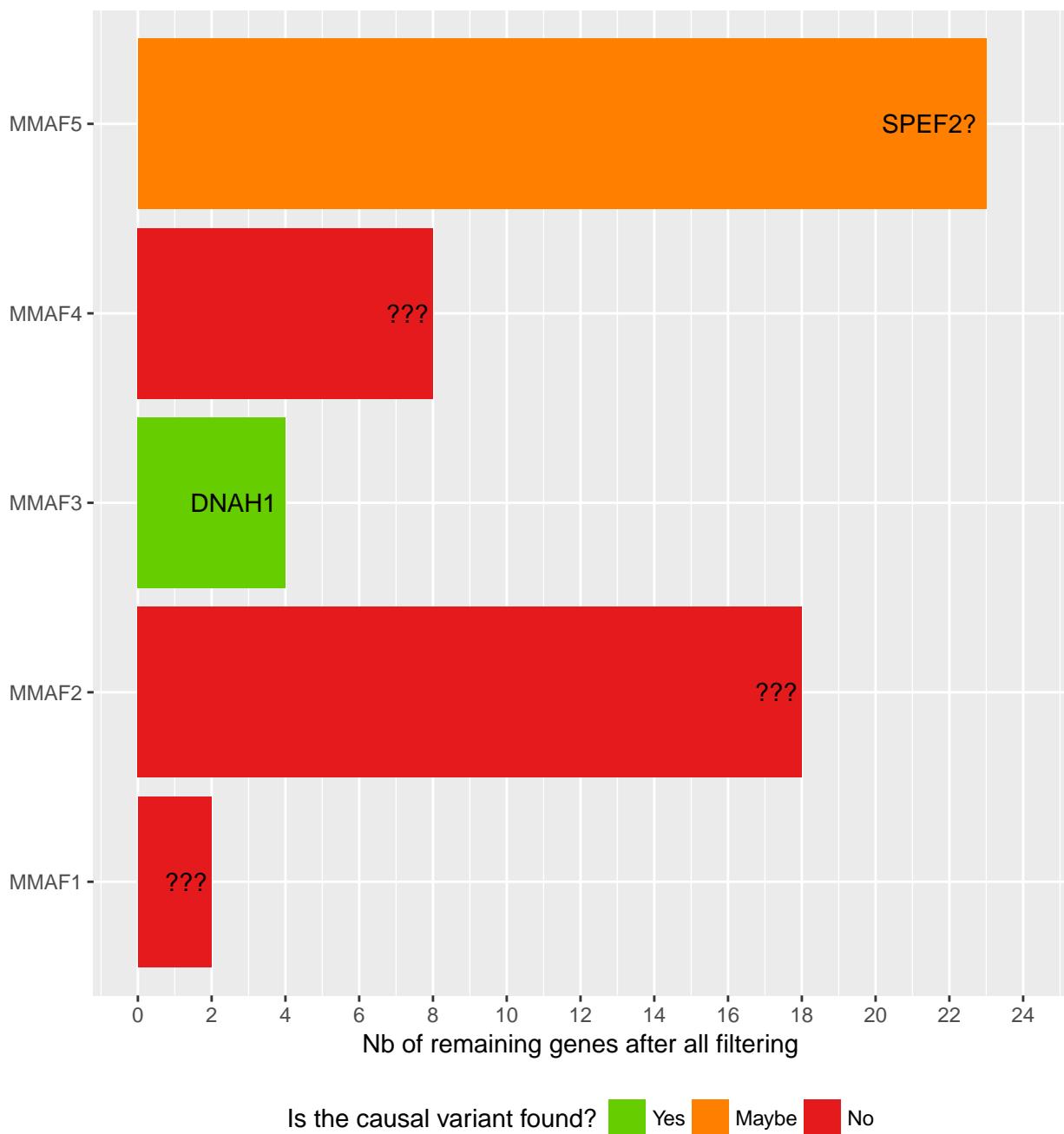


Figure 2.16 – Nombre de gènes passant l'ensemble des filtres par famille : Chaque barre représente une des familles analysées. La hauteur de cette barre correspond au nombre de gènes ayant passé l'ensemble des filtres pour chaque famille. Les barres vertes caractérisent les familles pour lesquelles le gène responsable de la pathologie a été identifié parmi la liste de gènes (dans ce cas le symbole du gène est écrit au-dessus de la barre). La barre orange caractérise la famille pour laquelle un candidat potentiel a été identifié (le symbole du gène est écrit au-dessus suivi d'un “?”). Les barres rouges indiquent qu'aucun des gènes ayant passé les filtres ne semble expliquer le phénotype des individus de la famille (dans ce cas il est écrit “???” au-dessus de la barre).

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

2.3.1 Article n°4

PATL2 Gene Mutation Causes Oocyte Meiotic Deficiency and Female Infertility

Christou-Kent M, Amiri-Yekta A, Kherraf ZE, Karaouzène T, Escoffier J, Guttin A, Martinez G, Le Blévec E, Lambert E, Fourati Ben Mustapha S, Cedrin-Durnerin I, Halouani L, Marrakchi O, Makni M, Latrous H, Kharouf M, Bottari S, Thierry-Mieg N, Coutton C, Zouari R, Issartel JP, Ray PF, Arnoult C

New England Journal of Medicine, 07 Juillet 2017 (soummis)

Contexte et objectifs

Entre 2013 et 2014 notre équipe a pris en charge l'étude génétique de 23 femmes nord africaines présentant toutes une déficience méiotique ovocytaire (DMO) caractérisée par un blocage de la méiose au stade M1 et entraînant une infertilité. À l'heure actuelle, seul le gène TUBB8 retrouvé muté à l'état hétérozygote chez des patientes chinoises avait pu être lié à ce phénotype. Cette étude a donc pour objectif de caractériser la cause génétique responsable du phénotype DMO de ces 23 femmes. Parmi celles-ci, 15 ont été analysées par séquençage haut-débit. Dans ce contexte, j'ai, au cours de ma thèse, été en charge de l'ensemble des analyses bio-informatiques de ces 15 femmes.



Please review the Supplemental Files folder to review documents not compiled in the PDF.

PATL2 Gene Mutation Causes Oocyte Meiotic Deficiency and Female Infertility

Journal:	<i>New England Journal of Medicine</i>
Manuscript ID	17-09001
Article Type:	Original Article
Date Submitted by the Author:	07-Jul-2017
Complete List of Authors:	Christou-Kent, Marie; Institute for advanced Biosciences, team GETI Amiri-Yekta, Amir; Institute for advanced Biosciences, team GETI Kherraf, Zine-Eddine; Institute for advanced Biosciences, team GETI karaouzène, Thomas; Institute for advanced Biosciences, team GETI Escoffier, Jessica; Institute for advanced Biosciences, team GETI Guttin, Audrey; Grenoble Neuroscience Institute Martinez, Guillaume; Institute for advanced Biosciences, team GETI Le Blévec, Emilie; Institute for advanced Biosciences, team GETI Lambert, Emeline; Institute for advanced Biosciences, team GETI Fourati Ben Mustapha, Selima; Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation cedrin-durnerin, Isabelle; Assistance Publique - Hopitaux de Paris Halouani, Lazhar; Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation Marrakchi, Ouafi; Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation Makni, Mounir; Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation Latrous, Habib; Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation Kharouf, Mahmoud; Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation Bottari, Serge; Institute for advanced Biosciences, team GETI Thierry-Mieg, Nicolas; Laboratoire Techniques de l'Ingenierie Medicale et de la Complexite Informatique Mathematiques et Applications Grenoble Coutton, Charles; Institute for advanced Biosciences, team GETI Zouari, Raoudha; Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation Issartel, Jean-Paul; Grenoble Neuroscience Institute Ray, Pierre; CHU de GRENOBLE, ; Université Joseph Fourier, ARNOULT, christophe; Institute for advanced Biosciences, team GETI
Abstract:	Background— Infertility impacts the life of over 70 million couples worldwide, yet its molecular basis remains largely unknown. Some women present primary infertility with no menstruating defects. They respond well to the IVF ovarian stimulation protocol with normal follicular growth however the collected oocytes all appear immature or degenerated and

1
2
3
4 cannot be fertilized even after recourse to assisted reproductive
5 technologies. The aim of this study was to identify the genetic defects
6 inducing this oocyte meiotic deficiency (OMD) and to characterize its
7 molecular pathogeny.

8 Methods and subjects
9

10 A total of 23 OMD subjects were recruited and whole exome and Sanger
11 sequencing was performed to identify candidate genes. Patl2 knock-out
12 (Patl2^{-/-}) mice were generated to study their reproductive phenotype and
13 to perform a comparative transcriptomic analysis against WT.

14 Results
15

16 Six out of the 23 tested subjects (26%) harbored a homozygous PATL2
17 truncating mutation. Patl2^{-/-}female mice exhibited a severe subfertility
18 associated with oocyte spindle anomalies and poor developmental
19 competence of oocytes and embryos. PATL2, as a translation factor, is
20 believed to be involved in storage, processing, regulation, and/or
21 degradation of mRNA. Transcriptomic analysis confirmed that the absence
22 of Patl2 impacts numerous genes known to be crucial for oocyte meiotic
23 progression and early embryonic development.

24 Conclusions
25

26 We demonstrate that PATL2 is crucial for RNA regulation in human and
27 mice oocytes and that PATL2 gene mutation is a main genetic cause of
28 oocyte meiotic deficiency.

29
30 SCHOLARONE™
31 Manuscripts
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

***PATL2* gene mutation causes oocyte meiotic deficiency and female infertility**

Marie Christou-Kent¹, Amir Amiri-Yekta^{1,2,3}, Zine-Eddine Kherraf¹, Thomas Karaouzène¹, Jessica Escoffier¹, Audrey Guttin⁴, Guillaume Martinez^{1,2,6}, Emilie Le Blévec¹, Emeline Lambert¹, Selima Fourati Ben Mustapha⁷, Isabelle Cedrin Durnerin⁸, Lazhar Halouani⁷, Ouafi Marrakchi⁷, Mounir Makni⁷, Habib Latrous⁷, Mahmoud Kharouf⁷, Serge P. Bottari¹, Nicolas Thierry-Mieg⁹, Charles Coutton^{1,2,6}, Raoudha Zouari⁷, Jean Paul Issartel⁴, Pierre F. Ray^{1,2,10,†} and Christophe Arnoult^{1,†,§}

¹Genetic Epigenetic and Therapies of Infertility, Institute for Advanced Biosciences, Inserm U1209, CNRS UMR 5309, Université Grenoble Alpes, F-38000 Grenoble, France

²UM GI-DPI, CHU de Grenoble, Grenoble, F-38000, France

³Department of Genetics, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine ACECR, Tehran, Iran.

⁴Grenoble Neuroscience Institute, INSERM 1216, Université Grenoble Alpes, F-38000 Grenoble, France

⁶UM de Génétique Chromosomique, CHU de Grenoble, Grenoble, F-38000, France

⁷Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation, Centre Urbain Nord, 1003 Tunis, Tunisia

⁸ Service de Médecine de la Reproduction, Centre Hospitalier Universitaire Jean Verdier, Assistance Publique - Hôpitaux de Paris, 93143 Bondy, France

⁹Univ. Grenoble Alpes / CNRS, TIMC-IMAG, F-38000 Grenoble, France

¹⁰UFR de Biochimie Génétique et Moléculaire, CHU de Grenoble, Grenoble, F-38000, France.

[†] Shared leadership

[§] Corresponding author: Dr Christophe Arnoult, PhD, DMV Tel: +33-476-637-408; E-mail: christophe.arnoult@univ-grenoble-alpes.fr

Key Words: Female sterility, oocyte maturation failure, oocyte maturation arrest, oocyte developmental competence, *Patl2*,

Abstract

Background— Infertility impacts the life of over 70 million couples worldwide, yet its molecular basis remains largely unknown. Some women present primary infertility with no menstruating defects. They respond well to the IVF ovarian stimulation protocol with normal follicular growth however the collected oocytes all appear immature or degenerated and cannot be fertilized even after recourse to assisted reproductive technologies. The aim of this study was to identify the genetic defects inducing this oocyte meiotic deficiency (OMD) and to characterize its molecular pathogeny.

Methods and subjects

A total of 23 OMD subjects were recruited and whole exome and Sanger sequencing was performed to identify candidate genes. *Patl2* knock-out (*Patl2*^{-/-}) mice were generated to study their reproductive phenotype and to perform a comparative transcriptomic analysis against WT.

Results

Six out of the 23 tested subjects (26%) harbored a homozygous *PATL2* truncating mutation. *Patl2*^{-/-} female mice exhibited a severe subfertility associated with oocyte spindle anomalies and poor developmental competence of oocytes and embryos. *PATL2*, as a translation factor, is believed to be involved in storage, processing, regulation, and/or degradation of mRNA. Transcriptomic analysis confirmed that the absence of *Patl2* impacts numerous genes known to be crucial for oocyte meiotic progression and early embryonic development.

Conclusions

We demonstrate that *PATL2* is crucial for RNA regulation in human and mice oocytes and that *PATL2* gene mutation is a main genetic cause of oocyte meiotic deficiency.

INTRODUCTION

In humans, oocyte production is a long process starting during embryonic development and characterized by a long diapause which lasts for decades until the production of a mature oocyte at each menstrual cycle. The quiescent oocytes are blocked in the prophase of the first meiosis and form an immature oocyte containing a germinal vesicle (GV) and embedded in a primordial follicle. Periodically, a group of follicles is recruited to form a growing follicular pool, where close interactions between follicular cells and the growing oocytes allow the production of a fully grown oocyte in an antral follicle¹ in approximately 100 days. At this stage, the oocyte is sensitive to hormonal stimulation, which triggers meiosis resumption associated with GV break down (GVBD) which entails the completion of the first meiosis with extrusion of the first polar body followed by a new arrest at the metaphase 2 (MII) of the second meiosis. Completion of meiosis with the exclusion of the second polar body is triggered by fertilization. Several reports describe that some infertile women repetitively produce mostly immature oocytes, a poorly-defined syndrome known as "oocyte factor" or "bad eggs syndrome"²⁻⁵. Our cohort consisted of patients who all had at least one IVF cycle with only GV, MI or atretic oocytes and we called this phenotype oocyte meiotic deficiency (OMD). The characterization of several KO mice has permitted the identification of several genes inducing a meiotic block at different stages. For instance, in mice with an invalidated Cdc25b gene, involved in cyclic AMP control, oocyte maturation arrests at the GV stage^{6,7}. Similarly, the absence of H1foo, a transcription factor, Mei1, required for normal meiotic chromosome synapsis and Ubb, an ubiquitin controlling the destruction of key cell cycle regulators, lead to MI arrest⁸⁻¹⁰. Finally, Smc1b, a meiosis-specific component of the cohesin complex, is involved in MII arrest¹¹ while Mlh3, which maintains homologous chromosome pairing at meiosis, induces mixed arrests¹². These genes are therefore suitable candidates for OMD but none of them have so far been associated with OMD in humans. Heterozygous mutations of *TUBB8*, an oocyte specific tubulin necessary for the meiotic spindle were recently identified in a cohort of Chinese patients with OMD¹³, establishing *TUBB8* as the first human gene identified in the context of OMD. In our present study, we analyzed a total of 23 unrelated

1
2 OMD patients coming from North and Sub Saharan Africa and found that 6 (26%) have a homozygous
3 truncating mutation in the *PATL2* gene indicating that absence of PATL2 is the main cause of OMD in
4 this region.
5
6
7
8
9
10
11

METHODS

Human subjects

A total of 23 patients were recruited. All subjects originated from North Africa, mainly Tunisia and Algeria and one patient from Mauritania, and underwent one or two cycles of ovarian stimulation for the purpose of egg collection for further *in vitro* fertilization. The patients responded normally to ovarian stimulation with a number of harvested oocytes similar to control patients (Figure S1AB). However, no MII oocytes were collected, the only oocytes harvested were either arrested immature oocytes (GV or MI stages) or oocytes presenting an irregular shape with a dark ooplasm and considered atretic (Figure S1C). The description of oocyte parameters of the patients harboring a *PATL2* mutation is described in Table 1.

Study design

Whole exome sequencing was performed for 15 subjects. Sanger sequencing of *PATL2* was then performed on an additional 8 subjects. The importance and role of PATL2 in oocyte maturation was assessed using *Patl2* knock-out mice (*Patl2*^{-/-}). Detailed methods are provided in the Methods section in the Supplementary Appendix.

Ethics

Informed consent was obtained from all the subjects participating in the study according to the local IRB protocols and the principles of the Declaration of Helsinki. All animal procedures were run according to the French guidelines on the use of animals in scientific investigations with the approval

1
2
3 of the local Ethics Committee (ComEth Grenoble N° 318, ministry agreement number #7128 UHTA-
4
5 U1209-CA).
6
7

8 **RESULTS**
9

10
11 **Whole-exome and Sanger sequencing identifies a homozygous truncating mutation in *PATL2* in 26% of the**
12 **tested subjects.**
13

14
15 In the present study, we analyzed a cohort of 23 infertile women presenting with OMD. Given the
16 fact that most of the patients are Tunisian and that 20-30% of marriages are consanguineous in this
17 country, we thus postulated that infertility was likely transmitted through recessive inheritance and
18 we focused on homozygous mutations.
19
20

21
22 Exome analysis was performed for 15 patients. After exclusion of common variants and application of
23 technical and biological filters, only three genes were found to be homozygously mutated in at least
24 two subjects (Table S1). We have not yet explored the variants found only in one patient. Two genes
25 were mutated in two patients and did not appear good candidates: *FAM58A* because the identified
26 variant is in fact common in the Genome Aggregation Database (gnomAD),
27 (<http://gnomad.broadinstitute.org/>) and *MGAM* because of its expression (primarily in the digestive
28 tract) and type of mutation (probably benign). Remarkably we identified 5 women (33%) with the
29 same homozygous loss of function variant p.Arg160Ter, c.478C>T in *PATL2* transcript
30 ENST00000434130. This variant introduces a premature stop codon in the coding sequence, resulting
31 (at best) in the production of a non-functional truncated protein with less than the first third of the
32 protein (Figure 1A). As the *xpat1a* gene, the xenopus ortholog of *PATL2*, has been described to be
33 expressed during oocyte growth^{14,15}, *PATL2* appeared an excellent candidate gene. This variant was
34 identified in a heterozygous state in 5 out of 148,732 alleles ([rs548527219](#)) in the Genome
35 Aggregation Database (gnomAD), a very low frequency of 0.003362% compatible with the expected
36 key role of *PATL2* in female reproduction. The presence of the variant was confirmed by Sanger
37 sequencing on the five mutated patients (Figure 1B). Sanger sequencing of *PATL2* coding sequences
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 was then performed on an additional 8 OMD subjects. One carried the same homozygous mutation,
4 giving 6 out of 23 analyzed subjects (26%) carrying the *PATL2* p.Arg160Ter variant.
5
6
7 Our exome data was then screened for the presence of *TUBB8* heterozygous mutations which have
8 been described to induce OMD¹³ and no deleterious *TUBB8* variants were identified in the 15
9 analyzed subjects.
10
11
12
13
14
15
16
17

18 ***Patl2* KO female mice display subfertility due to compromised oocyte maturation and poor developmental
19 competence of oocytes and embryos.**

20
21 Data bank analyses confirmed that *PATL2* is highly expressed in both human and mouse oocytes
22 (Figure S2) and we observed that its expression is very low in human follicular cells (Figure S2),
23 suggesting that the maturation defect is of oocyte rather than follicular origin. To confirm the
24 involvement of *PATL2* in our patient's phenotype, we assessed the reproductive phenotype of *Patl2*
25 deficient mice (*Patl2*^{-/-}) harboring an exon 7 deletion inducing a translational frameshift downstream
26 (Figure S3). We first studied the fertility of *Patl2*^{-/-} females crossed with WT males by measuring the
27 number of pups produced per litter, the number of pups produced over a period of 6 months and the
28 number of litters per month. Females lacking *Patl2* exhibited a severe subfertility: the number of
29 pups per litter dropped from 7.2 ± 0.9 n=13 for WT to 2.3 ± 0.4 n=7 for *Patl2*^{-/-} mice (Figure 2A) and
30 both the number of pups and of litters per month/per female were reduced (Figure S4AB).
31 Conversely no difference in litter size was observed between WT and *Patl2*^{-/-} males (7.2 ± 0.9 n=13 for
32 WT and 8.7 ± 0.5 n=35 for *Patl2*^{-/-} males, (Figure S4C)).
33
34

35 This severe subfertility could be due to oocyte and/or embryonic developmental defects. To better
36 characterize the female *Patl2*^{-/-} phenotype, we first performed a histological study of the ovaries. No
37 obvious defects were observed at this level and *Patl2*^{-/-} and WT ovaries had similar cellular structure
38 (Figure S5). Next, ovarian stimulation was performed and GV and MII oocytes were collected for
39 morphological and *in vitro* fertilization (IVF) studies. *Patl2*^{-/-} GV oocytes, which displayed a smaller
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1 diameter than WT GV oocytes, showed a well-defined apparent nucleolus (Figure S6AB). Similar
2 to observations in OMD women, *Patl2*^{-/-} and WT mice produced the same number of follicles and
3 oocytes after full ovarian stimulation (Figure S6C). Nevertheless, contrary to what happened in
4 women, part of the collected oocytes reached the MII stage (Figure S6D). It is worth noting that as
5 for GV, MII *Patl2*^{-/-} oocytes had a smaller diameter than control (Figure S6E), suggesting that oocyte
6 growth was impaired in the absence of Patl2. Next the developmental competence of MII oocytes
7 was challenged through IVF. The IVF outcomes were significantly altered in *Patl2*^{-/-} females with the
8 percentage of eggs reaching the 2-cell stage dropping from 67.7% ± 8.1, n=5 in WT mice to 36.4% ±
9 6.4, n=5 in *Patl2*^{-/-} mice (Figure 2C), demonstrating that the developmental competence of *Patl2*^{-/-}
10 oocytes was compromised. This decrease of 2-cell *Patl2*^{-/-} embryos was mostly due to defective
11 development of *Patl2*^{-/-} zygotes, which exhibited numerous defects including delayed pronucleus
12 formation, absence of sperm DNA decondensation or/and polyspermy whereas almost all fertilized
13 WT zygotes exhibited 2 pronuclei (Figure 2D-F). The poor developmental competence of *Patl2*^{-/-} eggs
14 and their abnormal fertilization finally impacted preimplantation development since only 27.2% ±
15 5.1, n=4 of 2-cell embryos generated with *Patl2*^{-/-} eggs reached the blastocyst stage in contrast to
16 87.1% ± 5.6, n=4 with WT eggs (Figure 2C and S6F). The impaired ability of *Patl2*^{-/-} oocytes to sustain
17 embryo development was likely due to numerous defects observed in MII oocytes such as abnormal
18 morphology of the spindle, misalignment of the chromosomes on the spindle and numerous
19 cytoplasmic asters (Figure 3AB). It is worth noting that the lack of Patl2 also impacts meiosis
20 resumption, blocking a significant number of oocytes in a defective MI stage (Figure 3CD).

50 Absence of Patl2 induces an important alteration of the transcriptome of GV and MII oocytes.

51
52 Interestingly, *xpat1a*, the xenopus ortholog of *PATL2*, has been reported to be expressed during
53 oocyte growth, to interact with the cytoplasmic polyadenylation element binding complex (CPEB)
54 and to repress RNA translation in xenopus oocytes^{14,15}. This suggests that PATL2 could be involved in
55
56
57
58
59
60

1
2 specific modulation of the oocyte transcriptome during oocyte maturation in Mammals. To address
3 this point, global gene expression analysis was performed on oocytes collected at the GV and MII
4 stages from WT and *Patl2*^{-/-} females.
5
6

7
8 The expression of nearly 66,000 transcripts was measured across the different oocyte groups with
9 Affymetrix microarrays. First, we verified that oocyte RNA purification was not contaminated by RNA
10 from follicular cells, by comparing expression levels of specific follicular and oocyte genes (Figure
11 S7A). The absence of exon 7 transcription in *Patl2*^{-/-} oocytes' extracts was also verified on microarray
12 data (Figure S7B). First, there was no difference in global purified RNA concentration between WT
13 and *Patl2*^{-/-} for GV and MII oocytes, indicating that the overall transcription process is not controlled
14 by Patl2 (Figure S7C). We however observed significant changes in specific transcripts, both at the GV
15 and MII stages (Figure 4B). At the GV stage, lack of Patl2 induces a >two-fold decrease of 95
16 transcripts ($p < 0.05$) and >two-fold increase of 39 transcripts (Table S2) and at the MII stage, a >two-
17 fold decrease of 124 transcripts and >two-fold increase of 122 transcripts (Table S3). Around one
18 third of the downregulated genes at the GV stage (32) were also downregulated at the MII stages,
19 and half of the upregulated genes at the GV stage (19) were also upregulated at the MII stage (Table
20 S4 and Figure 4C). The impact of the absence of Patl2 on gene expression at the GV and MII stages
21 was visualized using hierarchical clustering of genes with an absolute fold-change ($(aFC) > 2$, $p < 0.05$)
22 (Figure 4D).
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Discussion

Genetic alteration of *PATL2* is the main cause of OMD in North African women.

The genetic basis of OMD remained unexplored until last year when heterozygous mutations of *TUBB8* were described to be responsible for approximately 30% (7/23) of OMD Chinese Han familial cases¹³. Surprisingly, we found no *TUBB8* deleterious variant in any of the 15 subjects analyzed by WES, indicating that *TUBB8* mutations are not a frequent cause of OMD in North /Sub Saharan Africa.

This result underlines a geographical heterogeneity of genetic causes of OMD. Here, we observed that 26% of our patients present the same homozygous truncating mutation in *PATL2* demonstrating that *PATL2* is also one of the main genes leading to OMD. Because all the tested patients were of North African/Sub Saharan origin, we were unable to assess the frequency of *PATL2* gene mutation in subjects suffering from OMD from other origins. All women exhibit the same mutation, suggesting a founder effect. Analysis of variants from WES data in the *PATL2* surrounding regions shows a common homozygous haplotype thus confirming this hypothesis (Table S5).

Additionally, we did not observe any downregulation of the expression of the tubulin genes in GV and MII oocytes from *Patl2*^{-/-} mice indicating that *PATL2*-dependent OMD is not caused by a tubulin deregulation.

Importance of *Patl2* in mRNA regulation for oocyte maturation and zygote development

We have shown that the absence of *Patl2* induces a transcriptomic deregulation, affecting 135 genes at GV and 248 at MII stages presenting an absolute fold-change >2, ($p < 0.05$). A literature survey of the downregulated genes in *Patl2*^{-/-} oocytes enabled us to identify several groups of genes reported to be involved in oocyte maturation. The encoded proteins are actors of several signaling pathways involved in oocyte differentiation, oxidative stress, transcription and translation, exocrine modulation, and meiosis and spindle formation (Table S6). We therefore expect that the decreased expression of these genes should interfere with normal oocyte differentiation. Regarding signaling pathways activated during oocyte maturation, a significant downregulation of transcripts of proteins involved in mTORC1, Wnt, NF- κ B, MAP kinase and phosphatase pathways was spotted. It is worth noting that a > two-fold decrease of Pgrmc1 is also observed, a receptor necessary for slowing down oocyte meiotic progression¹⁶ (Table S6). Oocyte maturation is controlled by a bi-directional cross-talk between the follicular cells and the oocyte and the secretion of oocyte origin factors is necessary for follicular cell differentiation, which in turn secretes factors activating different signaling pathways

1
2
3 within the oocyte¹. Interestingly, two factors Cxcl14 and Adm2, known to play a crucial role in
4 cumulus cell maturation^{17,18}, are downregulated in *Patl2*^{-/-} oocytes (Table S6). We also observed a
5 strong deregulation of several transcription factors (Table S6). Among them, Sohlh2, which was
6 downregulated 2.2 folds and 5.3 folds at the GV and MII stages, respectively, is particularly
7 interesting as it has been described as being necessary for oocyte growth and KO mice were found
8 infertile¹⁹. Interestingly this factor does not affect meiosis I²⁰, in agreement with the phenotype of
9 *Patl2*^{-/-} females, which are able to generate MII oocytes. We also noticed that two glutathione-S-
10 transferases were repressed in *Patl2*^{-/-} GV oocytes, which may increase oxidative stress within the
11 oocyte (Table S6). Oocytes are very sensitive to oxidative stress, which leads to spindle
12 abnormalities²¹ and impacts developmental competence²². Moreover, glutathione-S-transferase is a
13 marker of oocyte maturity in pigs²³. Spindle defects may also be aggravated by Pak4 and Ccdc69
14 repression, two proteins known to affects spindle assembly through Ran-GTPase for Pak4^{24,25} (Table
15 S6). We expect that the overall decreased production of these proteins likely contributes to
16 abnormal meiosis. Some transcripts such as *Fgf9* and *Cdc25a* were strongly upregulated after the
17 GVBD and control meiosis II²⁶. Interestingly both corresponding transcripts are significantly
18 downregulated in *Patl2*^{-/-} MII oocytes (Table S6), such a decrease likely negatively impacts final
19 oocyte maturation. Among common upregulated genes (Table S4), the most upregulated gene in MII
20 is *Prr11*. In WT MII oocytes, its expression was low, suggesting that its role in oogenesis is minimal or
21 null. *Prr11* deregulation has however been shown to dramatically modify the cell cycle, although its
22 specific role remains unclear²⁷. Its strong upregulation may therefore interfere with meiosis or early
23 development. Another remarkable common GV-MII upregulated gene is *Ska2* (spindle and
24 kinetochore associated complex subunit 2), known to control spindle stability during meiosis²⁸. This
25 upregulation is in accordance with the numerous defects observed in MI and MII spindles. Altogether
26 these results underline the importance of PATL2 in the regulation of specific mRNAs necessary for
27 the generation of mature oocytes.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

If most of the mRNAs synthetized during oocyte growth are immediately translated and correspond to proteins supporting oocyte growth, around 30% of mRNA are stored for subsequent translation and are necessary for meiosis resumption and early zygote development. It is worth noting that a considerable number of genes known to be involved in preimplantation embryo development are also downregulated in both *Patl2*^{-/-} GV and MII oocytes and likely contribute to the poor preimplantation developmental competence observed for *Patl2*^{-/-} embryos (Figure 2). These results suggest that *Patl2* is essential in maintaining the integrity of a small pool of mRNA synthetized during oocyte growth and necessary after fertilization during early embryo development. This result is in agreement with its function as a translation repressor shown in *Xenopus* oocytes¹⁴.

Identifying key actors of oocyte maturation is mandatory for a better mastery of *in vitro* oocyte maturation.

By unraveling the molecular basis of OMD, this work will help the patients who will benefit from a better diagnosis and comprehension of their pathology. It will also be of tremendous interest to the fast growing field of clinical *in vitro* oocyte maturation, necessary for the development of different applications such as fertility preservation before engaging in reprotoxic chemotherapies for patients diagnosed with cancer and in particular for young prepubertal girls²⁹ or *in vitro* maturation of primary/secondary follicles for patients suffering from polycystic ovary syndrome and from premature ovarian failure^{30,31}.

In conclusion, we show for the first time here that PATL2 is one of the vital actors of oocyte maturation through the regulation of the levels of mRNAs encoding proteins which are crucial for oocyte maturation and early embryonic development and that its invalidation results in OMD in humans.

ACKNOWLEDGMENTS

We thank the IAB microscopy platform and Mylene PEZET, Alexei GRICHINE, Jacques MAZZEGA for their technical help. We thank Emeline FONTAINE PELLETIER (INSERM 1209, CNRS UMR 5309) for

1
2
3 antibodies. We thank Marie-Christine BIRLING for help in mouse genotyping. We thank Marcio
4
5 CRUZEIRO (Institut Cochin, Paris, France) for providing the EIIaCre transgenic mice. We thank our
6
7 patients and control individuals for their participation.
8
9

10
11 This work was mainly supported by the following grants: "Investigation of the genetic aetiology of
12
13 oocyte meiotic failure (OMF) by exome sequencing" funded by the fondation maladies rares (FMR)
14
15 for the program High throughput sequencing and rare diseases 2012 and by the "MAS-Flagella"
16
17 project financed by French ANR and the DGOS for the program PRTS 2014, the "Whole genome
18
19 sequencing of patients with Flagellar Growth Defects (FGD)".
20
21
22

23
24
25
26 **AUTHORS' ROLES**
27
28

29 CA and PFR analyzed the data and wrote the manuscript; Z-EK, AA-Y, CC, performed molecular work;
30 TK, NT-M analyzed genetic data; MCK performed IF and histological experiments; MCK, JE, ELB and
31 GM performed IVF experiments; J-PI, AG performed transcriptome analyses; EL and SB performed
32 biochemistry experiments; SFBM, IC-D, LH, OM, MM, HL, MK and RZ provided clinical samples and
33 data; CA and PFR designed the study, supervised all laboratory work, had full access to all the data in
34 the study and took responsibility for the integrity of the data and its accuracy. All authors
35 contributed to the report.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 LEGENDS OF TABLE/FIGURES
4
5
6

7
8
9
10
11
12
13
14
Table 1. Medical history, laboratory investigations and oocyte collection outcomes of six patients
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
presenting with OMD due to *PATL2* mutation. Control corresponds to couples where the male suffers from azoospermia (n=234), NDoR (not determined or recorded)

Figure 1. Identification of a truncating mutation in PATL2

(A) Location of the *PATL2* mutation in the intron-exon structure and in the protein structure representation. The identified variant, homozygous in the 6 patients, is in exon 6 and creates a STOP codon, ending the translation and giving a truncated protein of 158 amino acids (aa) lacking the PAT1 (Topoisomerase II-associated protein PAT1) domain instead of 543 aa for the full length protein. (B) Electropherograms of Sanger sequencing for *PATL2* mutated patients compared to reference sequence.

Figure 2. infertility of *Patl2* deficient female mice is due to low developmental competence of MII oocytes

(A) Comparative accumulation of live pups over a period of 6 months from 3 WT and 3 *Patl2*^{-/-} females crossed with WT males shows a severe hypofertility phenotype of *Patl2*^{-/-} females. (B) Histograms showing the number of pups per litter (mean ± SD) obtained by crossing WT and *Patl2*^{-/-} females with WT males. (C) IVF outcomes measured at the 2-cell, morula and blastocyst stages show that the developmental competence of *Patl2*^{-/-} oocytes is compromised. Oocytes were collected from stimulated WT and *Patl2*^{-/-} females and sperm from WT males. (D) Z-stack projections of confocal images of 2PN zygotes obtained from WT and *Patl2*^{-/-} oocytes 6-8 h after sperm-egg mixing. Fertilized WT oocytes exhibit normal 2PN stage (D1) whereas the number of fertilized *Patl2*^{-/-} oocytes exhibiting normal 2PN stage is strongly reduced (D2) and most of them show defects such as absence of male pronucleus (D3), partial decondensation of male PN (D4, D5, white arrows) or polyspermy (D5 white arrows). Scale bars 20 μm, pb polar body (E) The rate of 2PN at 6-8 h after fertilization

1
2
3 drops from 53.4% for WT to 14.4% for *Patl2*^{-/-} oocytes, which exhibit various fertilization defects
4
5 including partial decondensation of sperm DNA, polyspermy, abnormal number of PN or mixed
6
7 defects (**F**).
8
9

10 **Figure 3. MII oocytes from *Patl2*^{-/-} female mice exhibit various defects preventing normal embryo**
11 **development.**
12
13

14 **(A)** Oocytes collected after ovarian stimulation were labelled with a tubulin antibody (red) and
15 counterstained with DAPI for DNA staining (blue). In control MII oocytes, stack projections of
16 confocal images show that the spindle was symmetric and the chromosomes distributed in the
17 middle of the spindle. In contrast, in *Patl2*^{-/-} MII oocytes various defects were observed such as
18 irregular spindle shape, spindle rotation and numerous cytoplasmic asters. Slightly more oocytes
19 with abnormal chromosome distribution were also observed. **(B)** Histograms quantifying the % of
20 defects observed in *Patl2*^{-/-} MII oocytes. **(C)** An increase of non MII oocytes (blockage in first meiosis,
21 absence of pb) after full ovarian stimulation was observed in *Patl2*^{-/-} mice. **(D)** In *Patl2*^{-/-} oocytes
22 blocked in MI stage, various defects were observed as irregular shape of the spindle and abnormal
23 distribution of the chromosomes.
24
25
26
27
28
29
30
31
32
33
34
35
36

37 **Figure 4. Transcriptome analysis of GV and MII oocytes from WT and *Patl2*^{-/-} mice.**
38
39

40 **(A)** Comparison of the transcriptomic profiles in *Patl2*^{-/-} oocytes vs WT oocytes at the GV or MII
41 stages. GV oocytes were collected 44 h after PMSG injection and MII 13 h after HCG injections. For
42 MII and GV oocytes, 2 replicates of WT and 3 replicates of *Patl2*^{-/-} oocytes were analyzed **(B)** Venn
43 diagram representing down or up regulated genes in *Patl2*^{-/-} oocytes (absolute fold-change ((aFC)> 2,
44 p< 0.05,) with respect to WT oocytes at GV and MII stages. **(C)** Hierarchical clustering of gene
45 expression data for the down- and up-regulated genes (aFC> 2, p< 0.05) of *Patl2*^{-/-} and WT oocytes at
46 the GV (left) and the MII (right) stages, demonstrating the clustering of replicates to their respective
47 groups.
48
49
50
51
52
53
54
55
56
57
58
59
60

References

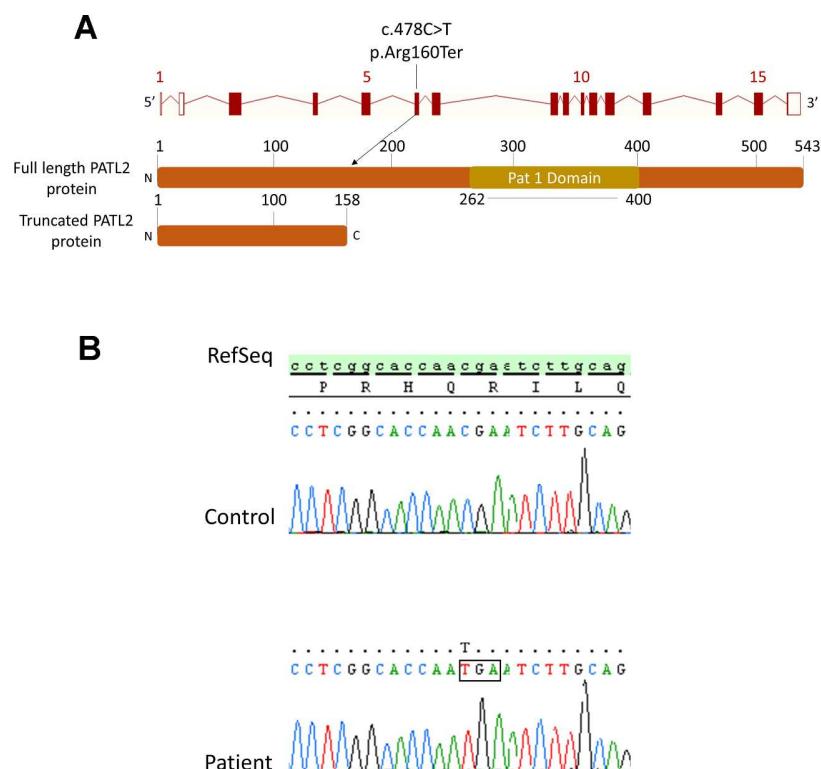
1. Li R, Albertini DF. The road to maturation: somatic cell interaction and self-organization of
2 the mammalian oocyte. *Nature reviews Molecular cell biology* 2013;14:141-52.
3. Levran D, Farhi J, Nahum H, Glezerman M, Weissman A. Maturation arrest of human oocytes
4 as a cause of infertility: case report. *Human reproduction (Oxford, England)* 2002;17:1604-9.
5. Hourvitz A, Maman E, Brengauz M, Machtlinger R, Dor J. In vitro maturation for patients with
6 repeated in vitro fertilization failure due to "oocyte maturation abnormalities". *Fertility and sterility*
7 2010;94:496-501.
8. Hartshorne G, Montgomery S, Klenzneris L. A case of failed oocyte maturation in vivo and in
9 vitro. *Fertility and sterility* 1999;71:567-70.
10. Beall S, Brenner C, Segars J. Oocyte maturation failure: a syndrome of bad eggs. *Fertility and*
11 *sterility* 2010;94:2507-13.
12. Lincoln AJ, Wickramasinghe D, Stein P, et al. Cdc25b phosphatase is required for resumption
13 of meiosis during oocyte maturation. *Nature genetics* 2002;30:446-9.
14. Vaccari S, Horner K, Mehlmann LM, Conti M. Generation of mouse oocytes defective in cAMP
15 synthesis and degradation: endogenous cyclic AMP is essential for meiotic arrest. *Developmental*
16 *biology* 2008;316:124-34.
17. Furuya M, Tanaka M, Teranishi T, et al. H1foo is indispensable for meiotic maturation of the
18 mouse oocyte. *The Journal of reproduction and development* 2007;53:895-902.
19. Libby BJ, De La Fuente R, O'Brien MJ, et al. The mouse meiotic mutation mei1 disrupts
20 chromosome synapsis with sexually dimorphic consequences for meiotic progression. *Developmental*
21 *biology* 2002;242:174-87.
22. Ryu KY, Sinnar SA, Reinholdt LG, et al. The mouse polyubiquitin gene Ubb is essential for
23 meiotic progression. *Molecular and cellular biology* 2008;28:1136-46.
24. Takabayashi S, Yamauchi Y, Tsume M, Noguchi M, Katoh H. A spontaneous smc1b mutation
25 causes cohesin protein dysfunction and sterility in mice. *Experimental biology and medicine*
26 (Maywood, NJ) 2009;234:994-1001.
27. Lipkin SM, Moens PB, Wang V, et al. Meiotic arrest and aneuploidy in MLH3-deficient mice.
28 *Nature genetics* 2002;31:385-90.
29. Feng R, Sang Q, Kuang Y, et al. Mutations in TUBB8 and Human Oocyte Meiotic Arrest. *The*
30 *New England journal of medicine* 2016;374:223-32.
31. Marnef A, Maldonado M, Bugaut A, et al. Distinct functions of maternal and somatic Pat1
32 protein paralogs. *RNA (New York, NY)* 2010;16:2094-107.
33. Nakamura Y, Tanaka KJ, Miyauchi M, Huang L, Tsujimoto M, Matsumoto K. Translational
34 repression by the oocyte-specific protein P100 in Xenopus. *Developmental biology* 2010;344:272-83.
35. Guo M, Zhang C, Wang Y, et al. Progesterone Receptor Membrane Component 1 Mediates
36 Progesterone-Induced Suppression of Oocyte Meiotic Prophase I and Primordial Folliculogenesis.
37 *Scientific reports* 2016;6:36869.
38. Bobe J, Montfort J, Nguyen T, Fostier A. Identification of new participants in the rainbow
39 trout (*Oncorhynchus mykiss*) oocyte maturation and ovulation processes using cDNA microarrays.
40 *Reproductive biology and endocrinology : RB&E* 2006;4:39.
41. Chang CL, Wang HS, Soong YK, Huang SY, Pai SY, Hsu SY. Regulation of oocyte and cumulus
42 cell interactions by intermedin/adrenomedullin 2. *The Journal of biological chemistry*
43 2011;286:43193-203.

- 1
2
3 19. Choi Y, Yuan D, Rajkovic A. Germ cell-specific transcriptional regulator sohlh2 is essential for
4 early mouse folliculogenesis and oocyte-specific gene expression. *Biology of reproduction*
5 2008;79:1176-82.
6 20. Shin YH, Ren Y, Suzuki H, et al. Transcription factors SOHLH1 and SOHLH2 coordinate oocyte
7 differentiation without affecting meiosis I. *The Journal of clinical investigation* 2017;127:2106-17.
8 21. Choi WJ, Banerjee J, Falcone T, Bena J, Agarwal A, Sharma RK. Oxidative stress and tumor
9 necrosis factor-alpha-induced alterations in metaphase II mouse oocyte spindle structure. *Fertility*
10 and sterility 2007;88:1220-31.
11 22. Rausell F, Pertusa JF, Gomez-Piquer V, et al. Beneficial effects of dithiothreitol on relative
12 levels of glutathione S-transferase activity and thiols in oocytes, and cell number, DNA fragmentation
13 and allocation at the blastocyst stage in the mouse. *Molecular reproduction and development*
14 2007;74:860-9.
15 23. Paczkowski M, Krisher R. Aberrant protein expression is associated with decreased
16 developmental potential in porcine cumulus-oocyte complexes. *Molecular reproduction and*
17 *development* 2010;77:51-8.
18 24. Bompard G, Rabeharivelo G, Cau J, Abrieu A, Delsert C, Morin N. P21-activated kinase 4
19 (PAK4) is required for metaphase spindle positioning and anchoring. *Oncogene* 2013;32:910-9.
20 25. Pal D, Wu D, Haruta A, Matsumura F, Wei Q. Role of a novel coiled-coil domain-containing
21 protein CCDC69 in regulating central spindle assembly. *Cell cycle (Georgetown, Tex)* 2010;9:4117-29.
22 26. Assou S, Anahory T, Pantesco V, et al. The human cumulus--oocyte complex gene-expression
23 profile. *Human reproduction (Oxford, England)* 2006;21:1705-19.
24 27. Li J, Sun P, Yue Z, Zhang D, You K, Wang J. miR-144-3p Induces Cell Cycle Arrest and Apoptosis
25 in Pancreatic Cancer Cells by Targeting Proline-Rich Protein 11 Expression via the Mitogen-Activated
26 Protein Kinase Signaling Pathway. *DNA and cell biology* 2017.
27 28. Zhang QH, Qi ST, Wang ZB, et al. Localization and function of the Ska complex during mouse
28 oocyte meiotic maturation. *Cell cycle (Georgetown, Tex)* 2012;11:909-16.
29 29. Kim SY, Kim SK, Lee JR, Woodruff TK. Toward precision medicine for preserving fertility in
30 cancer patients: existing and emerging fertility preservation options for women. *Journal of*
31 *gynecologic oncology* 2016;27:e22.
32 30. Kim JY. Control of ovarian primordial follicle activation. *Clinical and experimental*
33 *reproductive medicine* 2012;39:10-4.
34 31. Yin O, Cayton K, Segars JH. In Vitro Activation: A Dip Into the Primordial Follicle Pool? *The*
35 *Journal of clinical endocrinology and metabolism* 2016;101:3568-70.
- 36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

	Geographical origin	age (years)	Number of collected oocytes				Total	FSH, U/L >40 in case of POI	LH, U/L	TSH, U/L	Prolactine, µg/L	Menstruation	Comments
			GV	MI	MII	Atretic							
P1	Tunisia	35	4	0	0	1	5	NDR	NDR	1.36	NDR	NDR	
		34					2						Medical records not available
		34					8						Medical records not available
P2	Tunisia	28	9	0	0	11	20	NDR	NDR	NDR	NDR	YES	
		28.9	15		0	4	19						
P3	Tunisia	24	11	0	0	5	16	10.31	3.54	NDR	22.28	NDR	1 GV matures to M1 in vitro
P4	Tunisia	34.28	8	0	0	2	10	NDR	NDR	1.07	23	YES	
P5	Algeria	41	2	0	0	2	4	9.39	6.1	2.3	12.9	YES	
		42	3	1	0	1	5						
P6	Mauritania	36	2	10	0	4	16	3.01	3.38	2.88	25	YES	Cytoplasmic vacuoles in M1 oocyte
		36.8	0	0	0	5	5						
Patients' mean		34.00	6.00	1.38	0	3.89	10.00						
Normal values n=234		34.4	1.2	0.6	6	1.3	9.1	<10.2	<16.9	0.5-5	2-20		

Table 1

Confidential: Destroy when review is complete.



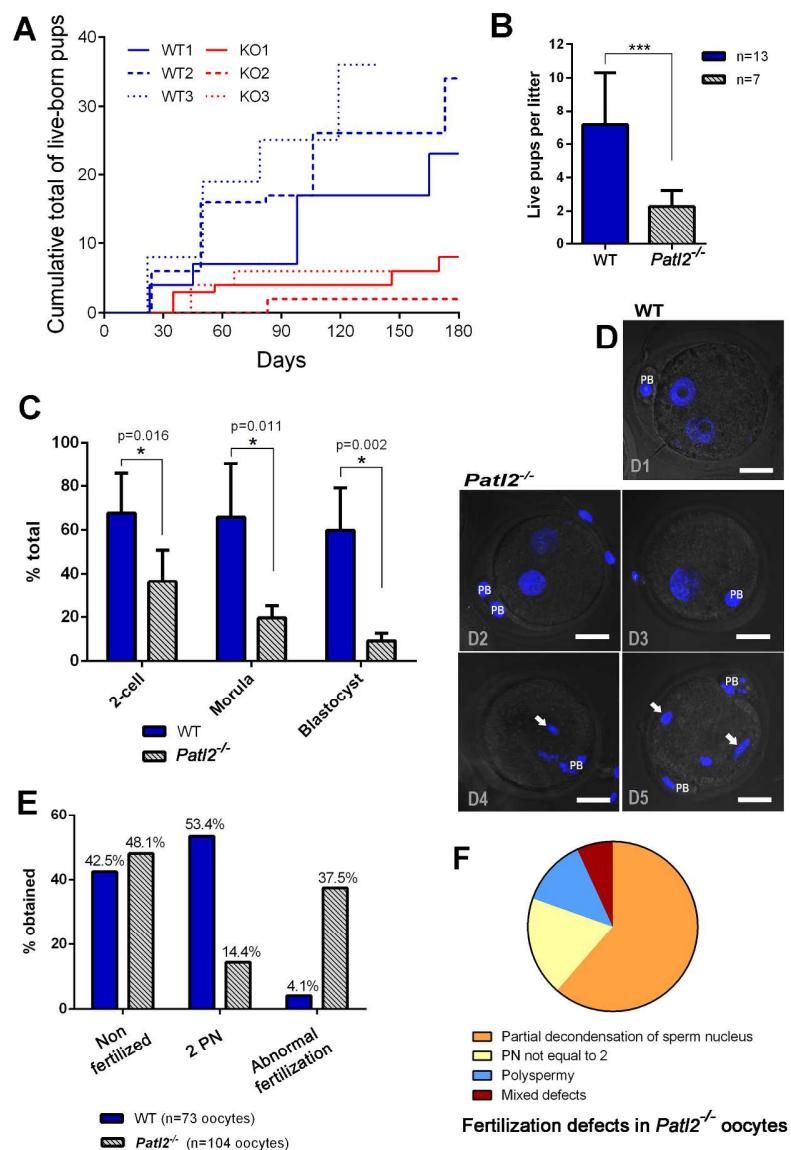
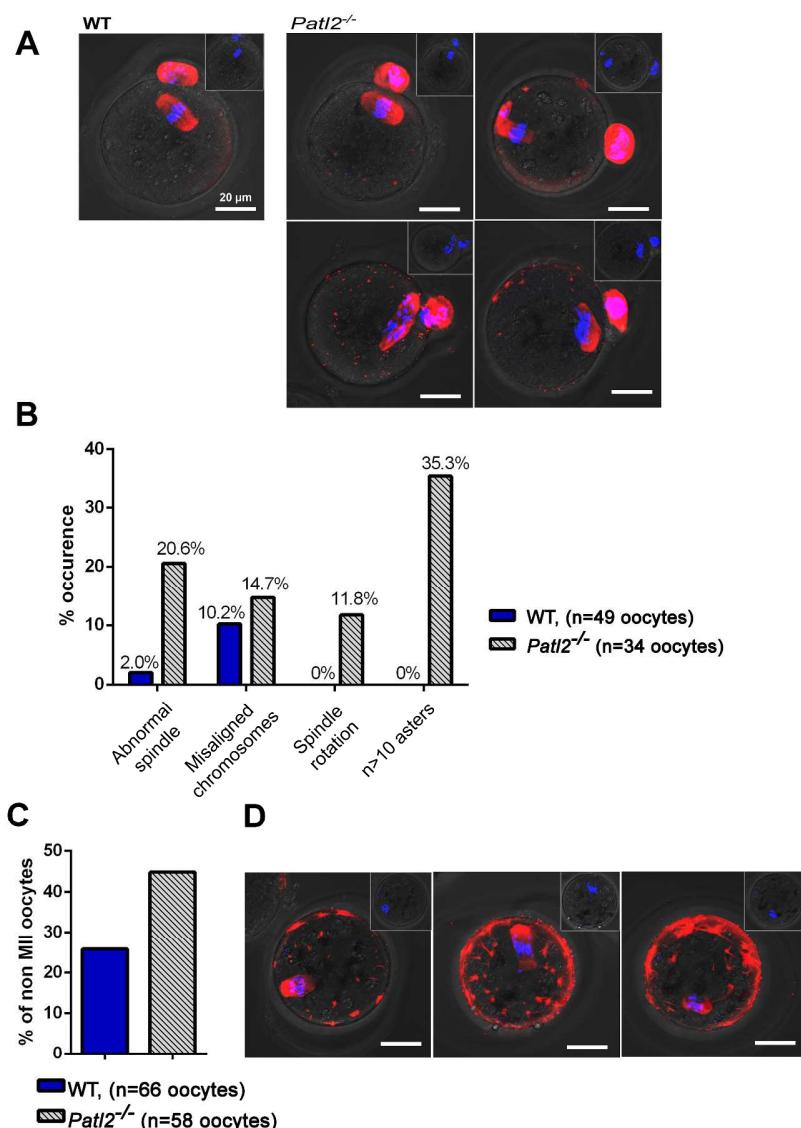


Figure 2. Infertility of *Patl2* deficient female mice is due to low developmental competence of MII oocytes

209x297mm (300 x 300 DPI)



Christou-Kent et al, Figure 3

Figure 3. MII oocytes from *Patl2*^{-/-} female mice exhibit various defects preventing normal embryo development.

297x420mm (300 x 300 DPI)

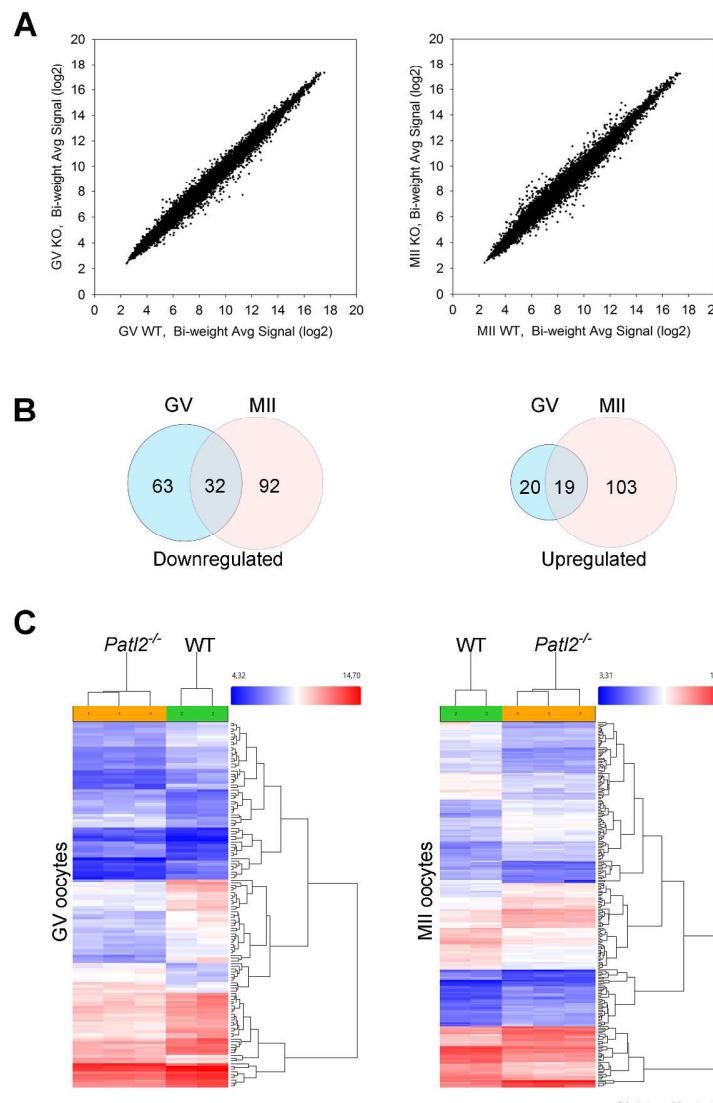


Figure 4. Transcriptome analysis of GV and MII oocytes from WT and *Patl2*^{-/-} mice.

209x297mm (300 x 300 DPI)

Principaux résultats

L'application de notre pipeline d'analyse sur les données de ces femmes nous a permis d'obtenir une liste de 316 variants impactant 299 gènes différents. Parmi ces variants, aucun n'impactait le gène *TUBB8*. Afin de restreindre à nouveau la liste de gènes, nous nous sommes concentrés sur ceux retrouvés mutés à l'état homozygote chez au moins 2 femmes. Seul 3 gènes ont passé ce nouveau critère : *FAM58A*, *MGAM* et *PATL2*. Aucune investigation n'a pour l'instant été effectuée sur les 296 autres. En raison de la fréquence élevée du variant retrouvé sur *FAM58A* et de l'impact peu délétère du variant chevauchant *MGAM*, ces deux gènes ont été considérés comme de mauvais candidats.

Ainsi, nous nous sommes dans un premier temps concentrés sur la caractérisation de *PATL2* dont l'orthologue *xpat1a* chez le xénophage a été décrit comme étant exprimé au cours du développement de l'ovocyte [206, 207] faisant de ce gène un excellent candidat. Nous avons observé que 5 de nos patientes (33.3%) étaient porteuses de la même mutation homozygote : c.478G>T, p.Arg160Ter induisant un codon stop prématûr dans la séquence codante du transcrit canonique *PATL2* : ENST00000434130. Au vu de ces résultats un séquençage Sanger de la séquence codante de ce gène fut réalisé pour ces 5 femmes afin de confirmer cette mutation, ainsi que sur 8 femmes supplémentaires souffrant du même phénotype. Parmi ces dernières, une s'est révélée porter la même mutation à l'état homozygote. Au final six femmes sur 23 (26%) étaient homozygotes pour la mutation stop de *PATL2*.

Dans un second temps, l'étude du modèle murin KO *Patl2*^{-/-} nous a permis de mettre en évidence une subfertilité importante chez les souris femelles tandis qu'aucun phénotype n'était observable chez les mâles.

Pour finir, *xpat1a*, l'orthologue de *PATL2* chez le xénophage, ayant été décrit comme réprimant la traduction de l'ARNm dans l'ovocyte, nous avons cherché à savoir si les souris femelles KO *Patl2*^{-/-} présentaient des dérégulations de leur transcriptome ovocytaire. Pour cela, nous avons procédé à une étude comparative des transcriptomes ovocytaires aux stades GC et MII murins sur puces Affymetrix mesurant les valeurs d'expression d'environ 66,000 transcrits différents. Ainsi, nous avons pu mettre en évidence 134 transcrits différemment exprimés au stade GV parmi lesquels 95 étaient sous-exprimés tandis que 39 étaient sur-exprimés. Au stade MII, ces dérégulations se révélèrent être plus impressionnantes puisque 124 étaient sous-exprimées et 122 sur-exprimées démontrant ainsi une forte implication de *Patl2* dans la transcription ovocytaire des gènes murins.

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

2.4.1 Article n°5

Whole exome cohort study and analysis of mouse and Trypanosoma models demonstrate the importance of WDR proteins in flagellogenesis and male fertility

Coutton C, Vargas A, Amiri-Yekta A, Kherraf ZE, Fourati Ben Mustapha S, Le Tanno P, Wambergue-LeGrand C, Karaouzène T, Martinez G, Daneshipour A, Hanieh Hosseini S, Mitchell V, Halouani L, Marrakchi O, Makni M, Latrous H, Kharouf M, Deleuze JF, Boland A, Hennebicq S, Satre V, Jouk PS, Bottari SP, Thierry-Mieg N, Conne B, Dacheux-Deschamps D, Schmitt A, Stouvenel L, Lorès P, El Khouri E, Fauré J, Wolf JP, Escoffier J, Gourabi H, Robinson DR, Nef S, Duloust E, Zouari R, Bonhivers M, Touré A, Arnoult C, Ray PF

Nature Communications. En révision

Contexte et objectifs

Après avoir mis en évidence l'implication du gène *DNAH1* dans le phénotype MMAF notre équipe s'est en partie spécialisée dans la caractérisation ce syndrome. Ainsi, entre 2012 et 2016, notre équipe a effectué le séquençage de 78 individus présentant tous ce phénotype afin d'en établir la cause génétique. Ces séquençages ont été effectués dans 4 centres différents que sont le Genoscope, Integragen, Novogene et la plateforme de séquençage de l'IGBMC de Strasbourg. La plupart de ces séquençages ont été effectués sur des Illumina HiSeq2000 sauf les 25 plus récents qui ont été effectués sur un Illumina HiSeq4000 (**Table : 2.6**).

Table 2.6 – Liste des différents individus présentant un phénotype MMAF séquencé en WES

Place	Year	Platform	Nb of individuals
IGBMC	2012	Illumina HiSeq2000	12
Genoscope	2013	Illumina HiSeq2000	12
Genoscope	2014	Illumina HiSeq2000	25
Genoscope	2015	Illumina HiSeq2000	3
Integragen	2016	Illumina HiSeq4000	15
Novogene	2016	Illumina HiSeq4000	10

AJHG The American Journal of Human Genetics

Whole exome cohort study and analysis of mouse and Trypanosoma models demonstrate the importance of WDR proteins in flagellogenesis and male fertility

--Manuscript Draft--

Manuscript Number:	
Full Title:	Whole exome cohort study and analysis of mouse and Trypanosoma models demonstrate the importance of WDR proteins in flagellogenesis and male fertility
Article Type:	Article
Keywords:	Infertility, spermatogenesis, flagella, gene defect, diagnosis
Corresponding Author:	Pierre F Ray, PhD CHU de Grenoble GRENOBLE, FRANCE
First Author:	Pierre F Ray, PhD
Order of Authors:	Pierre F Ray, PhD Charles Coutton Alexandra Vargas Amir Amiri-Yekta Zine-Eddine Kherraf Selima Fourati Ben Mustapha Pauline Le Tanno Clémentine Wambergue-Legrand Thomas Karaouzène Guillaume Martinez Abbas Daneshipour Seyedeh Hanieh Hosseini Valérie Mitchell Lazhar Halouani Ouafi Marrakchi Mounir Makni Habib Latrous Mahmoud Kharouf5 Jean-François Deleuze Anne Boland Sylviane Hennebicq Véronique Satre Pierre-Simon Jouk Serge P Bottari Nicolas Thierry-Mieg Beatrice Conne Denis Dacheux-Deschamps Alain Schmitt

	Laurence Stouvenel Patrick Lorès Elma El Khouri Julien Fauré Jean-Philippe Wolf Jessica Escoffier Hamid Gourabi Dereck R Robinson Serge Nef Emmanuel Dulioust Raoudha Zouari Mélanie Bonhivers Aminata Touré Christophe Arnoult
Abstract:	<p>Severe defects of spermatogenesis concern millions of men worldwide yet the vast majority remains without a diagnosis. Here we studied men with primary infertility due to multiple morphological abnormalities of the sperm flagella (MMAF) with severe anomalies of the sperm axoneme, a structure conserved between flagella and cilia. Whole exome sequencing was performed on 78 patients allowing the identification of 21 men with homozygous or compound heterozygous mutations in DNAH1 (n=5), the only previously described MMAF gene, and in CFAP43 (n=10) and CFAP44 (n=6). CFAP43 and CFAP44 are two WD repeat containing proteins of unknown function, highly expressed in the testis. Two knockout (KO) mouse lines were created using the CRISPR/Cas9 technique and protein functions were studied using <i>Trypanosoma brucei</i>. We demonstrate that homozygous males from both KO mouse lines were infertile and present severe flagellar defects. Moreover we observed that in heterozygous animals, Cfap43 mice had an impaired sperm motility and Cfap44 animals an increased proportion of morphologically abnormal spermatozoa. All studied models confirmed the importance of CFAP43 and CFAP44 in flagellogenesis and showed that they are functionally evolutionary conserved proteins that are critical to link the central-pair microtubules with the radial spokes.</p>
Suggested Reviewers:	<p>Hector Chemes CONICET Buenos Aires hchemes@cedie.org.ar He was one of the first to describe the MMAF syndrome</p> <p>François Vialard CHI Poissy, France fvialard@chi-poissy-st-germain.fr He is an expert in the genetics of male infertility</p> <p>Abraham Kierszenbaum City College of New York kier@med.cuny.edu He is an expert in spermatogenesis and flagella formation.</p>
Opposed Reviewers:	

Dear Editor,

I was stunned when I saw that you have just published the manuscript: "Biallelic Mutations in CFAP43 and CFAP44 Cause Male Infertility with Multiple Morphological Abnormalities of the Sperm Flagella" as we are ourselves in the process of publishing an extremely similar manuscript entitled : "Whole exome cohort study and analysis of mouse and Trypanosoma models demonstrate the importance of WDR proteins in flagellogenesis and male fertility" in which we describe that we identified 10 and 6 patients harboring biallelic mutations in CFAP43 and CFAP44 respectively. Our manuscript has just been evaluated by Nature Genetics and we were encouraged to transfer it to Nature Communication. In the light of the publication by Tang et al. we decided to submit it instead to American Journal of Human Genetics in the hope of providing your readers with a better overview of CFAP43 and CFAP44's role in flagellogenesis.

As you will see our work is more exhaustive than what is presented by Tang and colleagues and we provides many additional valuable information:

- Our cohort is more important with 78 patients with a total of 21 mutated patients (for CFAP43 and CFAP44 and DNAH1). Moreover our cohort is more international and not restricted to the Han population.
- We also confirmed these results by generating Crisp/Cas9 KO mice. We performed a finer characterization of sperm defects in KO animals and describe that heterozygous mice also present some minor anomalies.
- We have performed a more extensive study of the impact of the absence of protein on the different substructure of the axonema (Radial Spoke, outer and inner Dynein Arm , Central Pair,...), both by immunofluorescence and scanning electron microscopy. Our data strongly suggest that both proteins are important for central pair/radial spoke interactions.
- With trypanosoma, we present a new and interesting model bringing some invaluable results:
 - o We show that both proteins are located in the flagellum in mature axonema thus indicating that these proteins are necessary for the organization of the axonema.
 - o Their absence does not prevent axonemal growth, clearly indicating that these proteins are not mandatory for intra flagellar transport (IFT)
 - o Survival of trypanosoma is compromised, confirming that flagellum beat is defective in the absence of CFAP43/44

I therefore truly hope that you will accept to review our manuscript in a timely fashion, hopefully permitting the publication of our manuscript as back to back or with as short a delay as possible. I cannot tell you how important this manuscript is to us as we have, over the last four years, invested critical resources and it would be a tremendous disappointment to have our story snatched from us so close to the final line.

Best regards,

Pierre Ray

Whole exome cohort study and analysis of mouse and Trypanosoma models demonstrate the importance of WDR proteins in flagellogenesis and male fertility

Charles Coutton^{1,2}, Alexandra Vargas¹, Amir Amiri-Yekta^{1,3,4}, Zine-Eddine Kherraf^{1,3}, Selima Fourati Ben Mustapha⁵, Pauline Le Tanno^{1,2}, Clémentine Wambergue-Legrand^{1,3}, Thomas Karaouzène¹, Guillaume Martinez^{1,3}, Abbas Daneshipour⁴, Seyedeh Hanieh Hosseini⁶, Valérie Mitchell⁷, Lazhar Halouani⁵, Ouafi Marrakchi⁵, Mounir Makni⁵, Habib Latrous⁵, Mahmoud Kharouf⁵, Jean-François Deleuze⁸, Anne Boland⁸, Sylviane Hennebicq^{1,9}, Véronique Satre^{1,2}, Pierre-Simon Jouk¹⁰, Serge P. Bottari¹, Nicolas Thierry-Mieg¹¹, Beatrice Conne¹², Denis Dacheux-Deschamps^{13,14}, Alain Schmitt^{15,16,17}, Laurence Stouvenel^{15,16,17}, Patrick Lorès^{15,16,17}, Elma El Khouri^{15,16,17}, Julien Fauré^{18,19}, Jean-Philippe Wolf^{17,20}, Jessica Escoffier¹, Hamid Gourabi⁴, Dereck R Robinson¹³, Serge Nef¹², Emmanuel Duloust^{17,20}, Raoudha Zouari⁵, Mélanie Bonhivers^{13†}, Aminata Touré^{15,16,17 †}, Christophe Arnoult^{1†}, and Pierre F. Ray^{1,3 †,Σ}

† Shared leadership

¹Genetic Epigenetic and Therapies of Infertility, Institute for Advanced Biosciences, Inserm U1209, CNRS UMR 5309, Université Grenoble Alpes, F-38000 Grenoble, France

³CHU de Grenoble, UM de Génétique Chromosomique, Grenoble, F-38000, France

³CHU de Grenoble, UM GI-DPI, Grenoble, F-38000, France

⁴Department of Genetics, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran. PO Box 16635-148

⁵Polyclinique les Jasmins, Centre d'Aide Médicale à la Procréation, Centre Urbain Nord, 1003 Tunis, Tunisia

⁶Department of Andrology, Reproductive Biomedicine Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran. PO Box 16635-148

⁷EA 4308, Department of Reproductive Biology and Spermiology-CECOS, Lille University Medical Center, Lille, France.

⁸Centre National de Génotypage, Institut de Génomique, CEA, Evry, France.

⁹CHU de Grenoble, UF de Biologie de la procréation, Grenoble, F-38000, France

¹⁰CHU de Grenoble, UF de Génétique Médicale, Grenoble, F-38000, France

¹¹Univ. Grenoble Alpes / CNRS, TIMC-IMAG, F-38000 Grenoble, France

¹²Department of Genetic Medicine and Development University of Geneva Medical School, CH 1211 Geneva 4 Switzerland

¹³University Bordeaux, Microbiologie Fondamentale et Pathogénicité, CNRS UMR 5234, Bordeaux, France

¹⁴Institut Polytechnique de Bordeaux, Microbiologie Fondamentale et Pathogénicité, UMR-CNRS 5234, F-33000 Bordeaux, France.

¹⁵Institut National de la Santé et de la Recherche Médicale, INSERM U1016. Institut Cochin. Paris 75014, France

¹⁶Centre National de la Recherche Scientifique, CNRS UMR8104. Paris 75014, France

¹⁷Université Paris Descartes, Sorbonne Paris Cité, Faculté de Médecine. Paris 75014, France

¹⁸CHU de Grenoble, UF de Biochimie Génétique et Moléculaire, Grenoble, F-38000, France

¹⁹Grenoble Neuroscience Institute, INSERM 1216, Grenoble, F38000, France

²⁰Laboratoire d’Histologie Embryologie - Biologie de la Reproduction, GH Cochin Broca Hôtel Dieu, Assistance Publique-Hôpitaux de Paris. Paris 75014, France²¹Laboratoire d’Histologie Embryologie - Biologie de la Reproduction, GH Cochin Broca Hôtel Dieu, Assistance Publique-Hôpitaux de Paris. Paris 75014, France

Σ **Corresponding author:** Pierre F Ray, UM, 38043 Grenoble cedex 9, France. Tel: +33-4-76-76-55-73; E-mail: pray@chu-grenoble.fr

ABSTRACT

Severe defects of spermatogenesis concern millions of men worldwide yet the vast majority remains without a diagnosis. Here we studied men with primary infertility due to multiple morphological abnormalities of the sperm flagella (MMAF) with severe anomalies of the sperm axoneme, a structure conserved between flagella and cilia. Whole exome sequencing was performed on 78 patients allowing the identification of 21 men with homozygous or compound heterozygous mutations in *DNAH1* (n=5), the only previously described MMAF gene, and in *CFAP43* (n=10) and *CFAP44* (n=6). *CFAP43* and *CFAP44* are two WD repeat containing proteins of unknown function, highly expressed in the testis. Two knockout (KO) mouse lines were created using the CRISPR/Cas9 technique and protein functions were studied using *Trypanosoma brucei*. We demonstrate that homozygous males from both KO mouse lines were infertile and present severe flagellar defects. Moreover we observed that in heterozygous animals, *Cfap43* mice had an impaired sperm motility and *Cfap44* animals an increased proportion of morphologically abnormal spermatozoa. All studied models confirmed the importance of *CFAP43* and *CFAP44* in flagellogenesis and showed that they are functionally evolutionary conserved proteins that are critical to link the central-pair microtubules with the radial spokes.

INTRODUCTION

The global burden of infertility is substantial and is likely to further increase with the observation in the past decades of a widespread decline in semen quality^{1,2,3} which might be correlated with the global environmental change⁴. Medical treatment of infertility has rapidly

evolved over the past four decades but much remains to be accomplished⁵. Additional efforts should be pursued to better characterize male infertility, first focusing on gene identification, which is key to improve diagnosis efficiency, patient management and to develop personalized medicine in the field of reproduction. Most genetic causes of male infertility are currently uncharacterized, likely due to the large number of genes involved in human spermatogenesis⁶. The emergence of high-throughput, genome-wide genomic approaches such as whole-exome sequencing (WES) has revolutionized the identification of new genes involved in human diseases⁷ and now allows unprecedented progresses in the field of infertility^{8,9}. Genetic characterization of gross morphological abnormalities of the sperm, or monomorphic teratozoospermia, has been particularly successful^{10,11}. Teratozoospermia, which accounts among the most frequent and most severe phenotype of male infertility¹², represents a heterogeneous group including a wide range of abnormal sperm phenotypes affecting, solely or simultaneously, the head, the neck, and the sperm tail. We demonstrated previously that mutations in the *DNAH1* gene are responsible for multiple morphological abnormalities of the flagella (MMAF), a relatively frequent teratozoospermia phenotype characterized by severe asthenozoospermia (i.e. absence or reduction of sperm motility) due to a combination of flagellar defects including bent, curled, abnormal width, rolled or absent flagella¹³. *DNAH1* encodes an axonemal inner dynein arm heavy chain, the lack of which leads to a strong disorganization of the axoneme,¹³. In our primary study, *DNAH1* was found mutated in about one third the 20 patients analysed presenting with a severe MMAF phenotype¹³. Novel *DNAH1* mutations were subsequently found in MMAF patients from various geographical origins in two replicative studies with a prevalence ranging from 33% to 37.5%^{14,15}. These results confirmed the strong implication of *DNAH1* in the MMAF phenotype but also indicates that MMAF is genetically heterogeneous and that other genes are likely to be involved in this syndrome¹¹. In the present study, we analyzed 78 additional MMAF patients using WES and

showed that in addition to mutations in *DNAH1*, mutations in *CFAP43* and *CFAP44*, two genes encoding for WD Repeat domains (WDR) containing proteins, are responsible for MMAF syndrome and account for 20.5% of the cases. Most importantly, we investigated the role of these novel genes by performing gene invalidation and silencing in two evolutionary distant models, *Trypanosoma* and mouse, yet sharing an extremely conserved flagellar structure. Using this original approach we demonstrate the importance of WDR proteins for axonemal structure of the flagella and male fertility in humans.

RESULTS

Whole-exome sequencing identifies homozygous truncating mutations in *CFAP43* and *CFAP44* in MMAF patients.

In the present study we analyzed a cohort of 78 individuals presenting with a MMAF phenotype defined by the presence in the ejaculate of immotile spermatozoa with several abnormalities of the sperm flagellum including short, coiled, absent and flagella of irregular caliber¹³ (**Fig. 1**). The average semen parameters of all 78 MMAF patients included in the cohort is described in **Table 1**. Nearly no spermatozoa with normal morphology could be observed in the ejaculate of MMAF individuals (1.6%); an average of 20.7% and 43.7% of spermatozoa had no flagella and short flagella, respectively, and 31.7% of the spermatozoa had flagella with an irregular caliber. As a result, total sperm motility was dramatically reduced to 3.9% (normal value > 40%), which prevented natural conception for all individuals and led them to consult for infertility. Given the notion of consanguinity for most individuals from the cohort, we postulated that infertility was likely to be transmitted through recessive inheritance and therefore likely resulted from homozygous mutations. After exclusion of frequent variants and applying stringent filters, a limited list of homozygous variants was identified for each proband. First, we identified 6 patients (7.7%) with mutations in the *DNAH1* gene (Table 2)

which was previously identified as the main cause for the MMAF phenotype^{13,14}. We subsequently identified 10 subjects with variants in *CFAP43* (12.8%), 8 of which had a homozygous loss of function variant and two had two likely deleterious variants (Table2). In addition 6 subject (7.7%) had a homozygous loss of function variant in *CFAP44* (Table2). These two CFAP genes (for Cilia and Flagella Associated Protein) were reported in public expression databases as strongly expressed in the testis and appear to be connected with cilia and flagella structure and/or functions¹⁶. These results were confirmed by RT-qPCR experiments performed in human and mouse tissue panels. In both cases, the relative expression of *CFAP43* and *CFAP44* mRNA in testis was predominant and very significantly higher than in the other tested tissues (**Supplementary Fig. 1**). Taking into account the high number of mutated patients and the specific expression pattern of the two genes we focused on these two genes which appeared as the best candidates to explain the primary infertility observed for these individuals. All individuals with *CFAP43* or *CFAP44* mutations were unrelated and to our knowledge, none of them carried rare variants in genes that were previously reported to be associated with male infertility. *CFAP43* and *CFAP44* encoded proteins belong to the WDR protein family and are both composed of nine WD repetitions¹⁷.

CFAP43 (also known as *WDR96*, NM_025145) is localized on chromosome 10 and contains 38 exons encoding a predicted 1665-amino acid protein (Q8NDM7). We identified 9 different pathogenic variants in *CFAP43*, in 10 unrelated individuals from the cohort . The splicing variant c.3541-2A>C was identified in 2 unrelated individuals and affects a consensus splice acceptor base of *CFAP43* intron 27 (**Table 2, Fig. 2**). This variant was absent from the 60,706 unrelated individuals aggregated in the ExAC database (<http://exac.broadinstitute.org>), consistent with potential negative selection throughout evolution for such pathogenic mutation. Three other variants correspond to stop-gain mutations identified in three subjects: c.2658G>A and c.2680C>T are located in exon 21 and c.3352C>T

is located in the exon 26 (**Fig. 2**). These three nonsense mutations were found in the ExAC database with very low allele frequency ranging from 8.24e-06 to 9.89e-05 (1 to 12 mutated alleles). Two others mutations were small frameshift indels: c.1240_1241delGT (found in two unrelated patients), and c.3882delA (**Table 2, Fig. 2**), not listed in ExAC. All these mutations generate a premature stop codon and are considered to be “damaging” by prediction softwares. Apart from these obvious harmful mutations found in 8 patients, we also identified two patients harboring likely deleterious mutations in *CFAP43*. The first patient harbors a homozygous splicing mutation c.2141+5G>A, not listed in ExAC which, according to the splice site prediction algorithm Human Splicing Finder (<http://www.umd.be/HSF3>) , likely alter the consensus splice donor site of *CFAP43* exon 16. Unfortunately, we could not obtain any additional biological material from this patient and could not verify the effect of this variant on mRNA expression. The second patient is a compound heterozygous carrying one single-nucleotide duplication, c.1320dupT, located in the exon 11 and an additional missense mutation, Val347Ala, located in exon 8 of the *CFAP43* gene. Although prediction software classified this last variant as likely “benign”, this mutation affect a conserved residue located in the N-terminal part of the protein within a WD-repeat known to be important for protein /protein interactions¹⁸. Moreover, this missense variant is found at a very low prevalence in the general population estimated to 7.45e-5 (9/12078). Altogether the above identified truncating mutations clearly indicate that in humans, loss of function of *CFAP43* is associated with asthenozoospermia and sperm flagellum defects.

CFAP44 (also known as WDR52, NM_001164496) is localized on chromosome 3 and contains 35 exons encoding a predicted 1854-amino acid protein (Q96MT7). Five different homozygous variants were identified in 6 unrelated patients (**Table 2, Fig. 2**). The stop-gain mutations c.1387G>T and c.3175C>T are located in exon 12 and 23, respectively. The

frameshift mutations, c.4767delT, with a single-nucleotide deletion, is located in exon 31 and the variant c.2818dupG, with a single-nucleotide duplication, is located in exon 21. The last identified mutation is a splicing mutation, c.1890+1C>T, that affects a consensus splice donor base of *CFAP44* intron 15. All these mutations are predicted to generate a premature stop codon and are considered as “damaging” by prediction softwares. None of these variants were listed in the ExAC database. The presence of these variants was verified by bidirectional Sanger sequencing (**Supplementary Fig. 2**).

The detailed sperm parameters of the patients carrying *DNAH1*, *CFAP43* or *CFAP44* mutations were compared with each other. There was no significant difference between these three groups (**Table 1**), as illustrated by the similar morphologies of sperm from patient P₄₃-8 and P₄₄-3, mutated for CFAP43 and CFAP44, respectively (**Fig. 1**). Data from all these groups were therefore pooled and compared with data from the patients without an identified mutation. All patients with a mutation had a significant higher rate of spermatozoa with short or absent flagella and present a significant lower motility rate compared to patients with no identified mutations. There was no difference in the other sperm parameters (**Table 1**).

Mutations in CFAP43 and CFAP44 lead to severe axonemal disorganization

The internal cytoskeleton of motile cilia and flagella, named the axoneme, is a highly evolutionary conserved structure which consists of nine doublets of microtubules (DMTs) circularly arranged around the central pair complexe of microtubules (CPC) ('9+ 2' structure). Beating of cilia and flagella is orchestrated by multiprotein-ATPase complexes, located on the peripheral doublets, which provide the sliding force for sperm motility. In addition, the sperm flagellum harbors specific peri-axonemal structures, which are not found in other motile cilia, an helical mitochondrial sheath (MS) in the midpiece, the fibrous sheath (FS) in the principal piece (PP) and outer dense fibers (ODF) in the midpiece and the proximal part of the PP.

We studied the ultrastructure of sperm cells from mutated patients in *CFAP43* and *CFAP44* by transmission electron microscopy (TEM) (**Fig. 3**). Due to an insufficient amount of sperm cells collected from most *CFAP44* and *CFAP43* patients, only one patient for each gene was studied with TEM. For each patient, we could observe longitudinal sections and >20 cross-sections presenting a sufficient quality to observe the ultrastructure of the axonemal components. Longitudinal sections showed severe axonemal and peri-axonemal defects affecting the ODF, the fibrous sheath FS and the MS which appeared completely disorganized resulting in aborted flagella or their replacement by a cytoplasmic mass englobing unassembled axonemal components (**Fig. 3**). In *CFAP43* and *CFAP44* patients, approximately 95% of the cross-sections were abnormal and the main defect observed was an absence of the CPC (9+0 conformation) observed in 81.8 %, and 66.7% of the observations respectively, compared to 0% in control fertile subject (**Supplementary Table 1**). In the residual fraction (~5%) of normal axoneme (9+2 conformation), peri-axinonemal structures abnormalities were constantly observed (**Supplementary Table 1**). The lack of central pair defects is associated with peripheral doublets defects in 13.6 and 19% for *CFAP43* and *CFAP44*, respectively (**Supplementary Table 1**). Cross-sections with a single central microtubule (9+1 conformation) were observed in about 10% of cases, only for the *CFAP44* patient. Interestingly in *CFAP43* patients, the CPC, when present, was misoriented compared to control sections in which the CPC is normally parallel to the axis of the two longitudinal columns of the FS (**Fig. 3**).

To define the ultrastructural defects evidenced by TEM, we performed immunofluorescence (IF) experiments using antibodies targeting different axonemal proteins. In *CFAP43* and *CFAP44* patients, staining of SPAG6, a protein located in the CPC, was abnormal and atypical. In *CFAP43*, it was totally absent of the patient's spermatozoa (**Fig. 4b1-b3**), whereas in *CFAP44* patients' sperm cells, SPAG6 immunostaining was present but with

an abnormal and diffuse pattern concentrated in the midpiece of the spermatozoa, quickly vanishing along the flagellum (**Fig. 4c1-c3**). Additionally, staining of the radial spoke head protein RSPH1 presented significantly different patterns from controls. In CFAP43 patients, the RSPH1 staining was completely abnormal with a marked diffuse staining, concentrated in the midpiece, whereas the tubulin staining remains restricted to the axoneme (**Fig. 4e1-e3**). In CFAP44 patients' sperm cells, staining of RSPH1 was significantly reduced (**Fig. 4f1-f3**). For CFAP43 and CFAP44 patients, immunostaining for AKAP4, DNALI1, DNAI2 and GAS8 were all comparable with controls, suggesting that FS, outer dynein arms (ODAs), inner dynein arms (IDAs) and the nexin-dynein regulatory complex (N-DRC), respectively were not directly affected by mutations in *CFAP43* or *CFAP44*, in contrast to the central pair complex (**Supplementary Table 2-Human**).

Orthologues of CFAP43 and CFAP44 in *Trypanosoma* are located in the axoneme and are necessary for cell survival and axonemal integrity.

To overcome the absence of reliable antibodies against CFAP43 and CFAP44 proteins in human and murine cells, and to analyze CFAP43 and CFAP44 in a tractable model organism (which can be used for forward and reward genetics), we decided to characterise *TbCFAP44* and *TbCFAP43* in *Trypanosoma brucei* (*T. brucei*), a flagellated protozoan used as system model to study flagellar proteins.

BLASTp analysis on *T. brucei* genome database¹⁹ using human CFAP43 and CFAP44 sequences identified *T. brucei* orthologs *Tb927.7.3560* (named *TbCFAP44* in this study) and *Tb927.4.5380* (named *TbCFAP43* in this study), respectively. Previous functional genomics and proteomic studies identified *TbCFAP44* and *TbCFAP43* as flagellar proteins^{20,21}. In addition, *TbCFAP44* is the *Chlamydomonas* FAP44 ortholog, and is involved in flagellar

motility (also named *TbCMF7*)²². However, the function of *TbCFAP44* and *TbCFAP43* are currently unknown.

We first localized *TbCFAP44* and *TbCFAP43* in *T. brucei*, using myc-tagged proteins by generating Trypanosome cell lines expressing endogenous levels of C-terminal myc-tagged proteins *TbCFAP44*_{myc} and *TbCFAP43*_{myc}. Both proteins were found in the axoneme as substantiated by co-labelling with an antibody against the paraflagellar rod (PFR) structure (**Fig. 5a**).

To assess the function of *TbCFAP44* and *TbCFAP43* in the trypanosome flagellum, we knocked-down protein expression by inducible RNA interference (RNAi), either in the parental cell line (cell lines *TbCFAP44*^{RNAi}, *TbCFAP43*^{RNAi}), or in the cell lines expressing the myc-tagged proteins (cell lines *TbCFAP44*_{myc}^{RNAi}, *TbCFAP43*_{myc}^{RNAi}). Cell proliferation was assessed in parental, non-induced and induced *TbCFAP44*^{RNAi} and *TbCFAP43*^{RNAi} cells (or in *TbCFAP44*_{myc}^{RNAi} and *TbCFAP43*_{myc}^{RNAi} cell lines showing the same results, data not shown). Importantly, both *TbCFAP44*^{RNAi} and *TbCFAP43*^{RNAi} induced cells stopped proliferating after 24h and died (**Fig. 5b**). This growth defect was accompanied by a defect in cytokinesis producing multiflagellated cells (**Supplementary Fig. 3**), a phenotype previously described when proteins directly or indirectly involved in flagellar motility are knocked-down²³. Efficiency and specificity of RNAi knockdown of *TbCFAP44* and *TbCFAP43* was validated by RT-PCR (**Supplementary Fig. 4**) and by immunofluorescence, showing a clear decrease of myc labelling in the new flagellum of *TbCFAP44*_{myc}^{RNAi} and *TbCFAP43*_{myc}^{RNAi} induced cells (**Fig. 5a**). The impact of *TbCFAP44* and *TbCFAP43* knockdown was also investigated by TEM. In control longitudinal sections of parental cells, the flagellum exits the cell through the flagellar pocket (**Fig. 5c1**), and, in cross section, the canonical ultrastructure of the axoneme, composed of 9 doublets of microtubules (DMTs) and the central pair of microtubules (CPC), is observed (**Fig. 5c2**). In contrast, *TbCFAP44*^{RNAi} and *TbCFAP43*^{RNAi} induced cells were abnormal and

displayed more than 2 flagella in one abnormally enlarged flagellar pocket (**Fig. 5c3,c5**). In cross-section, flagella exhibited abnormal axonemes with either 90°rotated CPC, a defect never observed in control trypanosome cell lines²⁴ (**Fig. 5c3**, white arrows) or displaced CPC and DMTs (**Fig. 5c4** and **c6**). No obvious default in the basal body structure was observed (data not shown). All together these results confirm the essential role of CFAP43 and CFAP44 in the ultrastructure of the axoneme.

Male infertility and flagellar anomalies are also observed in *Cfap43* and *Cfap44* CRISPR/Cas 9 knock out (KO) mice.

Finally, we assessed the impact on spermatogenesis of *Cfap43* and *44* absence in mouse by generating KO mice using the CRISPR-Cas9 technology. We obtained 3 independent mouse strains for *Cfap43* and 2 for *Cfap44* with different insertions/deletions of a few nucleotides, all inducing a translational frameshift expected to lead to complete absence of the protein or production of a truncated protein. Because all strains for *Cfap43* presented the same reproductive phenotype, we restricted our study to a strain with a 4 pb deletion in the exon 21 (delAAGG). Similarly, both *Cfap44* lines presented the same reproductive phenotype, and we focused on a strain with a 7 pb insertion in exon 3 (InsTCAGATA). RT-PCR was performed on testis ARN from *Cfap43*^{-/-} and *Cfap44*^{-/-} mice which confirmed the production of abnormal transcripts in both mutants leading to a premature stop codon (**Supplementary Fig. 5**).

Reproductive phenotype was studied for both KO mice models. First homozygous KO females were fully fertile and gave litters of normal size (7.8±1.8 and 7.3±3.5 versus 6.7±0.5 pups/litter (mean ± SD) for *Cfap43*^{-/-}, *Cfap44*^{-/-} and WT, respectively), contrary to homozygous KO males, which exhibited complete infertility when mated with WT females (**Fig. 6a**). Sperm concentrations obtained from the two epididymes were 11.4±0.9, 11.2±0.4 and 14.4±3.1 10⁶ sperm/ml (mean ± SD), for *Cfap43*^{-/-}, *Cfap44*^{-/-} and wild-type (WT), respectively, and fall

within the normal values for mouse. We next studied sperm morphology. For *Cfap43*^{-/-} males, 100% of sperm showed a typical human MMAF phenotype with short, thick and coiled flagella (**Fig. 6b1**). Although slightly deformed, sperm head exhibited an overall normal hooked form (**Fig. 6c2**). In contrast, sperm from *Cfap44*^{-/-} males had normal flagellum length but most of them showed abnormal forms and irregular caliber of the midpiece (**Fig. 6b2c3**). Interestingly we observed that in heterozygous animals, *Cfap43* mice had an impaired sperm motility and *Cfap44* animals an increased proportion or morphologically abnormal spermatozoa (**Fig. 6e**).

To better characterize this irregular caliber, sperm were stained with an antibody against Mpc11, a sperm mitochondrial protein²⁵ (**Fig. 6d1-d6**). In sperm from *Cfap44*^{-/-} mice, Mpc11 staining was discontinuous or punctiform, strongly suggesting that the distribution of the mitochondria surrounding the axoneme was irregular (**Fig. 6d2**). For both *Cfap43*^{-/-} and *Cfap44*^{-/-} sperm, the observed structural flagellum defects were associated with severe motility deficiencies (**Fig. 6e**) with sperm from *Cfap43*^{-/-} males showing complete immobility whereas those from *Cfap44*^{-/-} presented tiny vibration only (**Supplementary Video 1**), in contrast to sperm from heterozygotes *Cfap44*^{+/-} (**Supplementary Video 2**).

To better characterize the molecular defects induced by the lack of *Cfap43* and *44* in mouse sperm, we studied by immunofluorescence the presence and localization of several proteins belonging to different sub-structures of the axoneme in sperm from both KO mouse models. The presence of the following proteins was investigated: Dnah5 and Dnali1 as markers of dynein arms, Rspn1 and Rspn4a as markers of radial spokes, Gas8 as a marker of nexin links and Spef2 as a marker of the central pair (**Supplementary Table 2-Mouse**). In *Cfap43*^{-/-} mutated animals Spef2, Rspn1 and Rspn4a were clearly missing (**Fig. 7a,b, Supplementary Table 2 and Supplementary Fig. 6**), indicating that the center of the axoneme, including the head of the radial spoke interacting with the central pair, was absent. In contrast no obvious

defects were observed in sperm from *Cfap44*^{-/-}, suggesting that all substructures are present (**Supplementary Table 2-Mouse**).

To better visualize the impact of the absence of these proteins on the flagellum ultrastructure, sperm from *Cfap43*^{-/-} and *Cfap44*^{-/-} males were analyzed by TEM. The organization of WT mouse flagellum has been described in detail²⁶⁻²⁸: In the midpiece, a 9+2 axoneme is surrounded by 9 outer dense fibers ODFs and the mitochondria sheath. The ODFs 1, 5, 6 and 9 are parallel and aligned with the central pair (**Fig. 7c1**, see also **Fig. 8a1**). In the principal piece, the axoneme is surrounded by 7 ODFs (3 and 8 are missing) and by a fibrous sheath containing 2 longitudinal columns. In addition, a transversal complex is composed of the central pair, the DMT 3 and 8. This complex includes ODFs 3 and 8 in the midpiece and is closely associated with the fibrous sheath in the principle piece, through two stalks emerging from the longitudinal column and linking DMTs 3 and 8²⁹. Contrary to what is usually observed in human sperm, this organization is highly reproducible and less than 5% of WT sections from mouse sperm flagellum present structural defects (**Supplementary Fig. 7a**).

Analyses of longitudinal sections from *Cfap43*^{-/-} sperm show that the observed short tail actually corresponds to a large cytoplasmic mass containing the different structural component of the flagellum, yet unorganized (**Supplementary Fig. 8**). When the axoneme was present, observations of transversal sections of *Cfap43*^{-/-} sperm revealed a deep disorganization of its structure with uneven distribution of the 9 DMTs and the absence of the CPC (**Fig.7c2**). Interestingly we did not observe cross-sections with a normal axoneme organization. In contrast, for *Cfap44*^{-/-} sperm, the 9+2 organization of the axoneme was preserved in 70% of cases compared with 95% in WT (**Fig. 8a2, c2-c3, d2-d5**). In the remaining 30%, several defects were observed including absence of peripheral doublets, external shift or mislocalization of the central pair and distorted circular distribution of the DMTs with CPC misorientation (**Fig. 8b1-b3**). *Cfap44*^{-/-} sperm also exhibited important defects of the ODFs and

of the fibrous sheath (**Supplementary Fig. 7b**). In the midpiece, the number of dense fibers is significantly higher and their localization and orientation are defective (**Fig. 8a2,a3**). In the principal piece, the ODFs 3 and 8 are abnormally positioned, preventing the normal anchoring of the stalks of the fibrous sheath on DMTs 3 and 8 (**Fig. 8c2,c3**). Longitudinal columns also present several defects: they are misaligned and not facing the 3-central-8 complex, leading to notable asymmetry of the structure (**Fig. 8 d2,d3**), which is reinforced in numerous sperm by the presence of a third longitudinal column (**Fig. 8 d4,d5**). Finally, the defects detected in the mitochondrial sheath by optic and fluorescence microscopy were specified by TEM and analyses of the midpieces of *Cfap44^{-/-}* sperm showed both uneven distribution and mitochondria fragmentation (**Supplementary Fig. 9**).

DISCUSSION

Original two-models strategy confirmed that CFAP43 and CFAP44 mutations are responsible for male infertility and MMAF phenotype.

Identification of disease-causing genes has recently been catalyzed by high-throughput genome-wide technologies. New challenges are now arising and in particular the validation of the involvement of candidate genes in the investigated phenotypes. This can be done by selecting the best cellular or animal models. In human diseases or syndromes linked to defects of motile cilia or flagellum five models are commonly used: protozoa *Trypanosoma* and *Tetrahymena*, the green alga *Chlamydomonas reinhardtii*, the sea urchin and mouse^{30,31}. Such a wide selection of distant models is possible because motile cilia or flagella are built on a canonical 9 + 2 axoneme which forms a highly organized protein network remarkably conserved during evolution³². In our study, we first used *Trypanosoma* mutants to confirm the protein localization in the axoneme using endogenous myc-tagged protein. This model enabled us to overcome the lack of reliable antibodies against CFAP43 and CFAP44. RNAi experiments

in our *Trypanosoma* model clearly demonstrated that the inactivation of *TbCFAP43* or *TbCFAP44* led to a severe axonemal disorganization. These results confirmed that these two proteins are essential to maintain axonemal integrity and therefore correct structure and function of the flagellum. The Trypanosoma model presents however some limits. *T. brucei* flagellum possesses a prominent structure called the paraflagellar rod (PFR), running along the entire axoneme and necessary for its motility, a structure which does not exist in human³³. This extra structure possibly modulating the impact of the absence of the studied proteins on the axonemal organization observed in *T. brucei* mutants. To reinforce the data obtained in Trypanosoma, we also produced two knockout mice lines using CRISPR/Cas9 technology^{34,35}. *Cfap43*^{-/-} male were infertile and 100% of their spermatozoa were immotile and morphologically abnormal with short and coiled tail, a phenotype very similar to human MMAF (**Fig. 6c2**). *Cfap44*^{-/-} male mice were also infertile due to flagellar immobility, yet presenting much subtler flagellar defects than patients with *CFAP44* mutations (**Fig. 6c3**) as the main morphological defect consist in the irregular caliber of the sperm flagellum. Such a phenotypic discrepancy between mouse and human is not uncommon. For example, *Dnah1* KO mice display asthenozoospermia without morphological defects of the flagellum³⁶ whereas in humans *DNAH1* mutations result in a MMAF phenotype¹³. Nevertheless, TEM analysis of *Cfap44*^{-/-} sperm indicated that both in human and in mouse CFAP43 protein is required for axonemal organization as we observed ultrastructural defects similar to those found in *CFAP44* patients such as the absence or mislocalization of the central pair complex, missing peripheral doublets and anomalies of the outer dense fibers and of the fibrous sheath (**Fig. 8**). The reason why this axonemal disorganization does not impact on the overall flagellum morphology is not known; however it is very unlikely that a functional truncated protein persists since the insertion of 7 bp in the 5'part of the mRNA (exon 3) is expected to produce a 49 amino-acid truncated protein compared to the 1854 amino-acid for the full protein (**Supplementary Fig. 5**). Overall, the use of two evolutionary distant animal models provides a robust confirmation of the involvement of CFAP43 and CFAP44 proteins in axoneme

formation and function; therefore clearly demonstrating that the truncating mutations identified in humans are responsible for the MMAF phenotype and the resulting primary male infertility.

CFAP43 and CFAP43 are two key axonemal proteins critical for radial spoke – central pair complex (RS-CPC) formation and stability.

The two genes identified in this work, *CFAP43* and *CFAP44*, encode proteins belonging to the WDR protein family. WDR proteins are composed of four or more tryptophan (W), aspartic acid (D) repeats (R), with generally a 40 amino acids-long core domains starting with a glycine-histidine (GH) dipeptide and ending with the WD repeats at the C-terminus. The sequence of these repeats is well conserved across all species in eukaryotes. The assembly of 6 to 7 WD-repeats forms a propeller-like structure serving as a platform for protein-protein interactions, macromolecular assembly but also RNA-binding^{17,37,38}. WDR proteins are involved in a wide range of cellular functions including signal transduction, RNA synthesis and processing, chromatin and cytoskeletal assembly, vesicular trafficking, cell cycle regulation and programmed cell death³⁷. WDR proteins are also involved in different functions of cilia and flagella such as intraflagellar transport³⁹ or assembly of the dynein arms⁴⁰. They are also reported to be associated with the calmodulin and spoke-associated complex (CSC)⁴¹ and with microtubules⁴². Moreover, several WDRs including WDR34⁴³, WDR35^{44,45}, WDR19⁴⁶ and WDR60⁴⁷ have been described to be involved in ciliopathies. Interestingly, mutations in *SPAG16*, a gene containing a WDR motif and localized in the bridge connecting the two singlet microtubules (C1 and C2) of the CPC, were reported to only impact spermatogenesis in mice^{42,48}. The sperm phenotype is strikingly comparable to the MMAF phenotype reported in our patients with abnormally shaped sperm flagella and missing central pairs, disorganization of outer doublet microtubules and outer dense fibers^{11,48,49}. This previous observation of a WDR protein linked to the CPC and involved in a MMAF-like phenotype echoes our results suggesting that CFAP43 and CFAP44 could be directly or indirectly associated with the CPC. The CPC complex consists of two

microtubules, named C1 and C2, which are structurally and biochemically distinct, surrounded by a central sheath made of projections that are unique for each microtubule⁵⁰. In patients harboring *CFAP43* and *CFAP44* mutations, co-immunostaining experiments in sperm cells evidenced severe defects on the CPC microtubules through the impairment of the SPAG6 staining, a protein located on the C1 singlet^{49,51} (**Fig. 4a and c**). These findings were in concordance with TEM analysis which revealed a high percentage of sections lacking the central pair (**Fig. 3, Supplementary Table 1**). Similar observations were described in the *Cfap43*^{-/-} mouse with abnormal staining of the Spef2 protein and the lack of CPC in TEM (**Fig. 7a,c**). CPC defects were also present but inconstantly found in the *Cfap44*^{-/-} mouse model in TEM (**Fig. 8**). These data are supported by the RNAi experiments in *T. brucei* leading to a disordered axoneme with rotated or displaced central pair and peripheral doublets (**Fig. 5c**). Furthermore, in both human and mouse, when the CPC persisted in the axoneme, the CPC was often shifted, mispositioned and not parallel to the axis of the axoneme defined by both longitudinal columns (**Fig. 3 and 8**). Such abnormal position of the CPC has already been reported in Trypanosoma mutants inactivated for CPC proteins²⁴ supporting the hypothesis that CFAP43 and CFAP44 could play a key role in RS-CPC interactions or organization. The lack of the CPC leading to an abnormal ‘9+0’ configuration of the axoneme is the main defect observed in most cases associated with MMAF and may be the hallmark of the MMAF phenotype^{11,52}. Such a defect was also observed in patients with *DNAH1* mutations^{13–15}. Furthermore, inactivation of different genes encoding CPC-associated proteins (SPAG6, SPAG16, SPEF2, ENO4, TTLL1, PGS1) leads to a MMAF-like phenotype (for review see¹¹). These recurrent observations confirm the hypothesis that CPC plays a major role in maintaining the global flagellum organization throughout spermatogenesis. Moreover, EM analysis of spermatozoa from non-genetically characterized MMAF subjects showed that axonemal defects were frequently associated with abnormalities of the peri-axonemal structures including FS, ODFs and mitochondrial sheath^{53,54}. Such defects were also reported in CFAP43 and CFAP44 mutated individuals from our cohort (**Fig. 3**) as well as in the two respective mice KO models we generated

(**Fig. 7c2** and **8**). This could suggest that CPC defects impact the overall flagellar biogenesis. Mammals have three different radial spokes (RS1, RS2, and RS3) which are T-shaped multi-unit proteins repeated every 96 nm along the axoneme connect the peripheral doublet microtubules to the CPC through its head. They very likely have a structural role while also contributing to the regulation of the dynein motors, and thus to flagellar beating⁵⁵. Interestingly, staining of RS head proteins such as RSPH1 was significantly altered in immunostaining experiments in spermatozoa from CFAP43 and CFAP44 mutated subjects (**Fig. 4b** and **d**) and in sperm from *Cfap43*^{-/-} mice (**Fig. 7b** and **Supplementary Fig. 6**). No RS head abnormalities were however observed by IF in *Cfap44*^{-/-} mouse spermatozoa but this is consistent with the higher frequency of normal 9+2 axoneme in this KO model (**Fig .8**). Interestingly, we observed a diffuse and atypical immunostaining of RSPH1 and of SPAG6 in spermatozoa from CFAP43 (**Fig. 4b**) and CFAP44 patients respectively (**Fig. 4c**). This suggests that these proteins are still present in the axoneme but are totally disorganized and/or unable to maintain their correct localization. Conversely, SPAG6 immunostaining is absent in the flagellum of spermatozoa from CFAP43 patients (**Fig. 4a**) and RSPH1 is strongly reduced in CFAP44 patients (**Fig. 4d**) demonstrating that these two proteins were not correctly addressed to the axoneme and may be eliminated during the early stages of spermiogenesis. CFAP43 and CFAP44 could thus be directly or indirectly involved in the formation and the stability of the radial-spokes/central-pair complex (RS-CPC) but may have different function and localization. Altogether, these data strongly suggest that CFAP43 and CFAP44 are essential for axonemal organization and that mutations in these genes severely impairs flagellum elongation resulting in MMAF phenotype and primary male infertility. The precise axonemal localization and function of CFAP43 and CFAP44, together with the possible interaction and functional cooperation between both proteins remain to be determine.

Whole-exome sequencing (WES) revealed that *CFAP43*, *CFAP44* and *DNAH1* are the main genes involved in MMAF phenotypes.

Our work illustrates the efficiency of the combination of WES with original workflow for the validation of the candidate genes that are identified in male infertility due to a MMAF phenotype. Our strategy allowed to identify two new genes responsible for MMAF syndrome. Novel mutations in *DNAH1* were also identified, reinforcing the importance of this gene in MMAF syndrome. The prevalence of mutations in *DNAH1* is however lower compared to our two previous studies^{13,14}. This may be explained by a wider geographic recruitment of patients in the current study compared to our previous work in which a founder effect may have led to an over-estimation of the prevalence of *DNAH1* mutations^{13,14}. This may be explained by a wider geographic recruitment of patients in the current study compared to our previous work in which a founder effect may have led to an increased prevalence of *DNAH1* mutations^{13,14}. Altogether, *DNAH1*, *CFAP43* and *CFAP44* mutations were identified in 28.2% of the analyzed subjects (n=78) originating from North Africa, Europe and Middle East. These results underlines the global importance of these 3 genes in the MMAF syndrome and will improve the genetic diagnosis efficiency of infertile MMAF patients.

Regarding phenotype genotype correlation, semen parameters analysis in patients carrying mutations in *DNAH1*, *CFAP43* and *CFAP44* did not show any significant differences between the three groups. A higher rate of spermatozoa with short and absent flagella and a lower motility rate were however evidenced in the mutated subjects compared to patients with no identified mutation. This suggests that mutations in these genes induces the most severe form of MMAF. Mutations in these three genes should therefore be first investigated in the severe form of MMAF syndrome with >50% of short flagella. The identified mutations are distributed throughout the whole *CFAP43* and *44* genes strongly suggesting that the entirety of both encoded proteins is necessary for preserving their functionality. Interestingly among *CFAP43* mutated patients, only one (P43-9) carried a missense variant (Val347Ala) and presents a milder phenotype compared to the others with homozygous truncating mutations.

This patient has the lowest percentage of sperm with short flagella (22%) with 10% of morphologically normal spermatozoa (versus 0% for the other *CFAP43* mutated subjects). These parameters suggest that the Val347Ala mutated protein is still present in the flagellum axoneme and thus may contribute to maintain a normal organization in a residual part of spermatozoa. This is coherent with the modification of a single amino acid in the N-part of the protein which could only mildly impair protein activity. All spermatozoa of this patient were however totally immotile (0%) indicating that this variant could abrogate the function of the protein resulting in complete flagellum immobility. This data supports the existence of a genotype-phenotype correlation as had been demonstrated for *DNAH1* mutated subjects^{13,14}.

Several questions remain concerning the prognosis of ICSI using sperm cells from *CFAP43* and *CFAP44* patients. We have previously demonstrated that MMAF patients with *DNAH1* mutations had a good prognosis using ICSI⁵⁶ but it remains difficult to predict the success rate for the other MMAF genotypes. Additional correlation studies should now be performed to take into account the individual genotypes in the counseling of MMAF patients.

Finally, these results indicate that we should now reanalyse the WES data from the remaining 56 undiagnosed individuals from the cohort. Mutations in other MMAF genes may have a lower prevalence than mutations in *CFAP43*, *CFAP44* and *DNAH1* genes and inclusion of additional MMAF subjects may be necessary to identify the other less frequent MMAF causing genetic alterations.

Exome sequencing of MMAF patients permitted a diagnosis efficiency of 27% confirming its interest as a diagnostic tool for infertility. Apart from infertility, studied patients did not present any obvious ciliopathies associated symptoms, our results thus contribute to characterize the molecular differences between cilia and flagella. Last, analysis of KO mice showed a small deterioration of sperm motility and morphology suggesting that heterozygous mutations in key spermatogenesis genes, might, alone or more likely through cumulative

genetic and/or environmental factors, contribute to the less severe but much more frequent phenotype of mild to intermediate oligoasthenozoospermia.

MATERIALS AND METHOD

Subjects and controls

We included 78 subjects presenting with asthenozoospermia due to a combination of morphological defects of the sperm flagella including: absent, short, bent, coiled flagella and of irregular width without any of the additional symptoms associated with primacy ciliary dyskinesia (PCD). The morphology of patients' sperm was assessed with Papanicolaou staining. Small variations in protocol might occur between the different laboratories. Subjects were recruited on the basis of the identification of >5% of at least three of the aforementioned flagellar morphological abnormalities (absent, short, coiled, bent and irregular flagella).

The global average of all semen parameters are presented in Table 1 and were compared between the different genotype groups using a two-tailed *t*-test. Forty eight patients are of North African origin who consulted for primary infertility at the Clinique des Jasmin in Tunis. Ten individuals originated from the Middle East (Iranians) and were treated in Tehran at the Royan Institute, (Reproductive Biomedicine Research Center) for primary infertility and eleven patients were recruited in France: 10 at the Cochin Institute and one in Lille.

All patients were recruited between 2008 and 2016. All subjects had normal somatic karyotypes. Approximately half of the patients declared to be born from related parents. Sperm analysis was carried out in the source laboratories during the course of the routine biological examination of the patient, according to World Health Organization (WHO) guidelines (World Health Organization, 2010). Saliva and/or peripheral blood was obtained for all participants. During their medical consultation, all subjects answered a health questionnaire focused on

primary ciliary dyskinesia (PCD) manifestations for infertility. Informed consent was obtained from all the subjects participating in the study according to the local protocols and the principles of the Declaration of Helsinki. In addition, the study was approved by local ethic committees. The samples were then stored in the CRB Germetheque (certification under ISO-9001 and NF-S 96-900) following a standardized procedure. Controls from fertile individuals with normal spermograms were obtained from CRB Germetheque. Consent for CRB storage was approved by the CPCP Sud-Ouest of Toulouse (coordination of the multi sites CRB Germetheque).

Whole-Exome Sequencing and bioinformatics analysis

Genomic DNA was isolated from saliva using Oragen DNA extraction kit (DNAgenotech®, Ottawa, Canada). Coding regions and intron/exon boundaries were enriched using the “all Exon V5 kit” (Agilent Technologies, Wokingham, UK). DNA sequencing was undertaken at the Genoscope, Evry, France, on the HiSeq 2000 from Illumina®. Sequence reads were aligned to the reference genome (hg19) using MAGIC⁵⁷. MAGIC produces quality-adjusted variant and reference read counts on each strand at every covered position in the genome. Duplicate reads and reads that mapped to multiple locations in the genome were excluded from further analysis. Positions whose sequence coverage was below 10 on either the forward or reverse strand were marked as low confidence, and positions whose coverage was below 10 on both strands were excluded. Single nucleotide variations (SNV) and small insertions/deletions (indels) were identified and quality-filtered using in-house scripts. Briefly, for each variant, independent calls are made on each strand, and only positions where both calls agree are retained. The most promising candidate variants were identified using an in-house bioinformatics pipeline, as follows. Variants with a minor allele frequency greater than 5% in the NHLBI ESP6500 [Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA] or in 1000 Genomes Project phase 1 datasets⁵⁸, or greater than 1% in ExAC⁵⁹,

were discarded. We also compared these variants to an in-house database of 94 control exomes obtained from subjects mainly originated from North Africa and Middle East corresponding to the geographical origin of most patients from this study and which is under-represented in SNP public databases. All variants present in homozygous state in this database were excluded. We used Variant Effect Predictor (VEP version 81⁶⁰) to predict the impact of the selected variants. We only retained variants impacting splice donor / acceptor or causing frameshift, inframe insertions / deletions, stop gain, stop loss or missense variants except those scored as "tolerated" by SIFT⁶¹ (sift.jcvi.org) and as "benign" by Polyphen-2⁶² (genetics.bwh.harvard.edu/pph2). Finally, identified mutations were validated by Sanger sequencing. PCR primers and protocols used for each patient are listed in the **Supplementary Table 3**. Sequencing reactions were carried out with BigDye Terminator v3.1 (Applied Biosystems). Sequences analysis were carried out on ABI 3130XL (Applied Biosystems). Sequences were analysed using SeqScape software (Applied Biosystems).

Quantitative real-time RT-PCR analysis

RT-qPCR was performed with cDNAs from various tissues of human and mouse including testes. A panel of 8 organs was used for mouse experiments: testis, brain, lung, kidney, liver, stomach, colon and heart. RNA extraction were performed from three DBA-C57 wild-type mice with the mirVana™ PARIS™ Kit (LifeTechnologies®). For human experiments, a panel of the 3 main ciliated tissues was used: testis, brain and lung. Human RNAs were purchased from Life Technologies®. Reverse-transcriptions were performed using the High Capacity cDNA Reverse Transcription Kit (Applied Biosystem®). Each sample was assayed in triplicate for each gene on a StepOnePlus (LifeTechnologies®), with Power SYBR®Green PCR Master Mix (Life Technologies®). The PCR cycle was as follows: 10 min at 95°C, 1 cycle for enzyme activation; 15 s at 95°C, 60 s at 60°C with fluorescence acquisition,

40 cycles for the PCR. RT-qPCR data were normalized using the reference housekeeping gene *ACTB* for human and mouse with the $-\Delta\Delta Ct$ method⁶³. The $2-\Delta\Delta Ct$ value was set at 0 in brain cells, resulting in an arbitrary expression of 1. Primers sequences and RT-qPCR conditions are indicated in the **Supplementary Table 4**. The efficacy of primers was checked using a standard curve. Melting curve analysis was used to confirm the presence of a single PCR product. Statistics were performed using a two-tailed *t*-test on Prism 4.0 software (GraphPad, San Diego, CA) to compare the relative expression of *CFAP43* and *CFAP44* transcripts in several organs. Statistical tests with a two-tailed P values ≤ 0.05 were considered significant.

Immunostaining in human and mouse sperm cells

We performed immunofluorescence (IF) staining on human and mouse spermatozoa as described by our laboratory⁶⁴. In human, immunostaining could be carried out on samples from two of the patients with a stop-gain mutation, one mutated in each gene. IF experiments were performed using sperm cells from control individuals, from the patient P₄₃-5 homozygous for the c.2658C>T variant in *CFAP43* and from the patient P₄₄-2 homozygous for the c.3175C>T variant in *CFAP44*. Sperm cells were fixed in PBS/4% paraformaldehyde for 1 min at room temperature. After washing in 1 ml PBS, the sperm suspension was spotted onto 0.1% poly L-lysine pre-coated slides (Thermo Scientific). After attachment, sperm were permeabilized with 0.1% (v/v) Triton X-100 –DPBS (Triton X-100; Sigma-Aldrich) for 5 min at RT. Slides were then blocked in 5% corresponding normal serum–DPBS (normal goat or donkey serum; GIBCO, Invitrogen) and incubated overnight at 4°C with primary antibodies. For human experiments, the following primary antibodies were used: DNAI2, DNALI1, RSPH1, RPSH4A, SPAG6, GAS8, AKAP4 and anti-acetylated- α -tubulin. For mouse experiments: Rspf1, Rspf4a, Gas8, Spef2, Dnali1, Dnah5, Mpc11 anti-acetylated- α -tubulin. Primary antibodies references, provider, species, and dilutions used are listed in the **Supplementary Table 5**.

Washes were performed with 0.1% (v/v) Tween 20–DPBS, followed by 1 h incubation at room temperature with secondary antibodies. Highly cross-adsorbed secondary antibodies (Dylight 488 and Dylight 549) were from Jackson Immunoresearch®. Appropriate controls were performed, omitting the primary antibodies. Samples were counterstained with 5 mg/ml Hoechst 33342 (Sigma-Aldrich) and mounted with DAKO mounting media (Life Technology). Fluorescence images were captured with a confocal microscope (Zeiss LSM 710).

Electron microscopy of human and mouse sperm cells

We performed transmission electron microscopy (TEM) experiments on human and mouse (*Cfap43^{-/-}* and *Cfap44^{-/-}* and wild-type) spermatozoa. In human, TEM experiments were performed using sperm cells from fertile control individuals, from the patient P₄₃-8 homozygous for the c.3352C>T variant in *CFAP43* and from the patient P₄₄-3 homozygous for the c.2818dupG variant in *CFAP44*. Sperm cells were fixed with 2.5% glutaraldehyde in 0.1 M sodium phosphate (pH 7.4) during 2 hours at room temperature. Cells were washed with buffer and post fixed with 1% osmium tetroxide in the same buffer for 1 hour at 4°C. After washing with distilled water, cells were stained overnight at 4°C with 0.5% uranyl acetate pH 4.0. Cells were dehydrated through graded alcohol (30%, 60%, 90%, 100%, 100%, and 100%; 10 minutes for each bath) and infiltrated with a mix of 1:1 Epon/alcohol 100% for 2 hours before 2 baths of fresh Epon for 2 hours. Finally, cells were included in fresh Epon and polymerized during 2 days at 60°C. Ultrathin sections of the cell pellet were done with an ultramicrotome (Leica). Sections were post-stained 10 minutes with 5% uranyl acetate pH 4.0, washed four times with distilled water (1 hour) and then stained with 0.4% lead citrate before being observed in an electron microscope at 80kV (JEOL 1200EX).

Trypanosoma brucei

Cultures and transfection

The trypanosome cell line used in this study derived from the blood stream form parental *T. brucei* 427 90-13 strain co-expressing the T7 RNA polymerase and tetracycline repressor⁶⁵. Cells were cultured at 37 °C and 5% CO₂ and transfected and cloned as described in ⁶⁶ in medium supplemented with puromycin (0.1 µg/ml) for constitutive expression of myc-tagged proteins, and with phleomycin (2.5 µg/ml) for RNA interference. RNAi interference was induced with tetracycline (10 µg.ml).

Cloning

For wild-type level of expression of 3xmyc C-terminal tagged proteins, parental cells were transfected as in ⁶⁶ with a tagging cassette that was obtained by PCR using a set of long primers containing 80 nucleotides from the 5'- and 3'-UTRs flanking regions of the *TbCFAP44* and *TbCFAP43* ORF and pMOTag23M vector as template⁶⁷ (**Supplementary Table 6**). For RNA interference, fragments of *TbCFAP44* ORF (bp 1880 to 2309) and *TbCFAP43* ORF (bp 131 to 596) were cloned into p2T7tiB⁶⁸ (**Supplementary Table 6**).

RNA expression analysis (RT-PCR)

Total RNA was isolated from 10⁸ cells of parental, non-induced, and tetracycline induced RNAi cells using the TRIzol reagent according to the manufacturer's instructions (Life Technologies). The constitutively expressed housekeeper Telomerase reverse transcriptase (TERT) was used as an internal control⁶⁹. RT-PCR was carried out with a SuperScript III one-step RT-PCR system with Platinum *Taq* high-fidelity polymerase (Life Technologies Ltd, UK) following the manufacturer's protocol. Briefly, 100 ng of total RNA were mixed with primers and reverse transcriptase solution in 50 ul total volume and using the following cycle protocol: 30 min at 55°C (reverse transcription); 2 min at 94°C (inactivation of reverse transcriptase and activation *Taq* polymerase); followed by 25 cycles of regular PCR (denaturation: 94°C for 15 s; annealing: 55°C for 30s; extension: 68°C for 40 sec); and finalized with a hold for 5 min at

68°C. The RT-PCR products were resolved on 1% agarose running gel with BET and visualized by UV light.

Immunofluorescence

Cells were collected, washed, and processed for immunolabelling on cytoskeletons (detergent extracted cells) as in ⁶⁸, except that the cytoskeletons were fixed in methanol at -20°C for 30 min. Samples were incubated with the primary antibodies anti-myc and rabbit anti-PFR2 (PFR2 is a protein of the para-axonemal structure called paraflagellar rod, 1: 2000 dilution in PBS) and with secondary antibodies anti-mouse FITC conjugated (Sigma F-2012, 1: 400 dilution in PBS) and anti-rabbit Alexa594 conjugated (Thermofischer A11012, 1: 400 dilution in PBS). Nuclei and kinetoplasts were stained with DAPI (10 µg/ml). Images were acquired on a Zeiss Imager Z1 microscope, using a Photometrics Coolsnap HQ2 camera, with Zeiss 100X or 63X objectives (NA 1.4) using Metamorph software (Molecular Devices), and processed with ImageJ. Primary antibodies references, provider, species, and dilutions used are listed in the **Supplementary Table 5**.

Electron microscopy

Cells were fixed in culture medium, by the addition of glutaraldehyde to a final concentration of 2.5%, for 60 min. They were pelleted (1,000 g 10 min), resuspended in fixation buffer (2.5 % glutaraldehyde, 2% PFA, 100 mM Phosphate buffer pH 7.4, 50 mM sucrose for 2 h). Fixed cells were washed in water for 10 min, post-fixed in 1% OsO₄ for 1 h, washed 3 times in water, then samples were stained in 2% uranyl acetate in water at 4°C overnight. Samples were next washed in water, dehydrated in ethanol, embedded in Spurr resin and polymerised overnight at 60°C. Sections were stained in 2% uranyl then lead citrate, and visualized on a TECNAI 12 TEM.

CRISPR/Cas9 KO mice

All animal procedures were run according to the French guidelines on the use of animals in scientific investigations with the approval of the local Ethical Committee (ComEth Grenoble N° 318, ministry agreement number #7128 UHTA-U1209-CA) (. Generation of CRISPR/Cas9 KO was done in collaboration with the University of Geneve (UNIGE.. All the procedures were done in Geneva until the birth of the modified litters. For each gene, three plasmids were injected directly into the nucleus of the zygotes. All plasmids had a U6 promoter which allowed the transcription of the inserted cDNA sequences. One plasmid expressed the Cas9 protein and the other two expressed two distinct RNA guides (single guide RNA, or sgRNA) targeting exons 2 and 21 of the *Cfap43* gene and exons 3 and 15 of the *Cfap44* gene. All plasmids (pGS-U6-sgRNA expression vector) were ordered from GeneScript (Piscataway, NJ, USA) with the different cDNA sequences already inserted. sgRNA sequences are indicated in the **Supplementary Table 7**. The Cas9 nuclease and sgRNAs were introduced into fertilized oocytes. Microinjected oocytes were introduced into pseudopregnant host females and carried to term. Edited founders were identified by Sanger sequencing from tail biopsies. Tail biopsies (2 mm in length) were digested in 200 µl lysis Direct PCR Lysis Reagent (Tail) (Viagen Biotech inc, CA, USA) and 0.2mg of proteinase K for 12–15 h at 55°C and 1 h at 85°C. The DNA was directly used for PCRs. The two targeted exons were amplified using the following PCR protocol: 59°C x 1 cycle, 58°C x 1 cycle and 57°C x 35 cycles with 1 min elongation. SANGER sequencing was then performed to identify CRISP/Cas9 induced mutations and genotypes were determined according to the sequence electropherograms. Mice carrying desired modification events are bred with C57BL6/J to ensure germline transmission and eliminate any possible mosaicism. From the second generation of mice, genotyping was performed by High Resolution Melting (HRM) using MeltDoctor™ HRM Master Mix with the following parameters: 10 min at 95°C for enzyme activation, 15 sec at 95°C for denaturation, 1 min at 60°C for annealing and

extension, then for the melting curve, 1 min at 95°C for denaturation, 1 min at 60°C for annealing, 15 sec at 95°C for high resolution melt and 15 sec at 60°C for annealing. List of primers used for mice genotyping with both methods is available in the **Supplementary Table 8**.

Heterozygous animals were mated to generate homozygous offspring. Approximately 25% of the offspring were homozygous for the mutated allele, indicating the absence of an increased embryonic or postnatal lethality in *Cfap43^{-/-}* and *Cfap44^{-/-}* mice. Mice were housed with unlimited access to food and water and were sacrificed by cervical dislocation after 8 weeks old, which means that they were pubescent and that their reproductive organs were fully established.

The edited gene expression in mutant mice was validated by RT-PCR followed by Sanger sequencing of testicular transcripts. RT-PCR experiments were performed using testis RNA from wild-type, heterozygous and homozygous animals. Reverse transcription was carried out with 5 µl of extracted RNA (~500 ng) using Macherey Nagel NucleoSpin® RNA II columns (Macherey Nagel, Hoerdt, France) according to the manufacturer's protocol. Hybridization of the oligo dT was performed by incubating for 5 min at 65°C and quenching on ice with the following mix : 5 µl RNA, 3 µl of poly T oligo primers (dT) 12-18 (10 mM, Pharmacia), 3 µl of the four dNTPs (0.5 mM, Roche diagnostics) and 2.2 µl of H₂O. Reverse transcription then was carried out for 30 min at 55° after the addition of 4 µl of 5X buffer, 0.5 µl RNase inhibitor and 0.5 µl of Transcriptor reverse transcriptase (Roche Diagnostics). Two microliters of the obtained cDNA mix was used for the subsequent PCR and Sanger sequencing. Primers sequences and RT-PCR conditions are indicated in the **Supplementary Table 9**.

Phenotypic analysis of mutant

To test the fertility, pubescent *Cfap43*^{-/-} and *Cfap44*^{-/-} males (8-wk-old) were mated with C57BL6/J females. To determine sperm concentration, sperm samples were collected from the cauda epididymis of 8-wk-old *Cfap43*^{-/-} and *Cfap44*^{-/-} and sperm number was determined using a hemocytometer under a light microscope. For sperm morphology, sperm was displayed over a slide, dried at room temperature, and then fixed in 75% ethanol. Harris-Schorr staining was performed according to the WHO protocol. Schorr staining solution was obtained from Merck and least 100 sperm per animal were analyzed. Mobility of sperm was assessed with computer-assisted motility analysis (CEROS I, Hamilton Thorn Research, Beverly, MA) in an analysis chamber (100 µm depth, 30 µl volume, Leja Products B.V., Netherlands) at 37 °C. The settings employed for analysis were as follows: acquisition rate, 60 Hz; number of frames, 100; minimum contrast, 25; minimum cell size, 10; low static size gate, 2.4; high static size gate, 2.4; low static intensity gate, 1.02; high static intensity gate, 1.37; minimum elongation gate, 12; maximum elongation gate, 100; magnification factor, 0.70. The motility parameters measured were curvilinear velocity (VCL), average path velocity (VAP), and straight-line velocity (VSL). At least 100 motile sperm were analyzed for each assay. Motile sperm and progressive sperm were characterized by VAP >1 µm/s, by average path velocity >30 µm/s and straightness (VSL/VAP) >70%, respectively. For each genotype, 5 mice were used.

Statistical analyses were performed with SigmaPlot Software. Unpaired t test was used to compare the different genotypes. Statistical tests with a 2-tailed P values ≤ 0.05 were considered significant.

AKNOWLEDGMENTS

We thank the GIN electron microscopy platform and Anne Bertrand, and the IAB microscopy platform and Alexei Grichine and Jacques Mazzega for their technical help. We thank K. Ersfeld (Bayreuth University) for the anti-myc antibody, N. Biteau (Bordeaux University) for the anti-PFR2 antibody, Vanderperre Benoît and Martinou Jean Claude (UNIGE university, Geneva, CH) for the Mpcl11 antibody. We thank Denise Escalier and Gérard Gascon for discussions and expertise on the topic. We thank our patients and control individuals for their participation. This work was mainly supported by the following grants: The “MAS-Flagella” project financed by French ANR and the DGOS for the program PRTS 2014 and the “Whole genome sequencing of patients with Flagellar Growth Defects (FGD)” financed by the fondation maladies rares (FMR) for the program Séquençage à haut débit 2012.

AUTHORS' ROLES

CC, MB, AT, CA and PFR analyzed the data and wrote the manuscript; Z-EK, AA-Y, LS, AL, PL, EEK, JF and CC performed molecular work; TK, JF-D, AB, NT-M analyzed genetic data; CC, AV, PLT, CW-L, SB performed IF experiments. AV, AS, AT performed EM experiments, Z-EK, AA-Y, AV, BC, SF, JE performed mouse work, DDD, DRR, MB performed Trypanosoma work, SFBM, GM, AD, SH, VM, LH, OM, MM, HL, MK, SH, VS, PSJ, J-PW, HG, ED, RZ provided clinical samples and data; MB, AT, CA and PFR designed the study, supervised all molecular laboratory work, had full access to all of the data in the study and takes responsibility for the integrity of the data and its accuracy. All authors contributed to the report.

CONFLICT OF INTEREST STATEMENT

The authors have declared that no conflict of interest exists.

FIGURE LEGENDS

Figure 1 Morphology of spermatozoa from the patients P₄₃-8 and P₄₄-3. All spermatozoa have a much shorter flagellum than those of controls. Additional features of MMAF spermatozoa are absent (*), thick (&) and rolled flagella (@).

Figure 2 Location of *CFAP43* and *CFAP44* mutations in the intron-exon structure and in the protein representation. **(a)** Mutations identified in the *CFAP43* gene. **(b)** Mutations identified in the *CFAP44* gene. Blue squares stand for WD-repeats domains and green squares for coiled-coiled domains as described by Uniprot server. Mutations are annotated in accordance to the HGVS's recommendations.

Figure 3 Transmission Electron Microscopy (TEM) analyses of sperm cells from CFAP43 and CFAP44 patients reveal that central pair is missing or presents major defects.

(Left panel) Longitudinal sections (scale bar 500 nm) and cross sections (scale bars 100 nm) of sperm flagellum from control. **(Central panel)** Longitudinal sections (scale bars 500 nm) and cross sections (scale bars 100 nm) of sperm flagellum from CFAP43 mutated patient (P₄₃-8). We can notice a short short tail corresponding to a cytoplasmic mass containing the different components of the flagellum, all unorganized. In CFAP43 upper cross section, the CPC is not aligned with DMTs 3 and 8 (red line) and is rotated by 90°. We can observe the absence of central pair of microtubules in other cross sections. **(Right panel)** Ultrastructure of CFAP44 mutated sperm (P₄₄-3) Longitudinal sections (scale bars 500 nm) show similar ultrastructure of short tail (cytoplasmic mass). In CFAP44 upper cross section, the central pair is disassembled and displaced (red arrow). We can observe the absence of central pair of microtubules in other cross sections. Scales bars for cross sections 100 nm.

Figure 4 Immunofluorescence staining in human spermatozoa from controls, CFAP43 and CFAP44 patients reveal that the organization of the axoneme is impaired in patients. **(a1-a3)** sperm cells from a fertile control stained with anti SPAG6 (Green) and anti acetylated tubulin (red) antibodies. DNA was counterstained with Hoechst 33342. **a3** corresponds to **a1-a2** overlay and shows that in control sperm, SPAG6 and tubulin staining superimpose. Scale bars 5 µm. **(b1-b3)** SPAG6 staining is absent in sperm from the patient P₄₃-5 homozygous for the c.2658C>T variant in *CFAP43*. **(c1-c3)** Similar IF experiments performed with sperm cells from the patient P₄₄-2 homozygous for the c.3175C>T variant in *CFAP44*. Scale bar 5 µm. Contrary to the control, the SPAG6 immunostaining (green) is abnormal with a diffuse pattern concentrated in the midpiece of the spermatozoa and is not detectable in the principle piece. **(d1-d3)** Sperm cells from a fertile control stained with anti RSPH1 (Green) and anti-acetylated tubulin (red) antibodies. DNA was counterstained with Hoechst 33342. **d3** corresponds to **d1-d2** overlay and shows that RSPH1 and tubulin staining superimpose in control sperm. Scale bar 5 µm. **(e1-e3)** In sperm from the patient P₄₃-5, the staining of the radial spoke head protein RSPH1 (green) is significantly different from control (**d1**) with a marked diffuse staining. **(f1-f3)** In sperm from the patient P₄₄-2 the intensity of the RSPH1 staining is strongly reduced.

Figure 5 Orthologues of *Trypanosoma brucei* *TbCFAP44* and *TbCFAP43* are flagellar proteins necessary for cell survival. **(a)** *TbCFAP44* and *TbCFAP43* are flagellar proteins. **(a1-a3)** Immunofluorescence on detergent extracted cells of parental *T. brucei* (non expressing myc-

tagged proteins) stained with anti-PFR (red) and anti-myc (green) antibodies. No myc staining was observed on parental cells, indicating the specificity of the anti-myc antibody. **(a4-a6)** Immunofluorescence on detergent extracted cells expressing constitutively myc-tagged *TbCFAP44* and non-induced for *TbCFAP44^{RNAi}*. The flagella were labelled with the anti-PFR (red), and myc-tagged protein with anti-myc (green), (refer to the panel: “Non induced”). Note the green staining of the entire flagellum showing that *TbCFAP44* localisation is restricted to the flagellum. **(a7-a9)** Inducible RNA interference (RNAi) was applied to knockdown *TbCFAP44*. *TbCFAP44_{myc}^{RNAi}*-induced (24H) cells that were constitutively myc-tagged *TbCFAP44* showed no, or weak myc-labelling (green) on the new flagellum (NF) whilst the old flagellum (OF) remained labelled. PFR is labelled in red. Note: cells with maximum 2 flagella were chosen for clear imaging (refer to the panel: “induced”). **(a10-a12)** Similar experiments as performed in **a4-a6** for *TbCFAP43*. **(a13-a15)** Similar RNA interference experiments as performed in **a7-a9** for *TbCFAP43*. Insets are enlargement of images of flagella taken from the main panels and display areas indicated by white arrows (scale bar 1 μ m). The old flagellum (OF) and the new flagellum (NF) are both labelled in non-induced cells. In induced cells, only the old flagellum is labelled with anti-myc (or a very weak signal is observed on the new flagellum). Nuclei and kinetoplasts (mitochondrial genome) are labelled with DAPI (blue). Scale bars represent 5 μ m. **(b)** *TbCFAP44* and *TbCFAP43* are necessary for *T. brucei* growth. Growth curves for parental cells, and *TbCFAP43^{RNAi}* and *TbCFAP44^{RNAi}* cells, non-induced or induced with tetracycline. Cells were counted every 24 h. The graph represents the cumulative number of cells per ml. Error bars represent the standard error from 3 independent experiments. **(c)** Ultrastructure of the flagellum of *TbCFAP44* and *TbCFAP43* RNAi-induced cells. Electron micrographs of stained thin sections of parental cells (**c1-c2**) bearing 1 flagellar pocket (FP) and 1 flagellum (*), and of 24h induced *TbCFAP44^{RNAi}* (**c3-c4**) and *TbCFAP43^{RNAi}* (**c5-c6**) cells. **c4** and **c6** are enlargements of **c3** and **c5** respectively. Scale bars 100 nm in **c2**,

c4, c6; 200 nm in **c1-c3**. Note: in *TbCFAP44^{RNAi}* and *TbCFAP43^{RNAi}* cells, the flagellar pocket is enlarged and bares more than two flagella (**c3, c5**). Some of these flagella present axonemal defects including displaced and rotated CPC and shifted DMTs (**c3** white arrows, **c4, c6**).

Figure 6 Reproductive phenotype of *Cfap43^{-/-}* and *Cfap44^{-/-}* males. **(a)** Fertility of *Cfap43^{+/-}*, *Cfap44^{+/-}* *Cfap43^{-/-}* and *Cfap44^{-/-}* males. Heterozygous and homozygous mutant males were mated with WT females and the numbers of pups per litter were measured. KO males were completely sterile. **(b)** Spermatocytograms showing the number of abnormal sperm in heterozygous and homozygous mutant males. **(c)** Images of typical sperm stained with Harris-Shorr staining from *Cfap43^{-/-}* and *Cfap44^{-/-}* males. Scale bars =10 µm. **(d)** The mitochondria sheath is fragmented in *Cfap44^{-/-}* males. Staining of WT sperm with an anti-MCPC11 (**d1**, green) and anti-acetylated tubulin (**d3**, red) antibodies. **(d5)** Overlay of MCPC11 and tubulin staining. Sperm were counterstained with Hoechst (blue). **(d2, d4, d6)**. Similar experiments on sperm from *Cfap44^{-/-}* males. Scale bars = 10 µm. **(e)** Total and progressive motilities of sperm from *Cfap43^{-/-}* and *Cfap44^{-/-}* males are abnormal. **a, b, e:** Data represent mean ± SD; the statistical difference was assessed with t-test, p value as indicated.

Figure 7 The axoneme of sperm from *Cfap43^{-/-}* males is fully disorganized. **(a)** Central pair is absent in sperm from *Cfap43^{-/-}* males. Staining of WT sperm with an anti-Spef2, a marker of a component of the projection 1b of singlet C1, (**a1**, green) and anti-acetylated-tubulin (**a3**, red) antibodies. **(a5)** Overlay of Spef2 and tubulin staining. Sperm were counterstained with Hoechst (blue). **(a2, a4, a6)** Similar experiments on sperm from *Cfap43^{-/-}* males. Scale bars = 5 µm. **(b)** Head of radial spoke are absent in sperm from *Cfap43^{-/-}* males. Staining of WT sperm with an anti-Rsph4a (**a1**, green) and anti-tubulin (**a3**, red) antibodies. **(a5)** Overlay of Rsph4a and tubulin staining. Sperm were counterstained with Hoechst (blue). **(a2, a4, a6)**. Similar experiments on sperm from *Cfap43^{-/-}* males. Scale bars = 5 µm. **(c)** Transversal section of a

sperm from WT (**c1**) and *Cfap43^{-/-}* (**c2**) males observed by EM. Note the specific arrangement of the ODFs around the axoneme in WT sperm and the complete disorganization of the DMTs and the absence of the central pair in *Cfap43^{-/-}* sperm. Scale bars = 240 nm. DMTs doublet of microtubules, ODF outer dense fiber and Mt Mitochondria.

Figure 8 Analyses of cross-sections of sperm from *Cfap44^{-/-}* males by EM reveal multiple structural axonemal defects. **(a1)** Cross-section of the midpiece of a WT sperm, showing the arrangement of the ODFs around the axoneme. **(a2)** Presence of extra ODFs in midpieces sections of sperm from *Cfap44^{-/-}* males. The orientation of ODFs is also defective, leading to an increase of the midpiece diameter. **a1** and **a2** same scale bars = 250 nm. **(a3)** Graph showing the increased number of ODF in the mutant. Data represent mean ± SD; the statistical difference was assessed with t-test, p value as indicated. **(b1-b3)** Diverse structural defects of the axoneme in sperm from *Cfap44^{-/-}* males. **(b1)** Three DMTs were missing. The central pair is shifted at the periphery (red arrowhead). **(b2)** DMTS 5, 6 and 7 were missing. Note the presence of a third longitudinal column. **(b3)** Irregular distribution of the DMTs associated with a rotation of the central pair (straight red line). Scale bars = 196 nm. **(c1)** In WT sperm, the fibrous sheath is linked to the 3-central-8 complex by stalks emerging from the longitudinal columns (white arrowheads). **(c2, c3)** The presence of extra ODFs facing DMTs 3 and 8 (red arrowheads) prevents a normal anchoring of the fibrous sheath's stalks on DMTs 3 and 8. Scale bars = 270 nm. **(d1-d5)** Longitudinal columns (LC) are not aligned with 3-8 CPC axis. In contrast to WT, where the 3-central-8 complex is aligned with LC to form almost a straight line (**d1**, red line), LC are misaligned in sperm from *Cfap44^{-/-}* males, leading to notable asymmetry of the structure (**d2, d3**, red lines). **(d4, d5)** The presence of a third LC increases asymmetry. Scale bars = 184 nm. **(d6)** Bar graph showing the % of defects observed in the principle piece as described in **b**, **c**, **d**. At least 50 Cross-sections were analyzed in each genotype.

TABLES

Table 1 Average semen parameters in the different genotype groups and for the 78 included MMAF subjects in the present study.

Table 2 *CFAP43* (*WDR96*), *CFAP44* (*WDR52*), and *DNAH1* variants identified by WES for all the analysed subjects (n=78).

SUPPLEMENTARY FIGURES

Supplementary Figure 1

Relative mRNA Expression of human and mouse *CFAP43* and *CFAP44* transcripts.

CFAP43 and *CFAP44* mRNA levels in a panel of human and mouse normal tissues. Results are presented as the mean of triplicates (ratio target gene/ACTB) \pm Standard Deviation (SD). RT-qPCR data were normalized using the reference gene ACTB with the - $\Delta\Delta Ct$ method. Brain expression is arbitrary set to 1. In human and mouse, *CFAP43* (blue columns) and *CFAP44* (red columns) has the strongest expression in testis compared to other organs. Unpaired t-test, ***P< 0.001.

Supplementary Figure 2

Electropherograms of Sanger sequencing for all *CFAP43* and *CFAP44* mutated patients compared to reference sequence.

Supplementary Figure 3

Knockdown of *TbCFAP43* and *TbCFAP44* induces cytokinesis defects and produces quadriflagellates cells.

Parental cells (**a**), and cells induced 24h for the RNAi of *TbCFPA44* (**b**) and *TbCFAP43* (**c**) were detergent extracted and methanol fixed. The flagella (F, asterisk) were immuno-labelled with mAb25, an anti-axonemal protein (green). Parental cells have 2 flagella whereas mutant cells have 4. The nuclei (N, triangle) and the kinetoplasts (K, arrow head) were labelled with DAPI. Scale bars 5 μ m. 4K4F2N means cell with 4 kinetoplasts, 4 flagella and 2 nuclei.

Supplementary Figure 4

Quantification of *TbCFAP43* and *TbCFAP44* RNAi knockdown.

Total RNA was isolated from parental and non-induced cells, and from 6 and 12h RNAi induced cells. mRNA decrease upon RNAi induction was tested by semi-quantitative RT-PCR using primer sets specific to TbCFAP44 (left panel) or TbCFAP43 (right panel) and quantified (lower panel) relative to the telomerase reverse transcriptase (TERT) level.

Supplementary Figure 5

Mutant mice generated by CRISPR/Cas9 technology.

Cfap44 and *Cfap43* KO mice were generated by targeting exon 3 and 21 respectively. Sequencing of transcripts obtained from the testis of mutant mice allowed us to identify a frameshift mutation InsTCAGATA in *Cfap44* and delAAGG in *Cfap43* (Red boxes) leading to a premature stop codon (Black box).

Supplementary Figure 6

Head of radial spoke RSPH1 are absent in sperm from *Cfap43*^{-/-} males.

(a1,a3) Staining of WT sperm with an anti-Rsph1 (a1, green) and anti-tubulin (a3, red) antibodies. (a5) overlay of Rsph1 and tubulin staining. Sperm were counterstained with Hoechst (blue). (a2, a4, a6) Similar experiments on sperm from *Cfap43*^{-/-} males. Scale bars = 5 μ m.

Supplementary Figure 7

Electron microscopy images of the ultrastructure of sperm from WT and *Cfap44*^{-/-} male at low magnification.

(a) Note the uniform structure of the WT sperm. (b) In contrast, in *Cfap44*^{-/-} sperm, shapes and forms are heterogenous. Scale bars = 402 nm.

Supplementary Figure 8

Electron microscopy images of sperm from *Cfap43^{-/-}* males show the absence of organized flagellum in short tail sperm. The short tail rather corresponds to a cytoplasmic mass containing the different components of the flagellum, yet unorganized. scale bar = 2 μ m.

Supplementary Figure 9

Longitudinal sections obtained by electron microscopy of sperm from *Cfap44^{-/-}* males showing that the mitochondria are fragmented (**a**) and irregularly layered (**b**). Left image scale bar = 510 nm and right image scale bar = 395 nm.

SUPPLEMENTARY TABLES

Supplementary Table 1.

Sperm axonemal abnormalities observed by transmission electron microscopy (TEM) in individuals P₄₃-8 carrying mutations in *CFAP43* and individual P₄₄-3 carrying mutations in *CFAP44*.

Supplementary Table 2

Absence of CFAP43 or CFAP44 modulate intensities of staining of different axonemal substructures in mouse and human flagella.

The following antibodies were used to identify sub components of the axoneme: α -RSPH1 and α -RSPH4A for radial spokes, α -GAS8 for N-DRC, α -DNALII for IDA, α -DNAI2 and α -DNAH5 for ODA, α -SPAG6 and α -SPEF2 for CPC and α -AKAP4 for fibrous sheath. ++ intense staining, + intermediate staining, +/- weak signal, - no staining.

Supplementary Table 3

Primers used for *CFAP43* and *CFAP44* patients Sanger sequencing verification

Supplementary Table 4

Primers used for RT-qPCR of *CFAP43* and *CFAP44* in human and mouse.

Supplementary Table 5

List of primary antibodies used in immunofluorescence experiments in human and mouse

Supplementary Table 6

Endogenous tagging and RNAi sequences for *T.brucei* experiments

Supplementary Table 7

Guide RNAs sequences for CRISPR/Cas9 mice generation.

Supplementary Table 8

Primers for CRISPR/Cas9 mice genotyping.

Supplementary Table 9

List of primers used for RT-PCR experiments for *Cfap43* and *Cfap44*.

SUPPLEMENTARY VIDEO

Supplementary Video 1. Video showing normal flagellum beat of sperm from *Cfap44⁺⁻* heterozygotes males

Supplementary Video 2 Video showing the defective flagellum beat of sperm from *Cfap44^{-/-}* males.

REFERENCES

1. Mascarenhas, M. N., Flaxman, S. R., Boerma, T., Vanderpoel, S. & Stevens, G. A. National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 health surveys. *PLoS Med.* **9**, e1001356 (2012).
2. Rolland, M., Le Moal, J., Wagner, V., Royère, D. & De Mouzon, J. Decline in semen concentration and morphology in a sample of 26,609 men close to general population between 1989 and 2005 in France. *Hum. Reprod. Oxf. Engl.* **28**, 462–470 (2013).
3. Carlsen, E., Giwercman, A., Keiding, N. & Skakkebaek, N. E. Evidence for decreasing quality of semen during past 50 years. *BMJ* **305**, 609–613 (1992).
4. Burton, A. Study Suggests Long-Term Decline in French Sperm Quality. *Environ. Health Perspect.* **121**, a46 (2013).
5. Coutton, C., Fissore, R. A., Palermo, G. D., Stouffs, K. & Touré, A. Male Infertility: Genetics, Mechanism, and Therapies. *BioMed Res. Int.* **2016**, 7372362 (2016).
6. Matzuk, M. M. & Lamb, D. J. Genetic dissection of mammalian fertility pathways. *Nat. Cell Biol.* **4 Suppl**, s41-49 (2002).
7. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755 (2011).
8. Escoffier, J. *et al.* Homozygous mutation of PLCZ1 leads to defective human oocyte activation and infertility that is not rescued by the WW-binding protein PAWP. *Hum. Mol. Genet.* **25**, 878–891 (2016).
9. Zhu, F. *et al.* Biallelic SUN5 Mutations Cause Autosomal-Recessive Acephalic Spermatozoa Syndrome. *Am. J. Hum. Genet.* **99**, 942–949 (2016).
10. Ray, P. F. *et al.* Genetic abnormalities leading to qualitative defects of sperm morphology or function. *Clin. Genet.* (2016). doi:10.1111/cge.12905
11. Coutton, C., Escoffier, J., Martinez, G., Arnoult, C. & Ray, P. F. Teratozoospermia: spotlight on the main genetic actors in the human. *Hum. Reprod. Update* **21**, 455–485 (2015).

12. Thonneau, P. *et al.* Incidence and main causes of infertility in a resident population (1,850,000) of three French regions (1988-1989). *Hum. Reprod. Oxf. Engl.* **6**, 811–816 (1991).
13. Ben Khelifa, M. *et al.* Mutations in DNAH1, which encodes an inner arm heavy chain dynein, lead to male infertility from multiple morphological abnormalities of the sperm flagella. *Am. J. Hum. Genet.* **94**, 95–104 (2014).
14. Amiri-Yekta, A. *et al.* Whole-exome sequencing of familial cases of multiple morphological abnormalities of the sperm flagella (MMAF) reveals new DNAH1 mutations. *Hum. Reprod. Oxf. Engl.* **31**, 2872–2880 (2016).
15. Wang, X. *et al.* Homozygous DNAH1 frameshift mutation causes multiple morphological anomalies of the sperm flagella in Chinese. *Clin. Genet.* (2016). doi:10.1111/cge.12857
16. Ivliev, A. E., 't Hoen, P. A. C., van Roon-Mom, W. M. C., Peters, D. J. M. & Sergeeva, M. G. Exploring the transcriptome of ciliated cells using in silico dissection of human tissues. *PLoS One* **7**, e35618 (2012).
17. Smith, T. F. Diversity of WD-repeat proteins. *Subcell. Biochem.* **48**, 20–30 (2008).
18. Smith, T. F. *Diversity of WD-repeat Proteins*. (Landes Bioscience, 2013).
19. Aslett, M. *et al.* TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Res.* **38**, D457-462 (2010).
20. Subota, I. *et al.* Proteomic analysis of intact flagella of procyclic *Trypanosoma brucei* cells identifies novel flagellar proteins with unique sub-localization and dynamics. *Mol. Cell. Proteomics MCPC* **13**, 1769–1786 (2014).
21. Broadhead, R. *et al.* Flagellar motility is required for the viability of the bloodstream trypanosome. *Nature* **440**, 224–227 (2006).
22. Baron, D. M., Ralston, K. S., Kabututu, Z. P. & Hill, K. L. Functional genomics in *Trypanosoma brucei* identifies evolutionarily conserved components of motile flagella. *J. Cell Sci.* **120**, 478–491 (2007).

23. Ralston, K. S., Kisalu, N. K. & Hill, K. L. Structure-function analysis of dynein light chain 1 identifies viable motility mutants in bloodstream-form *Trypanosoma brucei*. *Eukaryot. Cell* **10**, 884–894 (2011).
24. Branche, C. *et al.* Conserved and specific functions of axoneme components in trypanosome motility. *J. Cell Sci.* **119**, 3443–3455 (2006).
25. Vanderperre, B. *et al.* MPC1-like Is a Placental Mammal-specific Mitochondrial Pyruvate Carrier Subunit Expressed in Postmeiotic Male Germ Cells. *J. Biol. Chem.* **291**, 16448–16461 (2016).
26. Fawcett, D. W. The mammalian spermatozoon. *Dev. Biol.* **44**, 394–436 (1975).
27. Fawcett, D. W. A comparative view of sperm ultrastructure. *Biol. Reprod.* **2**, Suppl 2:90-127 (1970).
28. Fawcett, D. W. & Phillips, D. M. The fine structure and development of the neck region of the mammalian spermatozoon. *Anat. Rec.* **165**, 153–164 (1969).
29. Lindemann, C. B., Orlando, A. & Kanous, K. S. The flagellar beat of rat sperm is organized by the interaction of two functionally distinct populations of dynein bridges with a stable central axonemal partition. *J. Cell Sci.* **102** (Pt 2), 249–260 (1992).
30. Vincensini, L., Blisnick, T. & Bastin, P. 1001 model organisms to study cilia and flagella. *Biol. Cell* **103**, 109–130 (2011).
31. Ostrowski, L. E., Dutcher, S. K. & Lo, C. W. Cilia and models for studying structure and function. *Proc. Am. Thorac. Soc.* **8**, 423–429 (2011).
32. Silflow, C. D. & Lefebvre, P. A. Assembly and motility of eukaryotic cilia and flagella. Lessons from *Chlamydomonas reinhardtii*. *Plant Physiol.* **127**, 1500–1507 (2001).
33. Portman, N. & Gull, K. The paraflagellar rod of kinetoplastid parasites: from structure to components and function. *Int. J. Parasitol.* **40**, 135–148 (2010).
34. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).

35. Xiao-Jie, L., Hui-Ying, X., Zun-Ping, K., Jin-Lian, C. & Li-Juan, J. CRISPR-Cas9: a new and promising player in gene therapy. *J. Med. Genet.* **52**, 289–296 (2015).
36. Neesen, J. *et al.* Disruption of an inner arm dynein heavy chain gene results in asthenozoospermia and reduced ciliary beat frequency. *Hum. Mol. Genet.* **10**, 1117–1128 (2001).
37. Li, D. & Roberts, R. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell. Mol. Life Sci. CMLS* **58**, 2085–2097 (2001).
38. Lau, C., Bachorik, J. L. & Dreyfuss, G. Gemin5-snRNA interaction reveals an RNA binding function for WD repeat domains. *Nat. Struct. Mol. Biol.* **16**, 486–491 (2009).
39. Blacque, O. E. *et al.* The WD repeat-containing protein IFTA-1 is required for retrograde intraflagellar transport. *Mol. Biol. Cell* **17**, 5053–5062 (2006).
40. Hendrickson, T. W. *et al.* IC138 is a WD-repeat dynein intermediate chain required for light chain assembly and regulation of flagellar bending. *Mol. Biol. Cell* **15**, 5431–5442 (2004).
41. Dymek, E. E. & Smith, E. F. A conserved CaM- and radial spoke associated complex mediates regulation of flagellar dynein activity. *J. Cell Biol.* **179**, 515–526 (2007).
42. Zhang, Z. *et al.* Haploinsufficiency for the murine orthologue of Chlamydomonas PF20 disrupts spermatogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12946–12951 (2004).
43. Schmidts, M. *et al.* Mutations in the gene encoding IFT dynein complex component WDR34 cause Jeune asphyxiating thoracic dystrophy. *Am. J. Hum. Genet.* **93**, 932–944 (2013).
44. Smith, C. *et al.* A relatively mild skeletal ciliopathy phenotype consistent with cranioectodermal dysplasia is associated with a homozygous nonsynonymous mutation in WDR35. *Am. J. Med. Genet. A* **170**, 760–765 (2016).
45. Yamamura, T. *et al.* Rare renal ciliopathies in non-consanguineous families that were identified by targeted resequencing. *Clin. Exp. Nephrol.* **21**, 136–142 (2017).
46. Fehrenbach, H. *et al.* Mutations in WDR19 encoding the intraflagellar transport component IFT144 cause a broad spectrum of ciliopathies. *Pediatr. Nephrol. Berl. Ger.* **29**, 1451–1456 (2014).

47. McInerney-Leo, A. M. *et al.* Short-rib polydactyly and Jeune syndromes are caused by mutations in WDR60. *Am. J. Hum. Genet.* **93**, 515–523 (2013).
48. Zhang, Z. *et al.* Deficiency of SPAG16L causes male infertility associated with impaired sperm motility. *Biol. Reprod.* **74**, 751–759 (2006).
49. Sapiro, R. *et al.* Male infertility, impaired sperm motility, and hydrocephalus in mice deficient in sperm-associated antigen 6. *Mol. Cell. Biol.* **22**, 6298–6305 (2002).
50. Teves, M. E., Nagarkatti-Gude, D. R., Zhang, Z. & Strauss, J. F. Mammalian axoneme central pair complex proteins: Broader roles revealed by gene knockout phenotypes. *Cytoskelet. Hoboken NJ* **73**, 3–22 (2016).
51. Sapiro, R. *et al.* Sperm antigen 6 is the murine homologue of the Chlamydomonas reinhardtii central apparatus protein encoded by the PF16 locus. *Biol. Reprod.* **62**, 511–518 (2000).
52. Chemes, E. H. & Rawe, Y. V. Sperm pathology: a step beyond descriptive morphology. Origin, characterization and fertility potential of abnormal sperm phenotypes in infertile men. *Hum. Reprod. Update* **9**, 405–428 (2003).
53. Escalier, D. & David, G. Pathology of the cytoskeleton of the human sperm flagellum: axonemal and peri-axonemal anomalies. *Biol. Cell* **50**, 37–52 (1984).
54. Escalier, D. & Serres, C. Aberrant distribution of the peri-axonemal structures in the human spermatozoon: possible role of the axoneme in the spatial organization of the flagellar components. *Biol. Cell* **53**, 239–250 (1985).
55. Pigino, G. & Ishikawa, T. Axonemal radial spokes: 3D structure, function and assembly. *Bioarchitecture* **2**, 50–58 (2012).
56. Wambergue, C. *et al.* Patients with multiple morphological abnormalities of the sperm flagella due to DNAH1 mutations have a good prognosis following intracytoplasmic sperm injection. *Hum. Reprod. Oxf. Engl.* **31**, 1164–1172 (2016).

57. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
58. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
59. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
60. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
61. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
62. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
63. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods San Diego Calif* **25**, 402–408 (2001).
64. Escoffier, J. *et al.* Subcellular localization of phospholipase C ζ in human sperm and its absence in DPY19L2-deficient sperm are consistent with its role in oocyte activation. *Mol. Hum. Reprod.* **21**, 157–168 (2015).
65. Wirtz, E., Leal, S., Ochatt, C. & Cross, G. A. A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **99**, 89–101 (1999).
66. Dauchy, F.-A. *et al.* Trypanosoma brucei CYP51: Essentiality and Targeting Therapy in an Experimental Model. *PLoS Negl. Trop. Dis.* **10**, e0005125 (2016).
67. Oberholzer, M., Morand, S., Kunz, S. & Seebeck, T. A vector series for rapid PCR-mediated C-terminal in situ tagging of Trypanosoma brucei genes. *Mol. Biochem. Parasitol.* **145**, 117–120 (2006).

68. LaCount, D. J., Barrett, B. & Donelson, J. E. Trypanosoma brucei FLA1 is required for flagellum attachment and cytokinesis. *J. Biol. Chem.* **277**, 17580–17588 (2002).
69. Brenndörfer, M. & Boshart, M. Selection of reference genes for mRNA quantification in Trypanosoma brucei. *Mol. Biochem. Parasitol.* **172**, 52–55 (2010).

Table 1

Table 1. Average semen parameters in the different genotype groups and for the 78 included MMAF subjects in the present study.

Semen parameters	MMAF <i>CFAP43+</i> n=10	MMAF <i>CFAP44+</i> n=6	MMAF <i>DNAH1+</i> n=6	MMAF all mutations n=22	MMAF with unknown causes n=56	Overall MMAF n=78
Mean age (years)	37.7 ± 9.6 (n'=9)	41.3 ± 4.3 (n'=6)	41.5 ± 4.4 (n'=6)	39.8 ± 7 (n'=21)	42.3 ± 7.9 (n'=56)	41.6 ± 7.7 (n'=77)
Sperm volume (ml)	3.5 ± 1.4 (n'=8)	3.2 ± 0.87 (n'=6)	3.5 ± 1.2 (n'=6)	3.4 ± 1.2 (n'=20)	3.5 ± 1.5 (n'=55)	3.5 ± 1.4 (n'=75)
Sperm concentration ($10^6/\text{ml}$)	27.2 ± 23.4 (n'=8)	7.9 ± 8.4 (n'=6)	22.9 ± 15.2 (n'=6)	20.1 ± 18.8 (n'=20)	27.6 ± 35.7 (n'=55)	25.6 ± 32.1 (n'=75)
Motility (a+b+c) 1 h	0 ± 0 (n'=9)	0 ± 0 (n'=6)	2.6 ± 4.2 (n'=6)	0.7 ± 2.4* (n'=21)	5 ± 6.1 (n'=55)	3.9 ± 5.6 (n'=76)
Vitality	55.5 ± 24 (n'=8)	43.3 ± 22.6 (n'=6)	51.2 ± 23.1 (n'=5)	50.5 ± 22.7 (n'=19)	53.4 ± 20 (n'=53)	52.7 ± 20 (n'=72)
Normal spermatozoa	1.25 ± 3.5 (n'=8)	0 ± 0 (n'=6)	0 ± 0 (n'=6)	0.5 ± 2.3 (n'=20)	2.1 ± 4.1 (n'=54)	1.6 ± 2.7 (n'=61)
Absent flagella	21.8 ± 17.6 (n'=5)	36.8 ± 4.1 (n'=5)	25.6 ± 15.9 (n'=5)	28.1 ± 14.4* (n'=15)	18.5 ± 15.5 (n'=51)	20.7 ± 15.7 (n'=66)
Short Flagella	65.3 ± 31.7 (n'=8)	52.2 ± 27 (n'=6)	49.8 ± 24.3 (n'=5)	57.1 ± 27.9* (n'=19)	38.9 ± 25.7 (n'=53)	43.7 ± 27.3 (n'=72)
Coiled Flagella	8.2 ± 6 (n'=6)	14.4 ± 7 (n'=5)	9 ± 6.3 (n'=5)	10.4 ± 6.6 (n'=16)	13.5 ± 10 (n'=53)	12.8 ± 9.4 (n'=69)
Bent Flagella	10.3 ± 6 (n'=3)	9 (n'=1)	6 ± 8.5 (n'=2)	8.7 ± 5.8 (n'=6)	12 ± 9 (n'=20)	4.2 ± 8.4 (n'=26)
Flagella of irregular caliber	20.2 ± 19.3 (n'=5)	28.4 ± 16.9 (n'=5)	35 ± 22.7 (n'=5)	27.9 ± 19.4 (n'=15)	32.8 ± 26.7 (n'=52)	31.7 ± 25.1 (n'=67)
Multiple anomalies index	2.3 ± 0.2 (n'=4)	3.4 ± 0.4 (n'=5)	2.3 ± 1.3 (n'=5)	2.7 ± 1 (n'=14)	2.7 ± 0.6 (n'=47)	2.7 ± 0.7 (n'=61)

CFAP43+, CFAP44+ and DNAH1+ correspond to patient mutated in CFAP43, CFAP44 and DNAH1, respectively

Values are expressed in percent, unless specified otherwise.

Values are mean +/- SD; n= total number of patients in each group; n'= number of patients used to calculate the average based on available data.

* Significant P <0.05

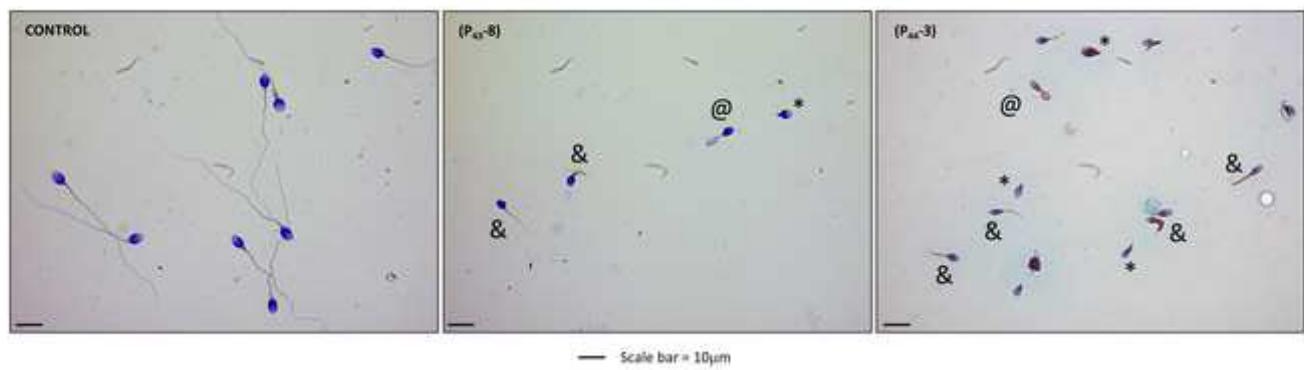
Table 2

Table 2. *CFAP43 (WDR96), CFAP44 (WDR52), and DNAH1* variants identified by WES for all the analysed subjects (n=78).

Gene	Variant coordinates	Transcript	cDNA Variation	Amino acid variation	Patients	Nationality	Hom./Het.
<i>CFAP43</i>	chr10:105912486	ENST00000357060	c.3541-2A>C	p.Ser1181lysfs*4	P ₄₃ -1, P ₄₃ -2	Tunisia	Homozygous
<i>CFAP43</i>	chr10:105956662	ENST00000357060	c.1240_1241delGT	p.Val414LeufsTer46	P ₄₃ -3, P ₄₃ -4	Afghanistan, Iran	Homozygous
<i>CFAP43</i>	chr10:105928535	ENST00000357060	c.2658G>A	p.Trp886Ter	P ₄₃ -5	Algeria	Homozygous
<i>CFAP43</i>	chr10:105928513	ENST00000357060	c.2680C>T	p.Arg894Ter	P ₄₃ -6	Algeria	Homozygous
<i>CFAP43</i>	chr10:105905296	ENST00000357060	c.3882delA	p.Glu1294AspfsTer47	P ₄₃ -7	Tunisia	Homozygous
<i>CFAP43</i>	chr10:105921781	ENST00000357060	c.3352C>T	p.Arg1118Ter	P ₄₃ -8	Tunisia	Homozygous
<i>CFAP43</i>	chr10:105953765	ENST00000357060	c.1302dupT	p.Leu435SerfsTer26	P ₄₃ -9	France	Heterozygous
<i>CFAP43</i>	chr10:105963485	ENST00000357060	c.1040T>C	p.Val347Ala	P ₄₃ -9	France	Heterozygous
<i>CFAP43</i>	chr10:105944769	ENST00000357060	c.2141+5G>A	p.Lys714Val*11	P ₄₃ -10	Turkey	Homozygous
<i>CFAP44</i>	chr3:113114596	ENST00000393845	c.1890+1G>A	p.Pro631Ile*22	P ₄₄ -1	Tunisia	Homozygous
<i>CFAP44</i>	chr3:113063450	ENST00000393845	c.3175C>T	p.Arg1059Ter	P ₄₄ -2	Tunisia	Homozygous
<i>CFAP44</i>	chr3:113082107_113082108	ENST00000393845	c.2818dupG	p.Glu940GlyfsTer19	P ₄₄ -3	Morocco	Homozygous
<i>CFAP44</i>	chr3:113119479	ENST00000393845	c.1387G>T	p.Glu463Ter	P ₄₄ -4, P ₄₄ -5	Algeria	Homozygous
<i>CFAP44</i>	chr3:113023990	ENST00000393845	c.4767delT	p.Ile1589MetfsTer6	P ₄₄ -6	Algeria	Homozygous
<i>DNAH1</i>	chr3:52414073	ENST00000420323	c.7531delC	p.Gln2511SerfsTer27	P _{DNA} -1	Tunisia	Homozygous
<i>DNAH1</i>	chr3:52382924	ENST00000420323	c.2127dupC	p.Ile710HisfsTer4	P _{DNA} -2	Tunisia	Homozygous
<i>DNAH1</i>	chr3:52395227	ENST00000420323	c.4744_4752delCCAGCTGGC	p.Pro1582_Gly1584del	P _{DNA} -3	Tunisia	Homozygous
<i>DNAH1</i>	chr3:52394055	ENST00000420323	c.4531G>A	p.Val1511Met	P _{DNA} -4	Tunisia	Heterozygous
<i>DNAH1</i>	chr3:52394397	ENST00000420323	c.4642C>G	p.Leu1548Val	P _{DNA} -4	Iran	Heterozygous
<i>DNAH1</i>	chr3:52409423	ENST00000420323	c.7153T>A	p.Trp2385Arg	P _{DNA} -5	Iran	Heterozygous
<i>DNAH1</i>	chr3:52423486	ENST00000420323	c.9505C>G	p.Arg3169Gly	P _{DNA} -5	France	Heterozygous

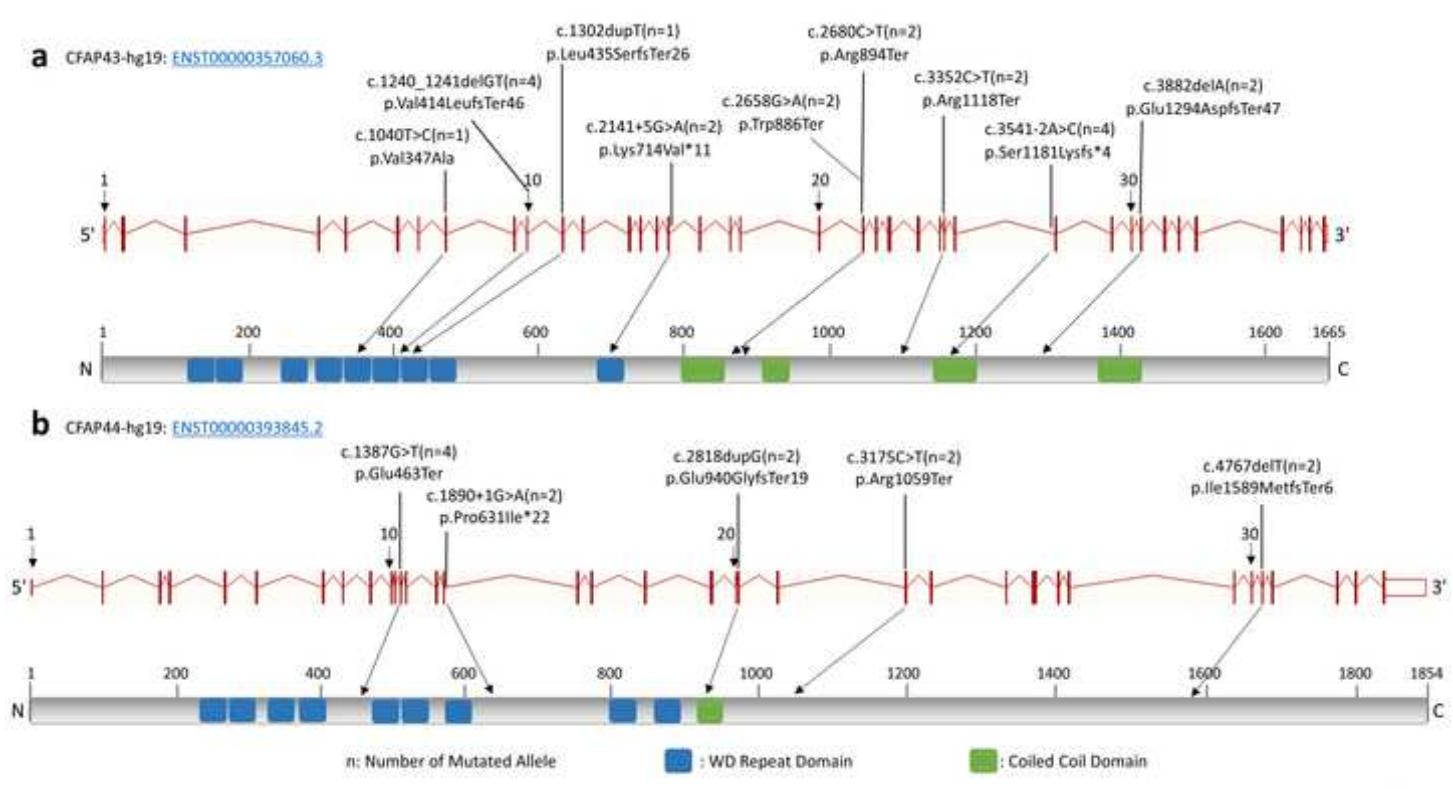
Figure 1

[Click here to download Figure Figure 1.tif](#)



Coutton et al., Fig. 1

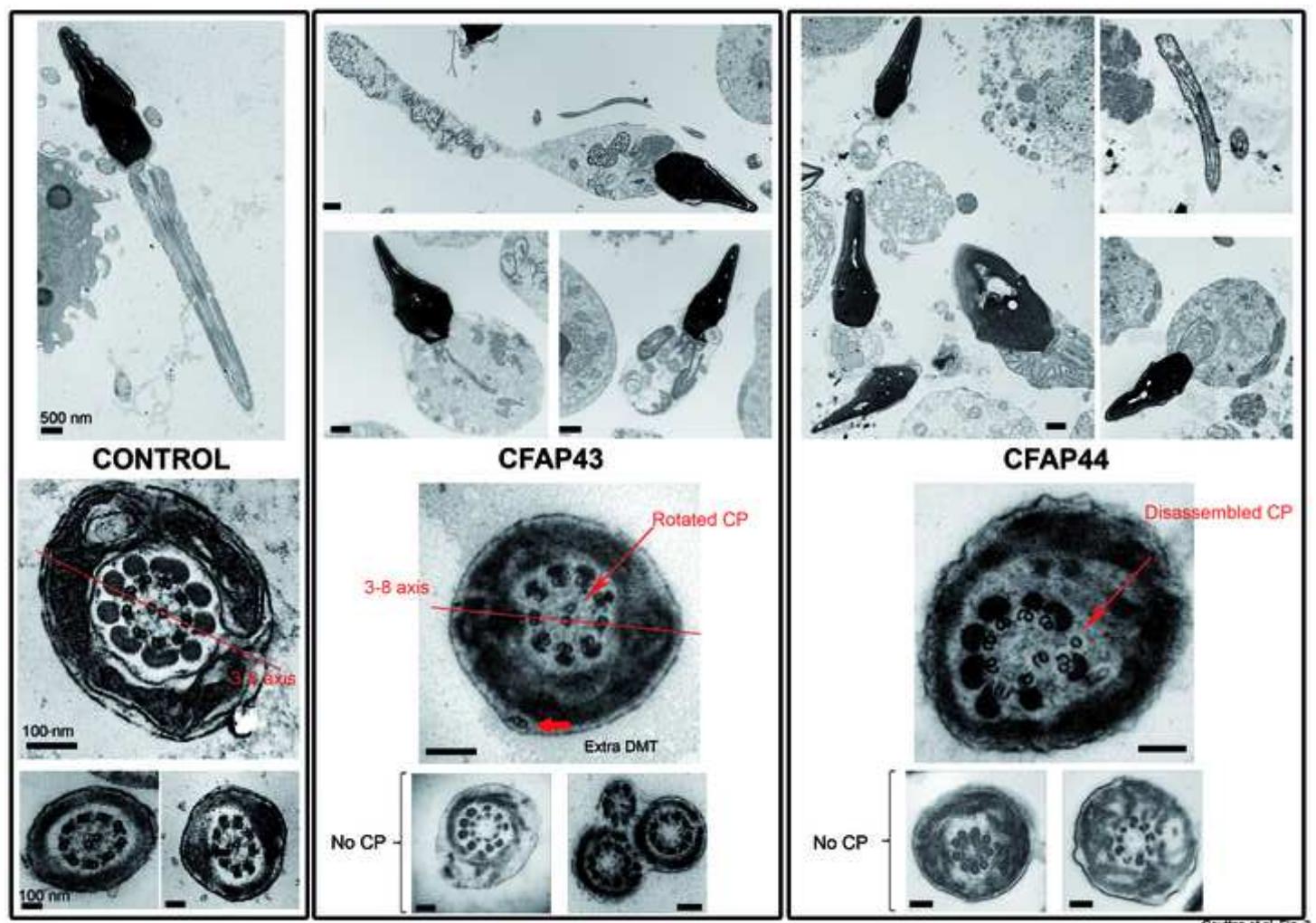
Figure 2

[Click here to download Figure Figure 2.tif](#)

Coutton et al., Fig.2

Figure 3

[Click here to download Figure figure 3.jpg](#)



Coutton et al. FIG. 3

Figure 4

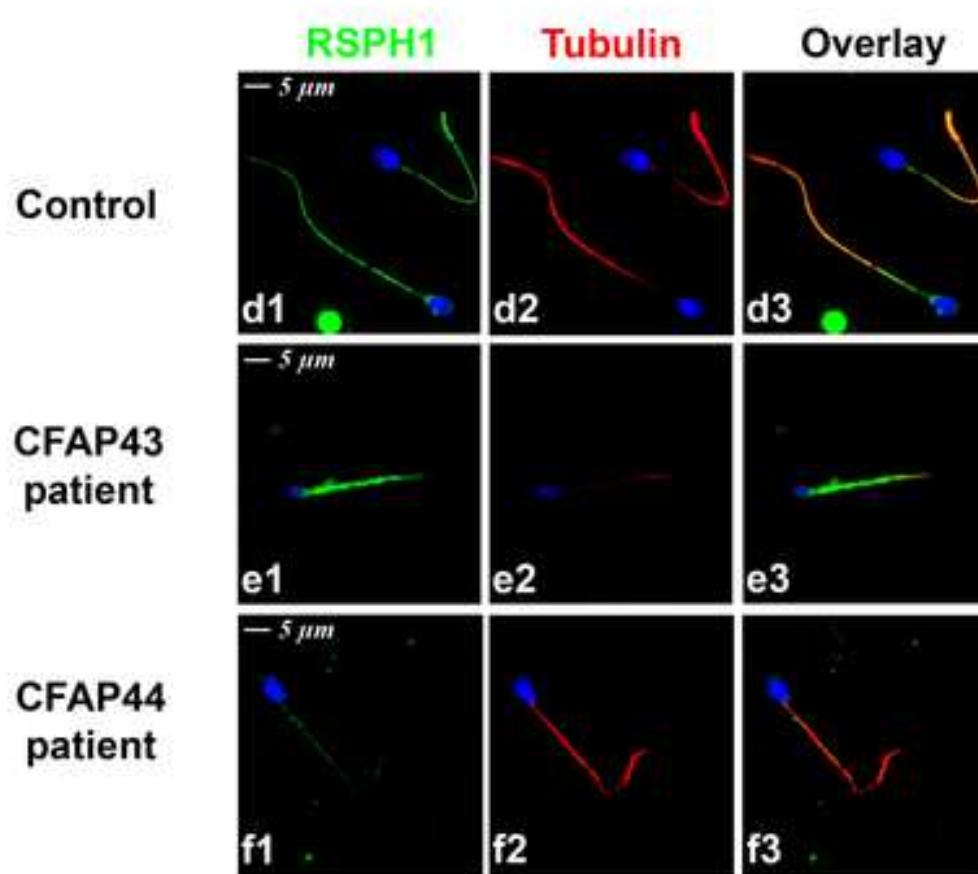
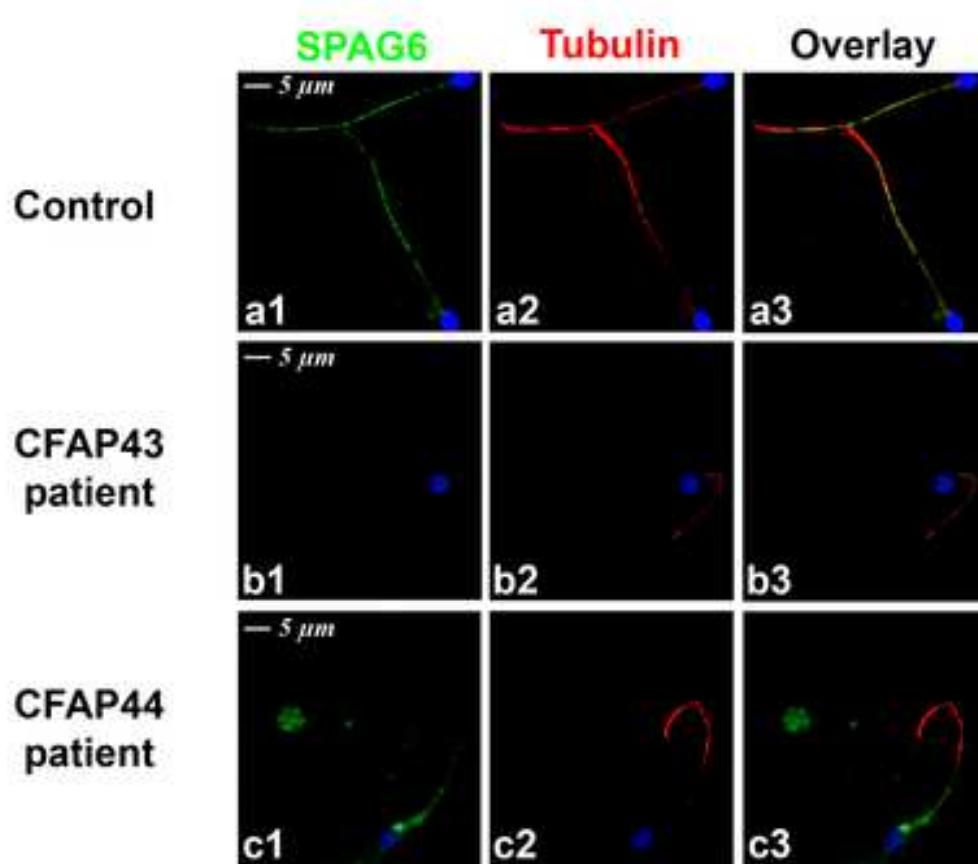
[Click here to download Figure figure 4.jpg](#)

Figure 5

Click here to download Figure figure 5.jpg

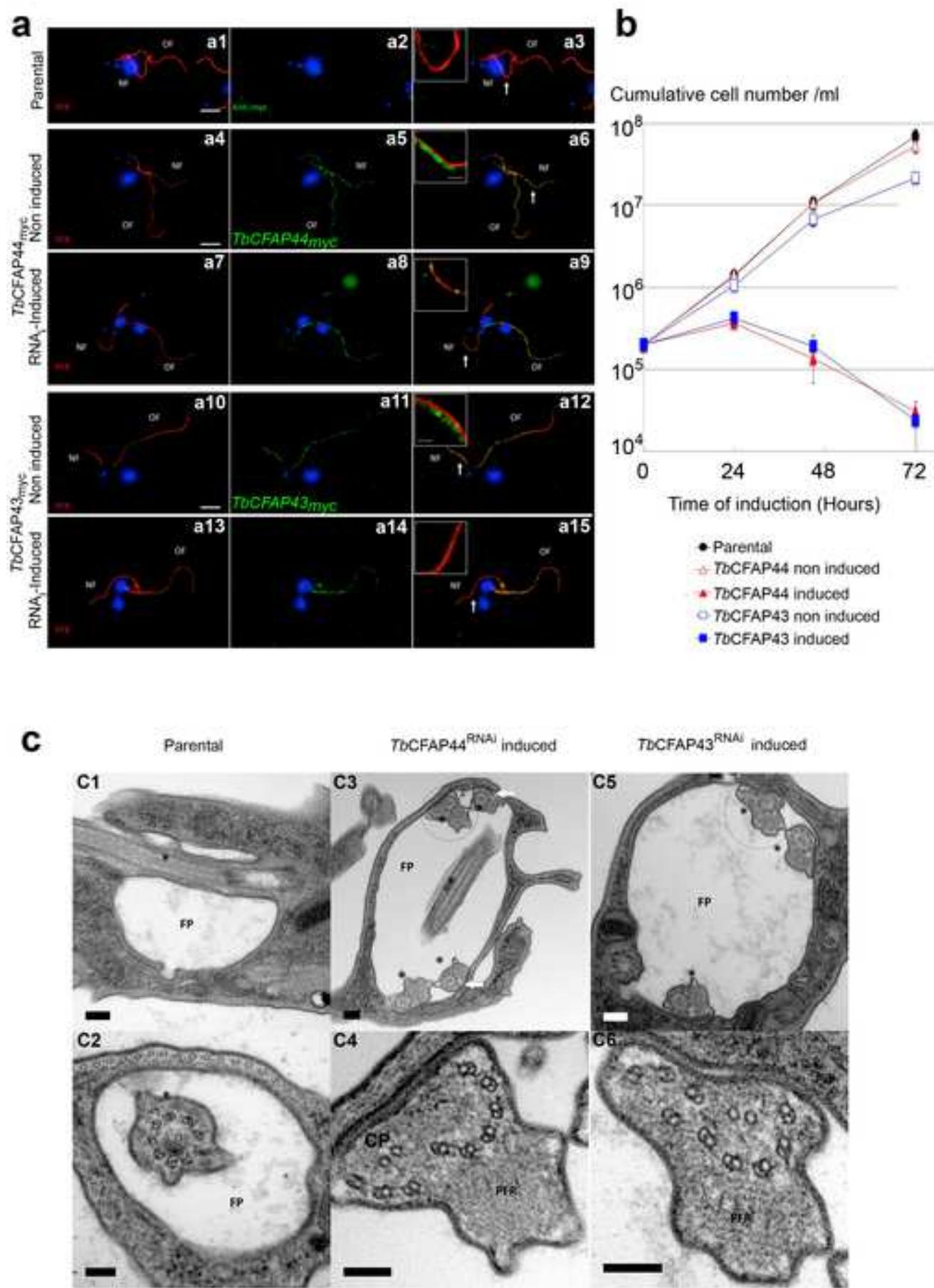


Figure 6

Click here to download Figure figure 6.jpg ↗

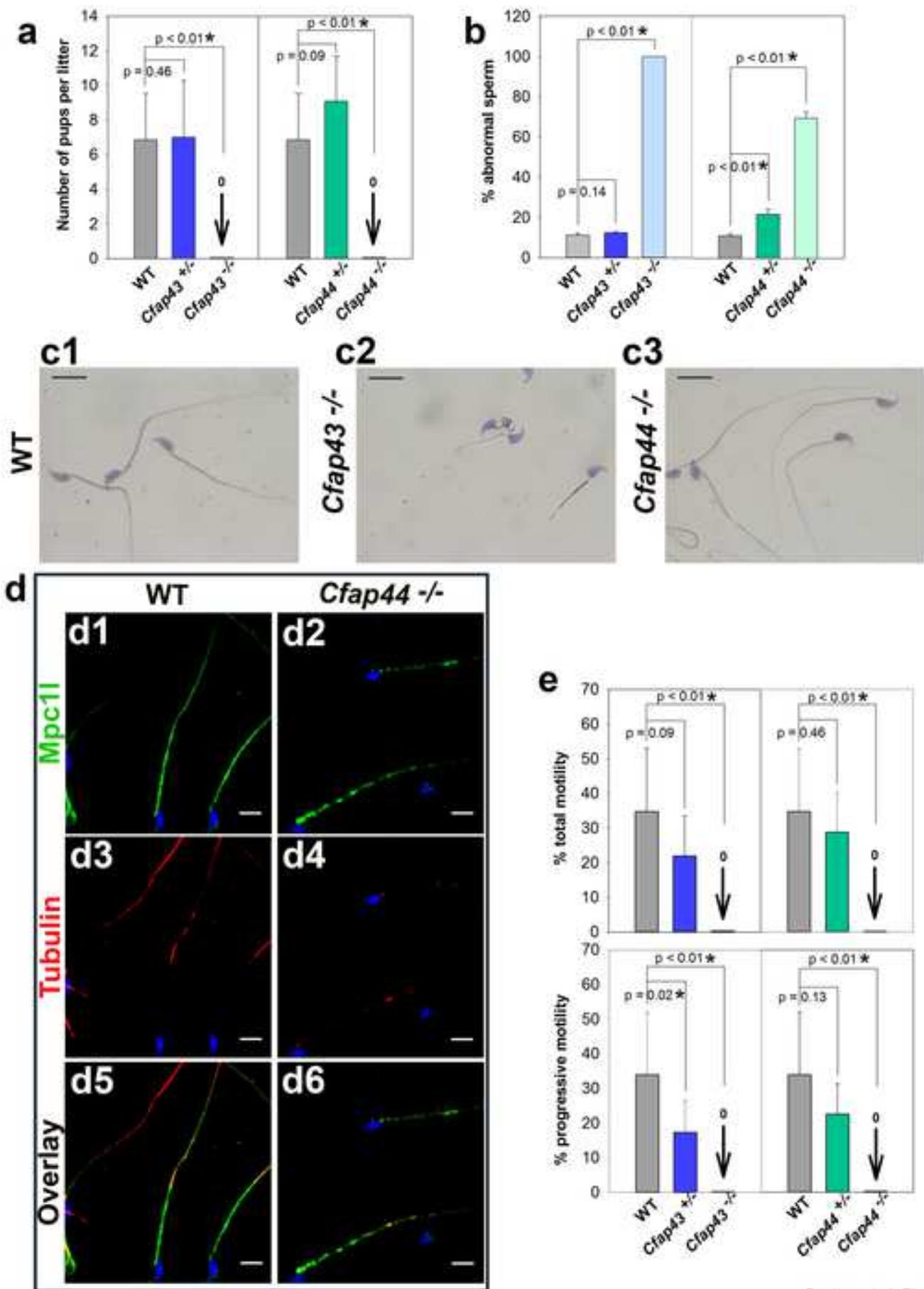


Figure 7

Click here to download Figure figure 7.jpg

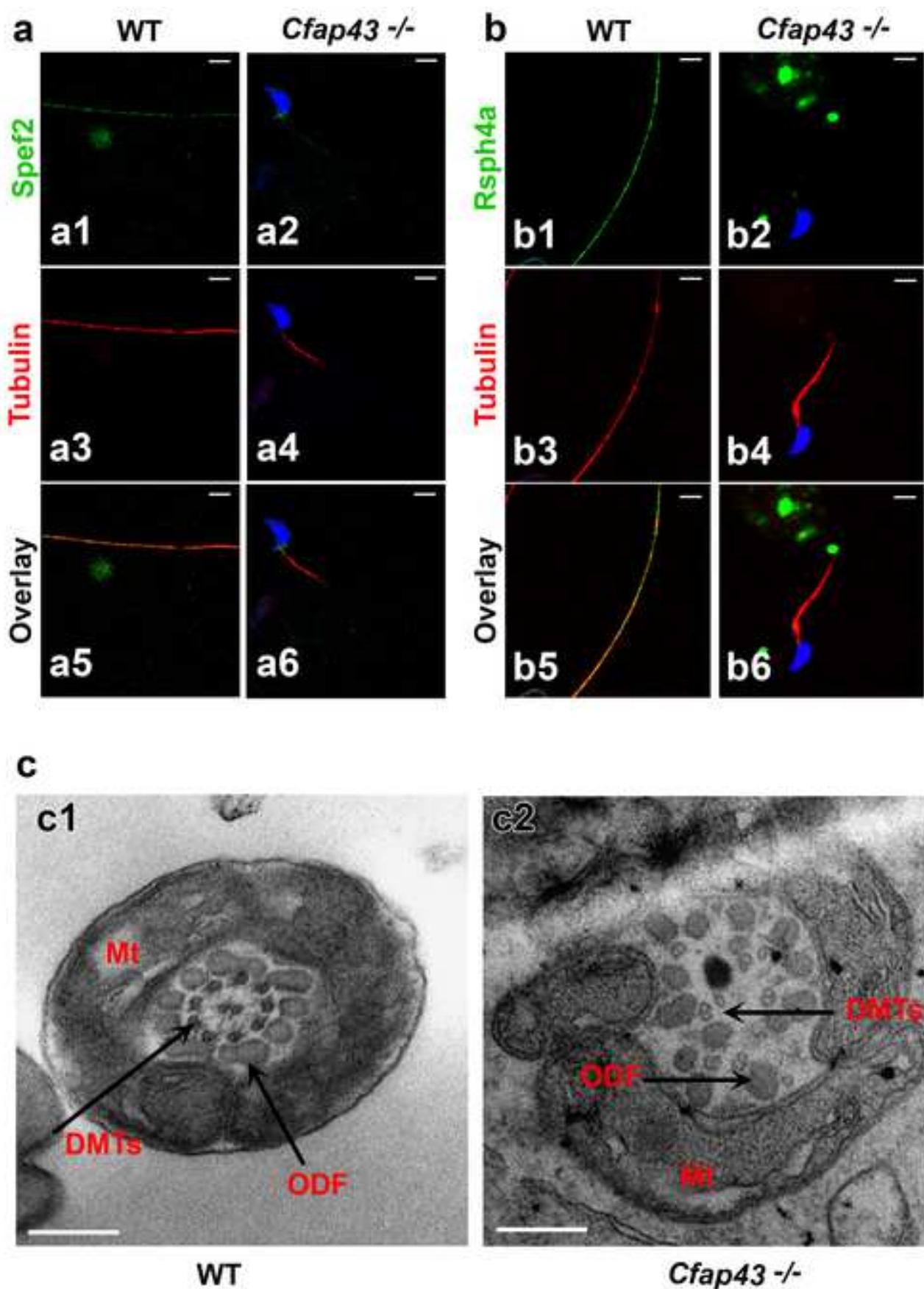
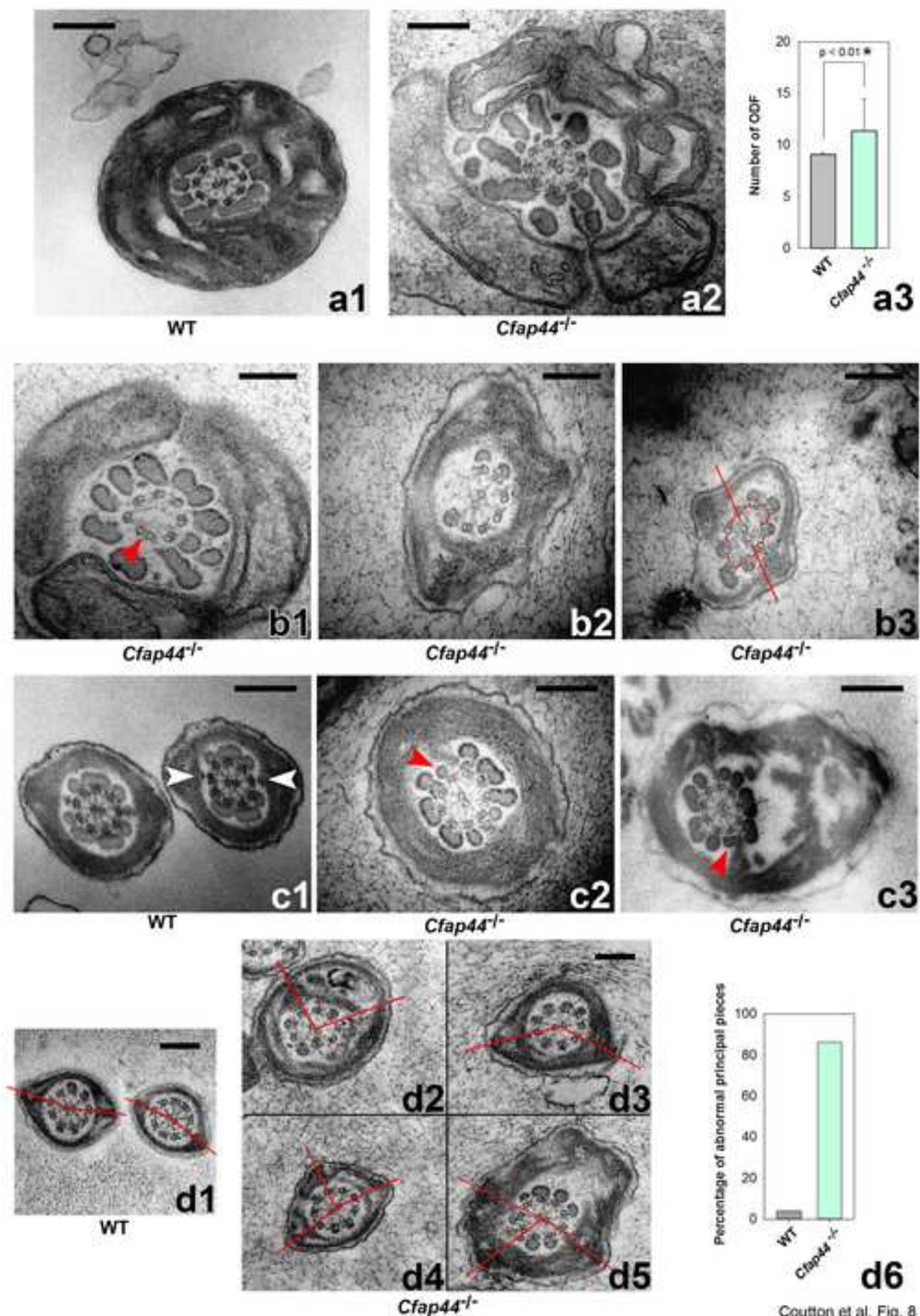


Figure 8

Click here to download Figure figure 8.jpg



Coutton et al, Fig. 8

Principaux résultats

L'application de notre pipeline d'analyse sur les données de ces 78 patients nous a permis d'obtenir une liste de 3630 variants distincts ayant passé l'ensemble de nos filtres (2903 SNPs et 727 indels), ceux-ci impactant un total de 2780 gènes différents.

Le gène *DNAH1* étant le candidat évident, nous avons cherché en priorité l'ensemble des variants retrouvés sur ce gène. Ainsi, nous avons obtenu une liste de 5 patients portant tous, soit au moins un variant homozygote sur le gène *DNAH1*, soit deux variants hétérozygotes sur ce même gène.

Suite à cela, au vu du nombre important de gènes restant et afin d'étudier en priorité ceux pouvant expliquer le phénotype d'un maximum de patients nous avons limité nos recherches aux gènes sur lesquels **au moins 3 patients portaient un variant homozygote tronquant**. Cela nous a ainsi permis de mettre en évidence les gènes *CFAP43* et *CFAP44* sur lesquels des variants homozygotes ont été retrouvés chez respectivement 9 et 6 patients auxquels viennent s'ajouter 1 patient portant deux variants hétérozygotes sur le gène *CFAP43* (**Tables : D.1 et D.2**). Ces deux gènes CFAP (pour *Cilia and Flagella Associated Protein*) avaient déjà été répertoriés dans les bases de données publiques comme ayant une forte expression testiculaire, et comme étant probablement impliqués dans la structure et / ou fonction du flagelle spermatique [208]. De plus, ces deux gènes codent tous deux pour des protéines appartenant à la famille des WDR et comportent tous deux neuf répétitions WD (tryptophane - acide aspartique) [209]. Ainsi, en tenant compte du nombre important de patients portant des variants sur un de ces deux gènes et le fait qu'ils codent tous deux pour des protéines appartenant à la même famille, nous avons décidé de nous concentrer dans un premiers temps sur la caractérisation de ces deux seuls gènes, ceux-ci étant les meilleurs candidats pour expliquer le phénotype d'infertilité de 16 de nos patients.

Un total de 9 a ainsi pu être identifié sur le gène *CFAP43* impactant 10 de nos patients n'ayant, à notre connaissance, aucun lien de parenté. Huit de ces variants ont un effet tronquant évident et le dernier est un variant intronique localisé 5 nucléotides après l'exon 16, non listé dans ExAC et prédit par Human Splicing Finder (<http://www.umd.be/HSF3>) comme altérant l'épissage de l'exon 16 de *CFAP43*.

Pour *CFAP44*, 6 de nos patients portaient des variants homozygotes ayant tous un effet tronquant.

L'analyse au microscope électronique à transmission des cellules spermatiques d'un patient portant un variant sur *CFAP43* et d'un autre portant un variant sur *CFAP44* révéla des défauts au niveau de l'axonème ainsi qu'une gaine fibreuse désorganisée pour chacun des deux patients.

Ensuite, afin de compenser l'absence d'anticorps anti-*CFAP43* et anti-*CFAP44* fiables chez l'humain comme chez la souris nous avons décidé de caractériser leur orthologues chez *Trypanosoma brucei* (*T. brucei*), un protozoaire flagellé utilisé comme organisme

modèle dans l'étude des flagelles chez qui, les protéines *TbCFAP43* et *TbCFAP44* respectivement orthologues de CFAP43 et CFAP44 avaient déjà été identifiées comme des protéines du flagelle [210, 211]. Ensuite, l'utilisation d'ARN interférence nous a permis de produire des organismes *knock-down* pour ces deux gènes *TbCFAP43^{RAi}* et *TbCFAP44^{RNAi}* nous permettant ainsi d'évaluer la fonction de ces deux gènes au sein du flagelle du trypanosome. Cela nous a permis d'observer un arrêt de la prolifération cellulaire au bout de 24h ainsi que de nombreux défauts au niveau des flagelles pour l'ensemble des lignées cellulaires *TbCFAP43^{RAi}* et *TbCFAP44^{RNAi}*. Cet arrêt de prolifération est typique des problèmes flagellaires chez le trypanosome qui dépend de son mouvement pour sa survie.

Pour finir, l'impact de l'absence des protéines CFAP43 et 44 sur la spermatogénèse murine a été déterminé grâce à la génération de modèle KO utilisant la technologie CRISPR-Cas9 qui nous a permis d'obtenir des phénotypes reproductibles pour nos deux modèles de souris KO. Ces modèles nous ont permis d'observer que les mâles *Cfap43^{-/-}* et *Cfap44^{-/-}* présentaient tous deux de nombreuses anomalies au niveau des flagelles tandis que les femelles *Cfap43^{-/-}* et *Cfap44^{-/-}* étaient parfaitement fertiles.

Pour conclure, cette étude portant sur la caractérisation du phénotype MMAF nous a permis d'identifier la cause génétique de 21 (26.9%) de nos patients, ceci n'ayant aucun lien de parenté. L'utilisation de notre pipeline pour l'analyse des données NGS nous a permis à la fois de confirmer l'importance du gène *DNAH1* dans la structure du flagelle et son implication dans ce phénotype, mais aussi d'identifier deux nouveaux gènes, *CFAP43* et *CFAP44* respectivement responsables du phénotype de 10 et 6 de nos patients soit 12.8 et 7.7% de notre cohorte.

CHAPITRE 3

Investigation génétique et physiologique de la globozoospermie

3.1 Introduction sur la globozoospermie

Comme expliqué précédemment, La globozoospermie est un phénotype rare (< 0.1% des patients infertiles) mais sévère [92] de teratozoospermie menant à l'infertilité masculine. Cette anomalie est caractérisée par la présence de spermatozoïdes présentant une tête ronde dépourvue d'acrosome et d'une pièce intermédiaire désorganisée dans l'éjaculat [212, 213] (**Figure : 3.1**). En plus des anomalies morphologiques, les spermatozoïdes globozoocéphales présentent également des désorganisations au niveau moléculaire. Par exemple, le facteur spermatique PLCZ1 requis pour l'activation ovocytaire, est absent ou présent en quantité infime dans les spermatozoïdes globozoocéphales [108, 214, 215] compromettant ainsi l'activation ovocytaire et expliquant le faible taux de fécondation observés en fécondation *in vitro* (IVF) et en ICSI (*intra cytoplasmic sperm injection*) [91]. On distingue la globozoospermie totale avec 100% des spermatozoïdes présentant le phénotype ou partielle en fonction du taux de spermatozoïdes atteints. La présence de cas familiaux dans les premières études présentant des patients atteints par un phénotype complet [92] suggéraient que la globozoospermie avait une cause génétique. De plus les caractéristiques morphologiques très typiques des spermatozoïdes laissaient penser à une cause monogénique. En 2007, une étude portant sur une famille juive ashkénaze comprenant six frères dont trois atteins a pu lier ce phénotype à une mutation homozygotes sur le gène *SPATA16* présente chez les trois frères atteint [216]. Cependant, dans la même étude, 29 autres patients présentant le même phénotype ont été analysés, et pour ceux-ci, aucun variant du gène *SPATA16* n'a pu être lié au phénotype [216] indiquant clairement que les mutations de ce gène n'étaient pas les seules responsables de ce phénotype. En 2011, une autre étude portant sur une cohorte de 20 patients Tunisiens a pu mettre en évidence une délétion homozygote de 200kb emportant la totalité du gène *DPY19L2* chez 15 des 20 patients analysés [96]. Les études effectuées ultérieurement sur ce phénotype ont ensuite confirmées que les altérations du gène *DPY19L2*, et notamment délétion décrite initialement, étaient responsables de la majorité des cas de globozoospermie [111, 217].

En 2012, le développement d'un modèle murin KO *Dpy19l2^{-/-}* a permis de mieux comprendre les mécanismes moléculaires impliqués dans la globozoospermie causée par la délétion du gène *DPY19L2* chez l'humain [218]. Ce modèle de souris KO présentait les mêmes caractéristiques que les patients humains. Ces souris étaient infertiles et présentaient des spermatozoïdes globozoocéphales (**Figure : 3.2**) ainsi que d'autre défauts secondaires retrouvés chez l'homme comme l'absence d'acrosome, des défauts morphologiques du noyau, de l'enveloppe nucléaire et de l'acoplaxome ainsi que le mauvais positionnement de la manchette [218]. Il a pu être démontré que la protéine Dpy19l2 étaient principalement exprimée dans au stade spermatide ronde et plus spécifiquement dans la membrane nucléaire interne faisant face à la vésicule acrosomale et que l'absence de cette protéine entraînait la déstabilisation de la jonction entre l'acoplaxome et l'enveloppe nucléaire [218].

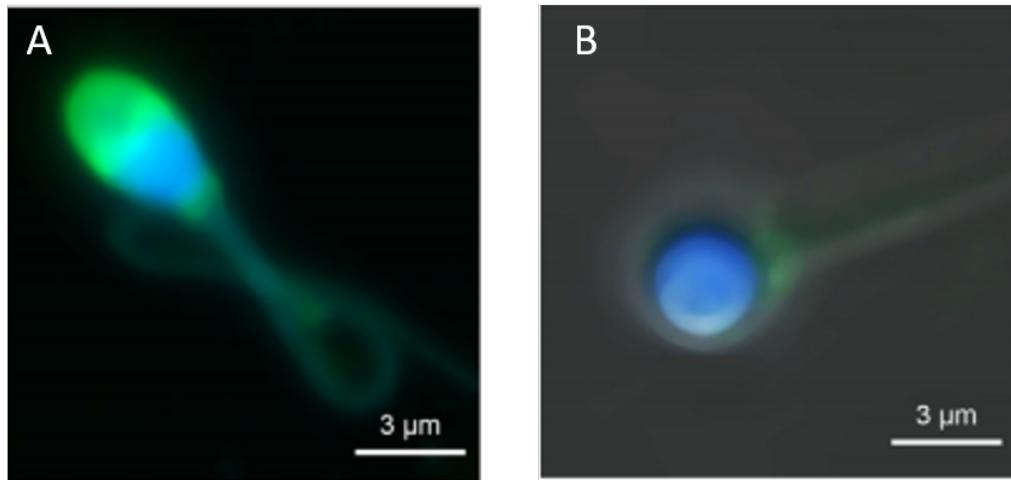


Figure 3.1 – Observation au microscope confocal de spermatozoïdes adapté d'après [96] : Sur ces deux photos, le noyau des spermatozoïdes est mis en évidence à l'aide de TopRo3 (en bleu). L'acrosome est coloré par la conjugaison de lectine de pois et de isothiocyanate de fluorescéine (PSA-FITC, Sigma Aldrich, France) (en vert). **A** : Spermatozoïde normal, l'acrosome forme une cloche entourant la tête du spermatozoïde. **B** : Spermatozoïde globozoocéphale, l'acrosome est ici absent

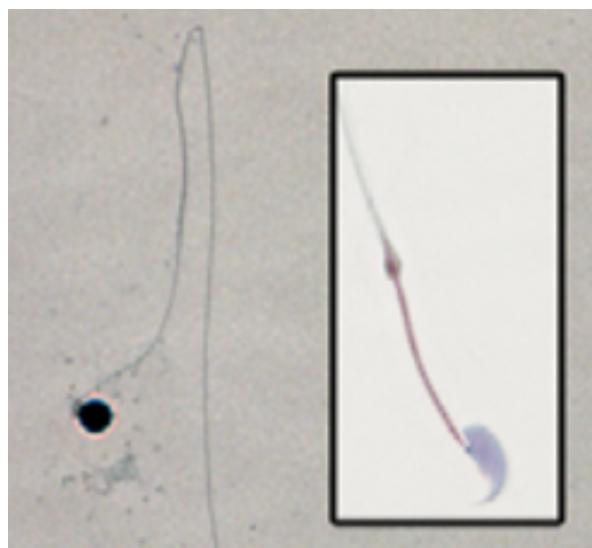


Figure 3.2 – Comparaison entre les spermatozoïdes des souris sauvages et globozoospermiques d'après [218] : À gauche, le spermatozoïde d'une souris globozoospermique *Dpy19l2*^{-/-}. À droite, celui d'une souris sauvage *Dpy19l2*^{+/+}.

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

3.2.1 Article n°6 :

Fine Characterisation of a Recombination Hotspot at the *DPY19L2* Locus and Resolution of the Paradoxical Excess of Duplications over Deletions in the General Population

Coutton C, Abada F, Karaouzène T, Sanlaville D, Satre V, Lunardi J, Jouk PS, Arnoult C, Thierry-Mieg N, Ray PF

* Co-premiers auteurs

PLOS GeneticS, Mars 2013

Contexte et objectifs

Chez les mammifères il existe trois paralogues de *DPY19L2* de fonctions encore inconnues et un pseudogène présentant une très forte homologie de séquence (> 95%) [219]. Chez l'homme, ce gène est flanqué de deux séquences présentant une forte homologie (>95%) d'une taille de 28kb. Ces séquences appelées LCRs (*low copy repeats*) représentent une large portion du génome humain [220, 221] et vont, de par leur homologie favoriser les duplications de gènes jouant ainsi un rôle important dans l'évolution des génomes des vertébrés [222, 223]. Dans le cas de *DPY19L2*, ces LCRs vont, au cours de la méiose, entraîner la survenue de recombinaisons homologues non-allélique (NAHR) donnant lieu soit à une délétion du gène *DPY19L2* et la formation d'un ADN circulaire comprenant le gène soit à un allèle possédant deux copies du gène tandis que l'autre n'en possède aucune [224].

Ce mécanisme de NAHR devrait, en théorie, engendrer la formation de plus d'allèles délétés que d'allèles dupliqués puisque seuls les remaniements interchromatidiens les induisent la formation d'un allèle dupliqué tandis que les délétions surviennent à la fois lors des remaniements interchromatidiens **et** intrachromatidiens [225] (**Figure : 3.3**). Cependant, les données mises à disposition par la base de données *Database of Genomic Variants* (DGV) [226] indiquent un excès de duplication puisque sur un total de 6575 individus analysés, 83 duplications et de 26 délétions hétérozygotes ont été observées pour le locus de *DPY19L2*.

Ainsi, dans cette étude, notre équipe a cherché à caractériser précisément le mécanisme génétique et les facteurs favorisant la survenue par NAHR de la délétion homozygote récurrente emportant totalement le gène *DPY19L2*. De même, nous avons tenté de résoudre le paradoxe observé entre le modèle théorique de NAHR et la fréquence des allèles observée dans la population générale afin de confirmer les données fournies dans les bases de données et ainsi écarter l'hypothèse d'un biais causé par la présence du pseudogène *DPY19L2P1* très homologue avec *DPY19L2* [219].

Dans ce contexte j'ai pu participer à diverses manipulations de biologie moléculaire tel que l'extraction d'ADN spermatique, quantification des délétions / duplications *de novo*. J'ai également contribué au diverses analyses statistiques.

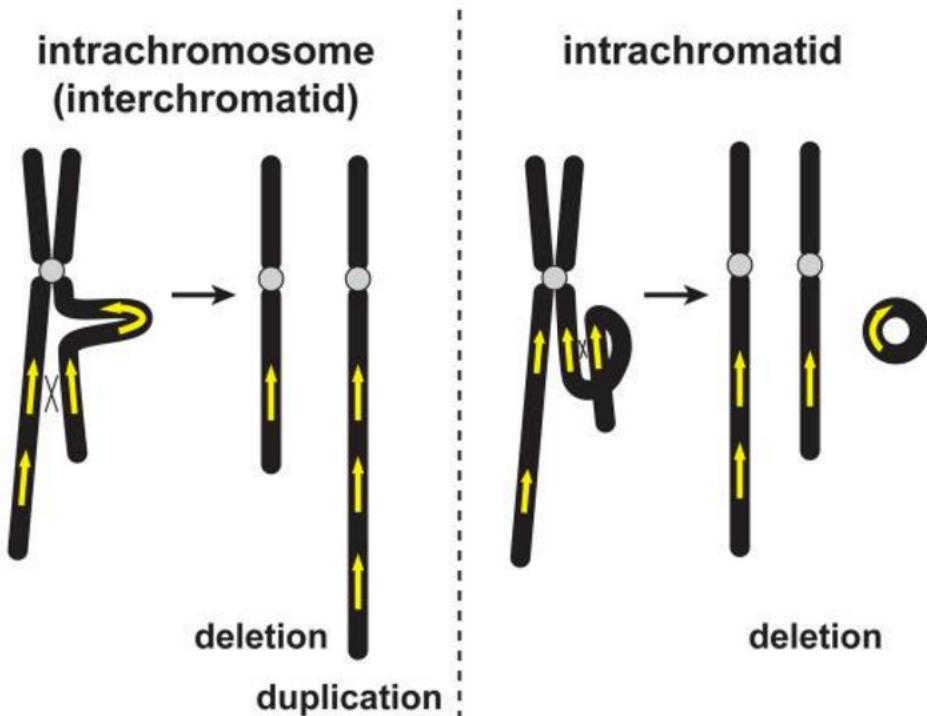


Figure 3.3 – Représentation schématique du mécanisme de NAHR adapté d'après [225] : Lors d'un NAHR interchromatidien, un allèle dupliqué et un allèle délété sont formés. Lors d'un NAHR intrachromatidien, seul un allèle délété est produit, en même temps qu'un petit ADN circulaire qui sera éliminé par la suite.

Fine Characterisation of a Recombination Hotspot at the *DPY19L2* Locus and Resolution of the Paradoxical Excess of Duplications over Deletions in the General Population

Charles Coutton^{1,2,3}, Farid Abada^{1,2}, Thomas Karaouzene^{1,2}, Damien Sanlaville⁴, Véronique Satre^{1,3}, Joël Lunardi^{1,2}, Pierre-Simon Jouk³, Christophe Arnoult¹, Nicolas Thierry-Mieg⁵, Pierre F. Ray^{1,2,3*}

1 Equipe "Génétique, Infertilité et Thérapeutiques," Laboratoire AGIM, CNRS FRE3405, Université Joseph Fourier, Grenoble, France, **2** Biochimie et Génétique Moléculaire, DBTP, CHU Grenoble, Grenoble, France, **3** Département de Génétique et Procréation, Hôpital Couple Enfant, CHU de Grenoble, Grenoble, France, **4** Hospices Civils de Lyon, Centre de Biologie et de Pathologie Est, Service de Cytogénétique Constitutionnelle, Lyon, France, **5** Université Joseph Fourier, Centre National de la Recherche Scientifique, Laboratoire TIMC-IMAG UMR 5525, Grenoble, France

Abstract

We demonstrated previously that 75% of infertile men with round, acrosomeless spermatozoa (globozoospermia) had a homozygous 200-Kb deletion removing the totality of *DPY19L2*. We showed that this deletion occurred by Non-Allelic Homologous Recombination (NAHR) between two homologous 28-Kb Low Copy Repeats (LCRs) located on each side of the gene. The accepted NAHR model predicts that inter-chromatid and inter-chromosome NAHR create a deleted and a duplicated recombined allele, while intra-chromatid events only generate deletions. Therefore more deletions are expected to be produced de novo. Surprisingly, array CGH data show that, in the general population, *DPY19L2* duplicated alleles are approximately three times as frequent as deleted alleles. In order to shed light on this paradox, we developed a sperm-based assay to measure the de novo rates of deletions and duplications at this locus. As predicted by the NAHR model, we identified an excess of de novo deletions over duplications. We calculated that the excess of de novo deletion was compensated by evolutionary loss, whereas duplications, not subjected to selection, increased gradually. Purifying selection against sterile, homozygous deleted men may be sufficient for this compensation, but heterozygotously deleted men might also suffer a small fitness penalty. The recombined alleles were sequenced to pinpoint the localisation of the breakpoints. We analysed a total of 15 homozygous deleted patients and 17 heterozygous individuals carrying either a deletion ($n=4$) or a duplication ($n=13$). All but two alleles fell within a 1.2-Kb region central to the 28-Kb LCR, indicating that >90% of the NAHR took place in that region. We showed that a PRDM9 13-mer recognition sequence is located right in the centre of that region. Our results therefore strengthen the link between this consensus sequence and the occurrence of NAHR.

Citation: Coutton C, Abada F, Karaouzene T, Sanlaville D, Satre V, et al. (2013) Fine Characterisation of a Recombination Hotspot at the *DPY19L2* Locus and Resolution of the Paradoxical Excess of Duplications over Deletions in the General Population. PLoS Genet 9(3): e1003363. doi:10.1371/journal.pgen.1003363

Editor: Nancy B. Spinner, University of Pennsylvania, United States of America

Received October 3, 2012; **Accepted** January 19, 2013; **Published** March 21, 2013

Copyright: © 2013 Coutton et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the program GENOPAT 2009 from the French Research Agency (ANR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pray@chu-grenoble.fr

Introduction

Several mechanisms have been proposed to cause genomic rearrangements, notably: Non Allelic Homologous Recombination (NAHR), Non Homologous End Joining (NHEJ), Fork Stalling and Template Switching (FoSTeS) and Break-Induced Replication (BIR) [1,2]. NAHR takes place between duplicated sequences with a high sequence identity (usually >95%) located in different genomic regions of the same chromosome [3]. These paralogous sequences or Low Copy Repeats (LCR) tend to generate polymorphic regions with deleted and duplicated alleles called Copy Number Variants (CNVs). The consensual NAHR model predicts that recombinations between LCRs located on the same chromatid result in the production of a deleted allele and a small circular molecule that will be lost by the end of the cell cycle. Recombinations between LCRs located on two distinct chromatids (whether sister-chromatids or chromatids from homologous chromosomes) result in the production of a deleted allele and a

complementary duplicated allele (Figure 1A). In consequence NAHR is expected to produce an excess of deletions over duplications. This has been verified for several NAHR hotspots using sperm typing assays: on average twice as many deletions as duplications were generated *de novo* [4]. One study however describes similar deletion and duplication frequencies at the 7q11.23, 15q11-q13 and 22q11.2 loci, suggesting a predominant inter-chromatid NAHR [5]. This study was carried out by fluorescent in situ hybridization (FISH) which allows the detection of all numerical anomalies occurring at these loci and not only the NAHR-mediated events. This could explain at least part of the discrepancy observed between the two studies, given that a recent study of the *RAII* locus suggested that complex genetic events generate an excess of duplications [6].

As NAHRs events occur at fixed LCRs they tend to be recurrent, and the recombined alleles normally share a common size defined by the distance separating the two LCRs. It is well-established that meiotic recombination events, whether resulting in crossing over or producing

Author Summary

We demonstrated previously that most men with globozoospermia, who produce only round acosomeless spermatozoa and are 100% infertile, had a homozygous deletion removing the totality of *DPY19L2*. We also showed that this deletion occurred by Non-Allelic Homologous Recombination (NAHR). NAHR results in the production of deletions and duplications of regions encompassed by two homologous sequences, normally with a higher occurrence of deletions over duplications. Analysis of public databases at the *DPY19L2* locus paradoxically revealed that, in the general population, duplications were approximately three times as frequent as deletions. Analysis of sperm DNA permits us to quantify *de novo* events that take place during male meiosis. We therefore measured the rates of *de novo* deletion and duplication in the sperm of three healthy donors. As predicted by the NAHR theoretical model and contrary to the allelic frequency observed in the general population, we identified an approximate 2-fold excess of deletions over duplications. We calculated that the measured rate of *de novo* deletion was compensated by evolutionary loss, whereas duplications, not subjected to selection, increased gradually. Purifying selection against infertile homozygously deleted men may be sufficient for this compensation, or heterozygotously deleted men may also suffer a small fitness penalty.

unbalanced alleles through NAHR, are not uniformly distributed along the human genome but occur preferentially at specific hot spots [7–9]. Myers et al. (2008) have characterized a degenerate 13 bp sequence motif (CCNCCNTNNCCNC) that is present in approximately 40% of the identified human crossover hotspots. A three nucleotide periodicity was observed within and beyond the 13-mer core, suggesting a direct interaction with a motif binding protein [10]. Subsequent work strengthened this hypothesis as it has been proposed that PRDM9, a multi-unit zinc finger binding protein expressed mainly during early meiosis in germ cells [11], specifies hotspot usage by binding specifically to this 13 bp consensus motif [12–14]. *PRDM9* was then shown to be highly polymorphic, different alleles seemingly providing preferred targeted recombination hotspots [14]. Berg et al. (2010) measured the recombination rate at ten crossover hotspots, five with a *PRDM9* recognition motif, five without a clear motif. Men with the rarer N allele showed a heavy reduction (>30-fold) at all hotspots, even at those which did not contain an obvious *PRDM9* motif [12]. Further work revealed that specific *PRDM9* alleles activated different hotspots [15]. The direct correlation between *PRDM9* recognition sequence and *PRDM9* genotype however remains elusive, indicating that the rules governing the interaction between *PRDM9* and its targeted sequences must be subtle and complex [12,15].

CNVs and other unbalanced micro recombination events are involved in the aetiology of many human pathologies such as Alpha Thalassemia, Potocki-Lupski Syndrome, Charcot-Marie Tooth, Williams-Beuren syndrome, Prader Willi/Angelman syndrome, and infertility through the production of Y-chromosome microdeletions [16]. Here we focus on the *DPY19L2* locus (12q14.2) which has recently been shown to be linked with Globozoospermia [17], a rare syndrome of male infertility [18] characterized by the presence of 100% round, acosomeless spermatozoa in the patient's ejaculate (MIM #102530). Reports of familial cases pointed to a genetic component to this pathology [19–21], and this assumption was confirmed as a homozygous mutation of *SPATA16* was identified in three siblings [22] and a homozygous missense mutation of *PICK1* was identified in a Chinese patient [23]. We demonstrated recently that *DPY19L2*

was in fact the main locus associated with globozoospermia as 15 out of 20 analysed patients presented a 200 Kb homozygous deletion removing the totality of the gene [17]. *DPY19L2* was described to have arisen, along with three other genes (*DPY19L1*, *L3* and *L4*), through the expansion and evolution of the *DPY19L* gene family from a single ortholog found in invertebrate animals [24]. We then identified *DPY19L2* point mutations and heterozygous deletions and demonstrated that 84% of the 31 globozoospermia patients analysed had a molecular alteration of *DPY19L2* [25]. Others find a slightly lower incidence of *DPY19L2* deletions in globozoospermia patients [26,27]. Comparison of the spermiogenesis between wild type and *Dpy19l2* knock out (KO) mice allowed us to demonstrate that *Dpy19l2* is expressed in the inner nuclear membrane only in the section facing the acrosome, and that it is necessary to anchor the acrosome to the nucleus. This indicates that DPY19 proteins (*DPY19L1-4* in mammals) might constitute a new family of structural transmembrane proteins of the nuclear envelope that likely participate in a function that was so far known to be only carried out by SUN proteins: constituting a bridge between the nucleoskeleton and cytoplasmic organelles and/or the cytoskeleton [28]. In our previous work we had demonstrated that *DPY19L2* was homozygously deleted in a majority of patients with globozoospermia and that this deletion occurred by NAHR between two highly homologous 28 Kb LCRs located on each side of the gene [17]. Strengthening the case for the occurrence of NAHR at the *DPY19L2* locus, heterozygous deletions and duplications have been identified in several large array CGH studies and this locus is classified as a CNV [29–33]. Surprisingly, considering that NAHR is known to generate an excess of deletions, these databases contain a large excess of duplications.

We developed a PCR assay to specifically amplify the recombinant LCRs corresponding to deleted and duplicated alleles allowing the precise localisation of the breakpoints (BP). We observed that all identified BPs clustered in the center of the LCR. We analysed this region and identified a 13-mer PRDM9 pro-recombination sequences in the middle of the hotspot. We also developed a digital PCR assay that enabled us to estimate the rates of *de novo* deletion and duplication at this locus. Contrary to the allelic frequency observed in the general population we measured an approximate 2 fold excess of deletions over duplications. We show that the negative selection against the deleted alleles could explain this apparent paradox.

Results

Estimation of the *DPY19L2* deleted and duplicated alleles' frequencies in the general population and assessment of the PCR assay's sensitivity

The *DPY19L2* CNV was analysed using array CGH data available from web servers [29–33] for a total of 6575 control individuals, mainly from the Database of Genomic Variants (<http://projects.tcag.ca/variation/>). A total of 83 gains and 26 heterozygous losses are reported for the *DPY19L2* CNV in this pool, indicating a threefold excess of duplications over deletions.

We wanted to confirm this result and exclude a potential technical bias towards duplications that could be caused by the presence on chromosome 7 of *DPY19L2P1*, a pseudogene highly homologous to *DPY19L2* [24]. To this end we re-analysed the array CGH data produced for the diagnosis of syndromic mental retardation in Grenoble and Lyon hospitals, and searched for *DPY19L2* deleted and duplicated alleles in this dataset. A total of 1699 array CGH profiles were re-analysed (see Figure S1 for illustration). We identified a total of 15 duplications and 3

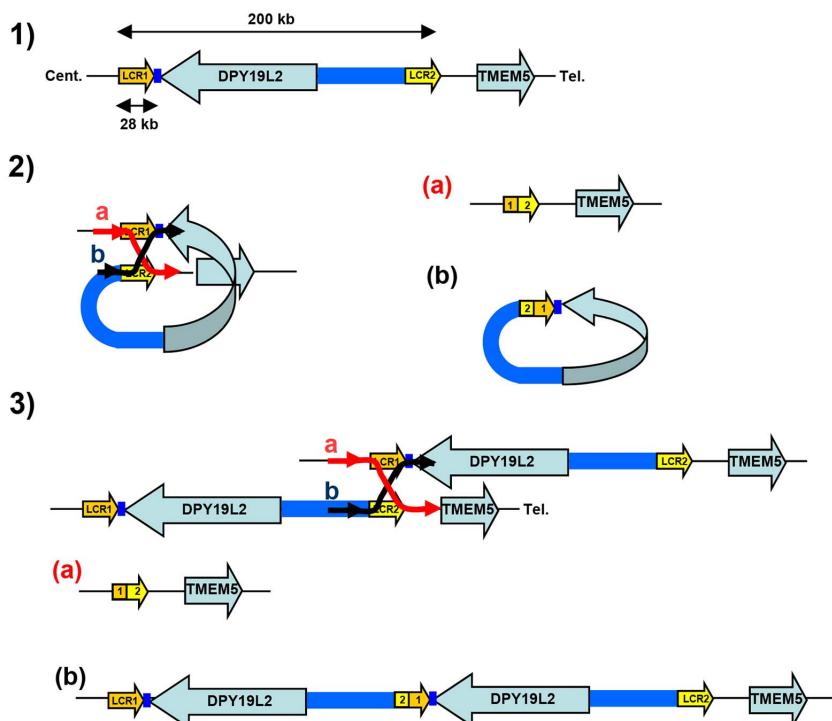
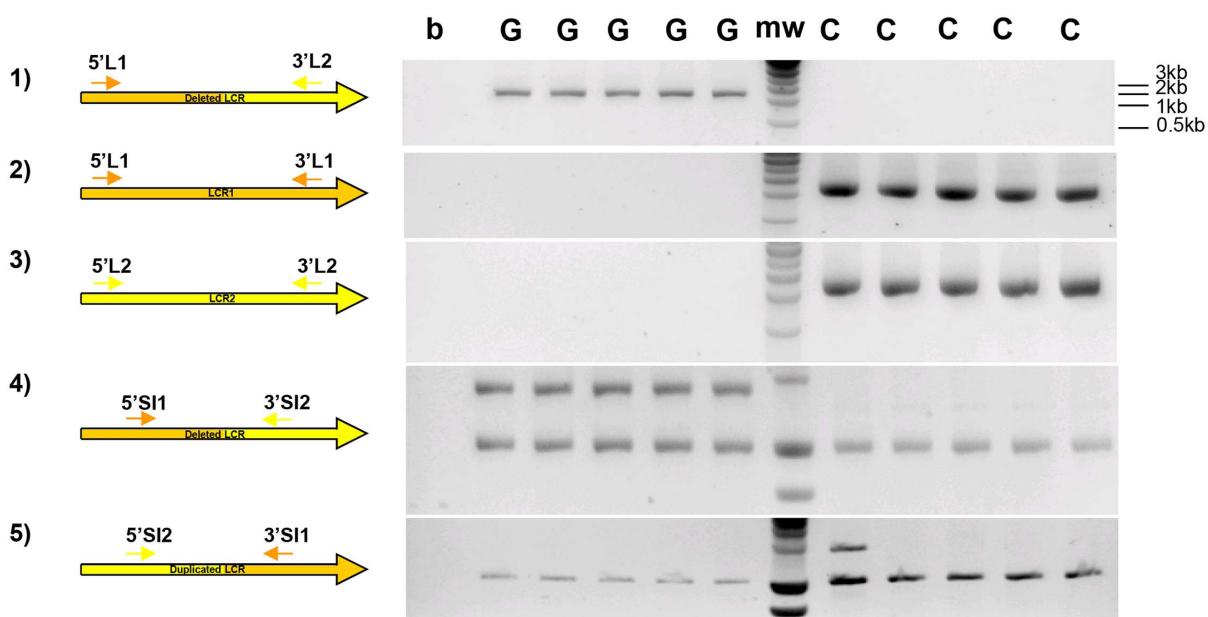
A.**B.**

Figure 1. Strategy and validation of the detection of *DPY19L2* recombinant alleles by PCR. (A) Schematic representation of NAHR at the *DPY19L2* locus. 1) LCR1 and LCR2 correspond to the centromeric and telomeric LCRs respectively. The two LCRs are separated by approximately 200 Kb and each measures 28 Kb. 2) NAHR can occur following the mis-alignment of Low Copy Repeats 1 and 2 located either on 1) the same chromatid and results in the production of a) a deleted allele with a recombinant 1-2 LCR, and b) a small circular molecule with a recombinant 2-1 LCR and the *DPY19L2* gene. This small molecule will not survive through the cell cycle. 3) NAHR can occur following the mis-alignment from two distinct chromatids (whether sister-chromatids or chromatids from homologous chromosomes). This results in the production of a) a deleted allele with a 1-2 recombinant LCR, and b) a complementary duplicated allele with a 2-1 recombinant LCR. (B) Illustration of the specificity of the LCR-specific amplification when amplifying DNA from *DPY19L2* homozygously deleted globozoospermic patients (G) and control individuals (C). 1) Primers specific to the deleted 1-2 LCR yield a 2088 nt fragment in globozoospermic patients only. 2,3) Specific amplification of LCR 1 and 2 is only obtained

from non-deleted controls. 4) Co-amplification of a control locus (bottom band) with a deleted 1-2 LCR-specific sequence. 5) Co-amplification of a control locus (bottom band) with a duplicated 2-1 LCR-specific sequence. A duplicated allele is identified in one control individual (first lane after the molecular weight markers (mw)).
doi:10.1371/journal.pgen.1003363.g001

heterozygous deletions. The recombined alleles were secondarily amplified with the long PCR primers to confirm the validity of the array CGH results. Presence of the deletion could be confirmed by our deletion-specific PCR in the three individuals putatively carrying a heterozygous deletion. DNA from 3 individuals expected to carry a duplicated allele could not be obtained. Ten out of the 12 remaining individuals putatively carrying a *DPY19L2* duplication were amplified by our duplication-specific PCR. For the two individuals that could not be amplified, the duplication was nevertheless confirmed by Multiplex Ligation-dependent Probe Amplification (MLPA). These results show that our reanalysis of the array CGH data did not yield any false positives. Overall reanalysis of these 15 individuals showed that 2 out of 15 recombinant alleles could not be detected by our PCR assay, indicating that the breakpoints of 2/15 recombined alleles fell outside of our amplified region.

We also wanted to obtain an estimation of the frequency of the deleted and duplicated alleles in the general population using our recombination-specific PCR assay. For that we designed primers that amplified a smaller sequence which could be co-amplified with an additional pair of primers (RYR2 primers) used as a positive amplification control (Figure 1B and Table S1). This duplex PCR setup controls for poor DNA quality or technical variations. We analysed 150 control individuals originating from North Africa and 150 individuals of European origin with these two duplex PCRs (for the detection of deleted and duplicated LCRs, respectively). We identified only one heterozygous deletion in an individual of North African origin and two duplications in one European and in one North African individuals.

Overall a total of 8574 individuals have been analysed, including 6575 individuals from array CGH public databases, 1699 individuals from Grenoble-Lyon array CGH data and 300 individuals analysed by recombination-specific PCR. From these cohorts we identified 30 deletions (frequency of approximately 1/290) and 100 duplications (approximate frequency 1/85) (Table S2). These values indicate that the allele frequencies of the recombined deleted and duplicated alleles are 1.7×10^{-3} (95% CI: 1.2×10^{-3} ; 2.5×10^{-3}) and 5.8×10^{-3} (95% CI: 4.7×10^{-3} ; 7.1×10^{-3}), respectively. Confidence intervals (CI) were calculated assuming a binomial model, with binom.test in R.

We note that our PCR-based assay only allows the identification of breakpoints occurring between the selected primers (1392 bp). The location of the breakpoints of each CNV detected by array CGH (an unbiased approach) located in the *DPY19L2* locus was scrutinised to establish if they were located within the LCR and hence were caused by NAHR (Table S3). This analysis shows that 87% of the deletions and 76% of the duplication fell within the LCR limits.

Overall, we believe that our PCR assay permits to identify the majority of recombinations occurring at the *DPY19L2* locus, since: 1) amplification was obtained for all 15/15 globozoospermia patients analysed, and 2) amplification was obtained for 13/15 (87%) recombined array CGH patients.

Determination of *DPY19L2* *de novo* recombination rates by digital PCR

As the previous results consistently showed an excess of duplications over deletions in the general population, we wanted to measure the rates of *de novo* duplications and deletions to verify if

the observed skew was due to the selection of duplications over deletions or if more duplications were produced *de novo*. The rate of genetic events occurring *de novo* can be measured on sperm DNA since each spermatozoon is the product of meiosis and corresponds to a new haploid genome. We first tried to develop a semi-quantitative PCR assay to directly measure the frequencies of deletions and duplications using sperm from control donors (with two copies of *DPY19L2*). The shortest fragment that could provide a reliable specific amplification and amplify the whole breakpoint area was 1392 nt long. Reliable quantitative PCR for fragments longer than 500 nt is difficult with current techniques. We therefore resorted to performing a digital PCR. First, the DNA was serially diluted and distributed in 96-well plates so that approximately 25% of the wells produced an amplicon. The appropriate quantity of sperm DNA was determined by trial experiments for each of the two PCR assays: 50 ng of sperm DNA per well (corresponding to approximately 17,000 copies of chromosome 12, assuming one haploid genome represents 3 pg of DNA) were used for the PCR specific of the *DPY19L2* deletion, and 100 ng per well ($\sim 33,000$ copies under the same assumption) were used for the duplication-specific PCR.

For example, for donor A the deletion-specific PCR produced 26 positive wells. The deletion recombination frequency λ and its 95% confidence interval were then calculated as described (see Methods), resulting in a rate of *de novo* *DPY19L2* deletion for donor A estimated at 1.9×10^{-5} (95% CI: 1.3×10^{-5} ; 2.7×10^{-5}). Similarly, the duplication-specific PCR for donor A produced 23 positive wells, but because there was twice as much starting DNA this results in a rate of *de novo* *DPY19L2* duplication estimated at 8.1×10^{-6} (95% CI: 5.3×10^{-6} ; 1.2×10^{-5}) for this donor (Table 1 and Figure 2).

When pooling the results from the three sperm donors, more robust estimates are obtained: the *de novo* *DPY19L2* deletion rate is estimated at 1.8×10^{-5} (95% CI: 1.4×10^{-5} ; 2.2×10^{-5}), while the *de novo* duplication rate is estimated at 7.7×10^{-6} (95% CI: 6.1×10^{-6} ; 9.7×10^{-6}) (Table 1). There is a significant approximately two-fold enrichment of deletions over duplications at the *DPY19L2* NAHR hotspot.

We investigated whether differential amplification efficiency between the deletion and duplication assays could explain the observed difference between deletion and duplication *de novo* rates. To this end, we performed a control experiment as described (see Methods). No significant difference in amplification efficiency was observed: the deletion-specific control PCR amplified 37 wells, and the duplication-specific PCR amplified 40 wells.

Precise localisation of the recombined allele's breakpoints

Amplification of the LCRs in the deleted alleles had not been achieved in our previous study and the breakpoint minimal region had only been narrowed down to a 15 Kb region within the LCRs (8). Here we designed and validated PCR primers that amplify a 2 Kb product in deleted individuals only (Figure 1B). We quickly realised that mapping the breakpoints was complicated by the fact that many of the nucleotides that differed between LCR1 and LCR2 in the reference sequence were in fact not specific to one or the other LCR. Since mapping the breakpoints requires markers specific to each LCR, we decided to amplify and sequence the 2 Kb breakpoint region for each LCR in 20 control individuals.

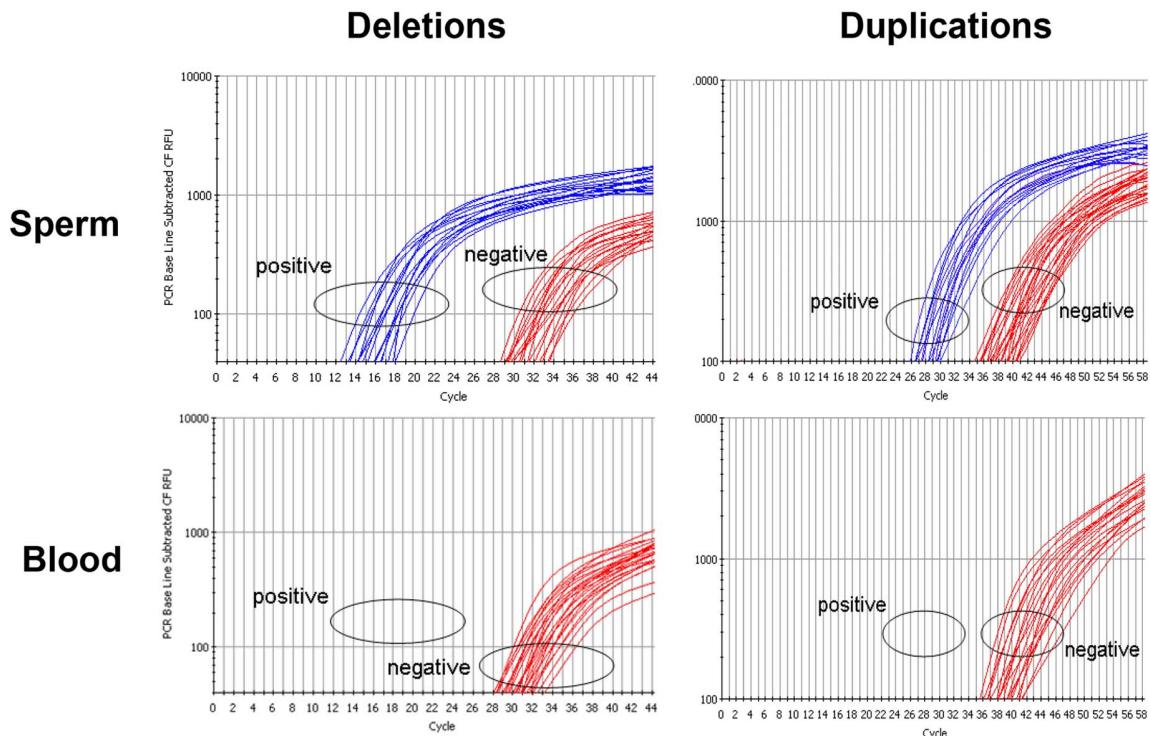
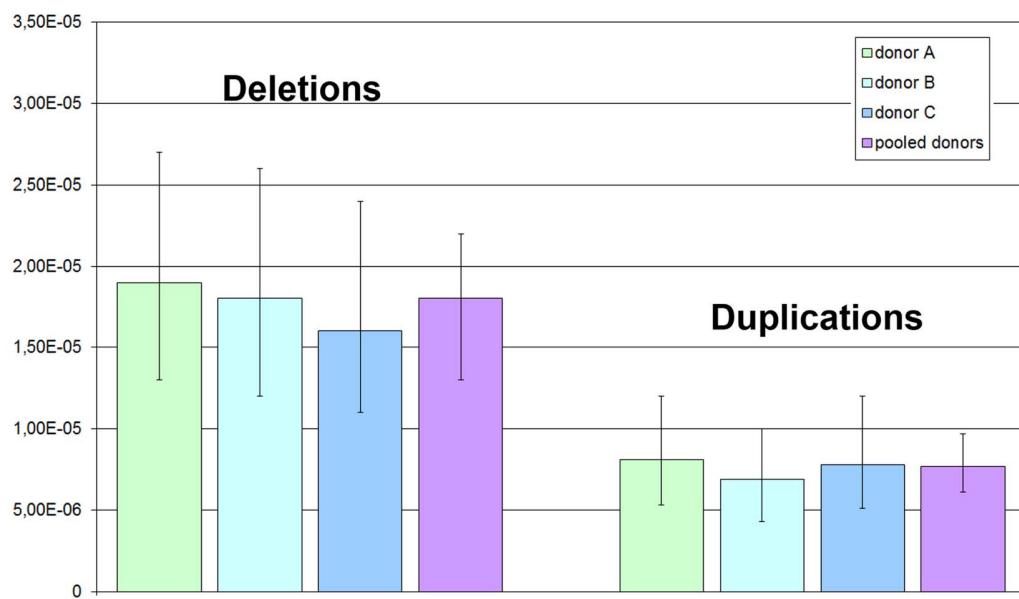
A.**B.**

Figure 2. Rate of *de novo* deletion and duplication events occurring at the *DPY19L2* NAHR hotspot determined by digital PCR on sperm from 3 control donors. (A) Illustration of PCR results obtained by real time PCR. The left plots show amplification profiles obtained with primers specific to the recombinant deleted LCR, the right plots show profiles obtained with the duplication-specific primers. No amplification was observed with either pairs of primers from 200 ng of somatic (blood) DNA, indicating that the NAHR did not occur during mitosis. Sperm DNA was diluted in order to obtain a positive amplification in approximately 25% of the wells. (B) The number of positive wells allowed estimating the frequency of *de novo* deletion and duplication events in three control sperms. Error bars represent 95% CIs.

Table 1. Frequency of deleted and duplicated alleles in sperm from three control donors.

	Deletion				Duplication			
	donor A	donor B	donor C	Pooled	donor A	donor B	donor C	Pooled
Positive wells	26	25	23	74	23	20	22	65
Nb of recombinants	30	29	26	85	26	22	25	74
Total nb of alleles		1.6E+6		4.8E+6		3.2E+6		9.6E+6
λ	1.9E-5	1.8E-5	1.6E-5	1.8E-5	8.1E-6	6.9E-6	7.8E-6	7.7E-6
95% CI inf	1.3E-5	1.2E-5	1.1E-5	1.4E-5	5.3E-6	4.3E-6	5.1E-6	6.1E-6
95% CI sup	2.7E-5	2.6E-5	2.4E-5	2.2E-5	1.2E-5	1.0E-5	1.2E-5	9.7E-6

doi:10.1371/journal.pgen.1003363.t001

To achieve the specific amplification of LCR 1 and 2 we had to rely on the reference human genome sequence to design the primers. We had no way of confirming that the targeted LCRs were specifically amplified in control individuals, but no amplification was obtained when assaying twenty homozygous deleted patients, vouching for the specificity of the primers. We then amplified and sequenced LCR1 and 2 from a total of 20 control individuals: 10 of North African origin and 10 of European origin. Thirty-four nucleotides were indicated as specific to either LCR 1 or 2 in hg19 reference sequence but 14 of these were in fact arbitrarily found in the two LCRs (Table S4): we consider that these are non-LCR-specific single nucleotide polymorphisms (SNPs). The remaining 20 nucleotides were indeed LCR-specific: these 20 fixed markers were used to map the recombination breakpoints, and we used the 14 SNPs to establish a haplotype map of the patients' deleted alleles (Table S4).

Allele-specific amplification of the deleted LCR was carried out on 15 homozygotously deleted globozoospermia patients. Each amplification yielded a single 2088 bp product, while the PCR was negative for all the healthy controls tested ($n = 20$). We sequenced all the amplicons in order to better characterize the breakpoint region. Fourteen out of the 15 patients analysed were homozygous for all markers tested. Three different breakpoints (BPs) were identified based on the presence of the 20 invariant markers. The three recombination events (BP1–3) were included in a 1153 bp maximal region (Table S4 and Figure 3). The breakpoints could not be mapped more accurately for lack of nucleotides specific to each LCR. One patient was heterozygous for markers 13 and 14, indicating that this patient was heterozygous and carried two different deleted alleles (BPs 2 and 3). If we consider that the other 14 patients carried two recombined deleted alleles each, we have a total of 14 alleles with BP1 (between markers 17 and 18), 13 alleles with BP2 (between markers 18 and 24) and 3 with BP3 (between markers 25 and 28) (Figure 3B). The 14 identified SNPs were then used to map the different haplotypes in patients presenting the same breakpoint (Table S4). This shows the presence of a total of 7 distinct haplotypes, indicating that at least 7 recombination events are at the origin of our patients' pathology (15 patients). We also observe that 5 patients with BP2 have the same haplotype and that two groups of 3 patients with BP1 have the same haplotype, suggesting the presence of several founding deletions in our patients' population. This is not surprising as all our patients came from the same region (Tunis area) and a majority had related parents (often first cousins).

One and three deletions were identified respectively in the 300 individuals analysed by PCR and in the 1699 Grenoble-Lyon array CGH patients group. There were 3 occurrences of BP2 and 1 of BP3. Overall, including the globozoospermia patients, a total

of 34 somatic deleted alleles were examined, resulting in the detection of three different recombination breakpoints. Fourteen alleles (41.2%) had a deletion between markers 17 and 18 (BP1), 16 alleles (47.0%) between markers 18 and 24 (BP2), and 4 alleles (11.8%) were recombined between markers 25 and 28 (BP3) (Figure 4 top left).

Two and fifteen genomic duplicated alleles were detected respectively in the 300 control individuals analysed by PCR and in the Grenoble-Lyon array CGH patients. Only 12 duplicated alleles could be sequenced (for lack of DNA from 3 control subjects and because two of the subjects had breakpoints falling outside the range of the duplication-specific PCR). Seven alleles (58.3%) corresponded to the reciprocal alleles of deletion 2 (BP2) with a recombination between markers 18 and 24, and 5 alleles (41.70%) corresponded to the reciprocal alleles of deletion 3 (BP3) with a recombination between markers 25 and 28 (Figure 3B).

The position of the meiotic recombination events (deletion and duplication) obtained from three sperm donors were also characterized by DNA sequencing. A total of 74 *de novo* deleted alleles and 65 *de novo* duplicated alleles were sequenced. All recombination events (from both duplications and deletions) clustered into five breakpoints (Figure 4). Two of them are new (BP4 and BP5) i.e. not previously identified in globozoospermic patients or in the CGH control cohort. The number and percentages of deleted and duplicated breakpoints respectively are: BP1: 2 (2.7%) and 4 (6.1%); BP2: 56 (75.7%) and 38 (58.5%); BP3: 10 (13.5%) and 13 (20%); BP4: 2 (2.7%) and 3 (4.6%) and BP5: 4 (5.4%) and 7 (10.8%) (Figure 4). BP2 is by far the most frequent BP, followed by BP3, explained by the fact that these two breakpoints correspond to the largest regions. Interestingly in sperm, the distributions of the deleted and duplicated breakpoints are quite similar. This is logical as the duplicated alleles are expected to be the reciprocal alleles of some of the deleted alleles. In genomic DNA the correlation is not as good, and we note that the frequency of the deleted BP1 is particularly high. Most of the deleted alleles come from globozoospermia patients (and a few detected in CGHarray patients) most of whom were recruited in Tunis. As suggested by the shared haplotypes observed between some deleted patients (Table S4) a founder's effect is likely to account for some of the most frequent deletions, in particular BP1.

PRDM9 genotyping of the sperm donors

Sequencing of the PRDM9 ZF array was performed in the 3 sperm donors. All three donors were homozygous for the A allele which represents over 90% of the European alleles. It comprises 13 copies of the 84-bp ZF repeat that binds the 13-bp Myers recombination motif [12,14]. This result is concordant with the ethnicity of the donors.

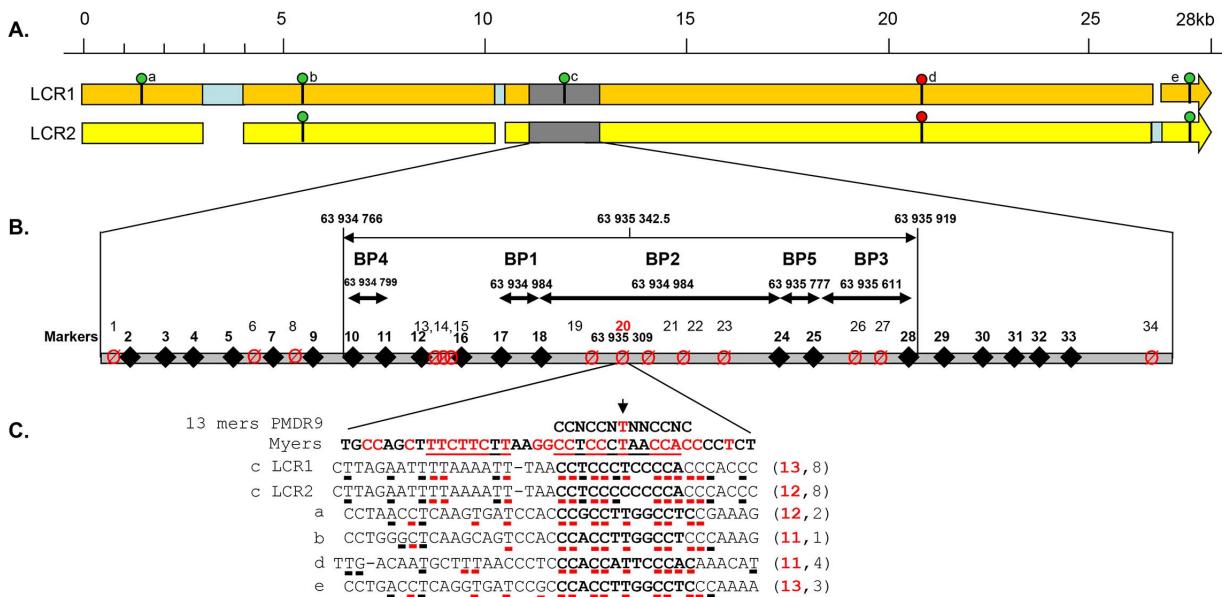


Figure 3. Details of the *DPY19L2* LCR1 and 2 and of the NAHR hotspot. (A) Detailed scaled representation of the 28.2 Kb LCR 1 (orange) and 28 Kb LCR2 (yellow). Pale blue rectangles correspond to sequences specific to one of the LCRs facing a gap in the other LCR. The presence of a 13 bp consensus PRDM9 recognition site (CCNCNTNNCCNC) on LCR1 or LCR2 is indicated by a green circle when identified on the forward DNA strand and by a red circle when identified on the reverse strand (GTGGNNAGGGTGG). The LCR arrows point toward the chromosome 12 telomere. (B) The analysed recombination region is represented in grey. The positions of LCR-specific markers (diamonds and bold numbering) and variable nucleotides (crossed circles) are represented. Details of the markers' sequences and localisations are indicated in Table S2. The five identified breakpoints (BP1-BP5) are shown as double arrows. One PRDM9 consensus sequence is localised in the centre of BP2, the central and most frequent breakpoint. (C) The central nucleotide from the consensus sequence corresponds to one of the identified SNPs (snp 20). A perfect match for the consensus sequence is present on LCR1, while the central thimine is replaced by a cytidine in LCR2. The 39 nt surrounding the 5 matches to the PRDM9 consensus sequence identified in LCR1 and 2 (sites a-e) are compared with the consensus sequence described in Myers et al [8,11]. Highly conserved nucleotides are red. For each locus the number of nucleotides identical to the consensus sequence is indicated on the right.
doi:10.1371/journal.pgen.1003363.g003

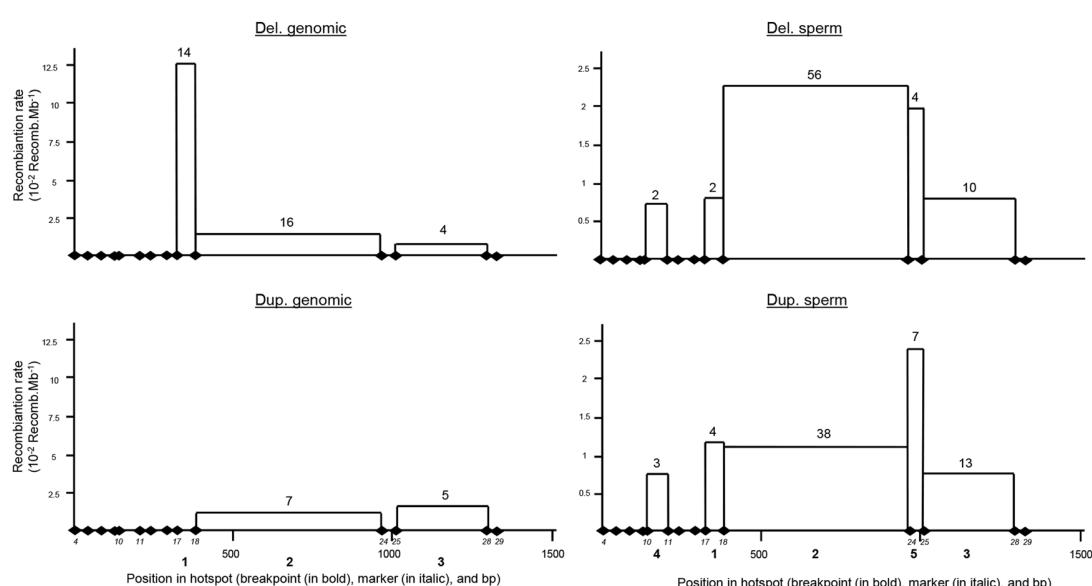


Figure 4. Distribution of deleted and duplicated breakpoints observed from somatic DNA (left two panels) and sperm DNA (right two panels). Somatic deletions were identified from sequence analysis of 15 homozygous deleted patients and two heterozygous deleted control individuals. Somatic duplications were identified from 12 positive control individuals. Data from sperm were pooled from three control donors.
doi:10.1371/journal.pgen.1003363.g004

Detailed analysis of LCR1 and 2

A comparison of the two LCRs is presented in Figure 3A. The illustration was produced from the results of a megablast search (<http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM>). All identified recombined alleles ($n = 185$) cluster between markers 10 and 28 within a 1153 bp region. This recombination hotspot is roughly located in the middle of the 28 Kb LCR (Figure 3A). Five 13 bp PRDM9 consensus recognition sites (CCNCCNTNNCCNC) are present along the LCR (Figure 3A). One of these sites is located in the centre of the 1153 bp hotspot (less than 35 nt away from the hotspot median position) (Figure 3B). We note that the most central BP (BP2) which encompasses the 13 bp site, represents 117 out of 185 recombined alleles or 63% of the detected recombined alleles (Figure 4). Given that five PRDM9 consensus recognition sites are found within the 28 Kb LCR1 sequence, the probability that a site would occur by chance less than 35 bp away from the centre of the hotspot is $1 - (1 - 5/28000)^{70} = 0.012$.

The Thymine at the centre of the consensus recognition site (CCNCCNTNNCCNC) was present only in the reference sequence of LCR1. Sequence analysis of our control individuals showed that this nucleotide was in fact a SNP (Marker 20 in Figure 3 and Table S4) with a T allele frequently found in both LCR1 and LCR2 (Table S4). In our globozoospermia patients we observed that all patients with the BP1 (with marker 20 located after the breakpoint thus on LCR2 sequence) have the T allele, indicating the presence of a T allele on LCR2 of the original unrecombined allele (Table S4). Conversely patients with BP3 (with marker 20 located before the breakpoint thus on LCR1 sequence) have the C allele indicating the presence of a C allele on LCR1 of the original unrecombined allele. All patients with BP2 have the C allele. As marker 20 is located within the breakpoint maximal sequence we can only conclude that at least a C allele was present on either LCR1 or LCR2 of the original unrecombined allele.

We sequenced LCR1 and LCR2 of our three sperm donors and realised that all were homozygous for the C allele at both LCR1 and LCR2, suggesting that the presence of the thymine in the CCNCCNTNNCCNC consensus sequence is not necessary to initiate recombination in the *DPY19L2* LCR central region. Myers et al. [8,11] indicated that although the core 13-mer recognition sequence was associated with recombination hotspots, the recognition motif extended beyond the core sequence with preferentially associated nucleotides identified within a 39 bp sequence encompassing the PRDM9 core sequence. We therefore aligned this extended motif with the sequence of the 5 PRDM9 motifs identified within the LCR (Figure 3C). We observe a good correlation within all 5 sequences, especially for the nucleotides that had been shown to be significantly associated with hotspots (indicated in red in Figure 3C). We also observe that the sequence central to our recombination hotspot (motif c) presents the highest homology (53%) with Myers' extended recognition sequence (Figure 3C).

Discussion

It appears paradoxical that *de novo* deletions are produced twice more frequently than *de novo* duplications during meiosis, while duplicated alleles are three times more frequent than deleted alleles in the general population. We investigated whether this could be explained parsimoniously through the combined effects of selection and mutation. Men carrying a homozygous deletion of *DPY19L2* are 100% infertile, but currently there is no evidence that a heterozygous deletion of *DPY19L2* causes a phenotype or that homozygous women are affected. Additionally, the deleted

allele is rare. Under these assumptions, according to the General Selection Model (GSM), natural selection results in a decrease in the frequency of the deleted allele of approximately $q^2/2$ per generation, where q is the frequency of the deleted allele (see Methods). Given that the deleted allele has a frequency of 1.7×10^{-3} (95% CI: 1.2×10^{-3} ; 2.5×10^{-3}) in the general population according to our combined control data, the GSM predicts that this frequency decreases by 1.5×10^{-6} (95% CI: 7×10^{-7} ; 3.1×10^{-6}) per generation. Conversely, deleted alleles are produced *de novo* by NAHR at an estimated rate of 1.8×10^{-5} (95% CI: 1.4×10^{-5} ; 2.2×10^{-5}) according to our digital PCR data. Assuming the allele frequency is at an equilibrium, these two rates should balance out. In fact they are somewhat similar but the 95% confidence intervals do not overlap. However the CIs only represent the uncertainty induced by the sampling procedure, i.e. the fact that the allele frequency and recombination rate are estimated from a sample of the whole population: they do not take into account experimental biases or imperfections that may exist at various steps. In addition, the GSM is a theoretical model that assumes an infinite population size and panmixia, whereas in practice stochastic effects and population structure (including for example any potential consanguinity or local founder effects) come into play. These could result in a significantly increased impact of purifying selection on the deleted allele, so that the frequency decrease resulting from selection and the *de novo* production of deleted alleles through NAHR may in fact cancel out.

Alternatively, it is possible that heterozygously deleted men suffer a fitness penalty. This can be taken into account within the GSM, and one can calculate the relative fitness of heterozygous individuals such that the GSM-predicted decrease of the deleted allele's frequency compensates the measured NAHR-induced production of new deleted alleles. In fact, assuming women are not affected, a 98% relative fitness of heterozygous men is sufficient (see Methods). Such a small effect could have easily remained undetected, and this scenario cannot be ruled out. This potential selection could be caused by meiotic segregation distortion as was observed for the T/t mouse locus [34]. Finally we only studied the recombination rate in male germ cells and we cannot exclude the possibility that the frequency and ratio of deletion and duplication might be different in female gametes.

All in all we believe the rates are reconcilable: whether the discrepancy observed when assuming heterozygous individuals have no phenotype is due to imperfections in the data and/or to population structure which disrupts the theoretical GSM model, or whether heterozygously deleted men suffer a small fitness penalty, we propose that the frequency decrease due to purifying selection and the *de novo* production of deleted alleles through NAHR cancel out, and that the frequency of the deleted *DPY19L2* allele is today at a selection-recombination equilibrium in the population. On the other hand, to the best of our knowledge there is no evidence that the duplicated *DPY19L2* allele is either deleterious or advantageous. We therefore assume that the duplicated *DPY19L2* allele is not under selection, so its frequency can increase in the population by recurrent NAHR. This resolves the paradox.

Liu and colleagues (2011) proposed that the frequency of NAHR occurring between two paralogous LCRs was proportional to the LCR length and sequence homology but inversely proportional to the distance between the LCRs [6]. The authors logically proposed that the probability of ectopic chromosome synapsis increases with LCR length, and that ectopic synapsis is a necessary precursor to ectopic crossing-over. Here we measured that the average rate of *de novo* recombination (deletion plus duplication) by NAHR at the *DPY19L2* recombination hotspot was 2.6×10^{-5} . This rate is higher than what was measured at

other loci such as the Williams-Beuren syndrome (WBS) locus or the LCR17p locus [4]. In our case the relatively small LCR size (28 Kb) is compensated by the proximity of the repeats (200 Kb) compared with much greater distances separating the paralogous LCRs for WBS and LCR17p. *DPY19L2* LCR1 and 2 also present a very high sequence identity (98%) which could also reinforce their synapsis and recombination. Our results are in agreement with previous work suggesting that the distance separating the two LCRs, as well as their sequence homology and length are parameters likely influencing recombination frequency.

We observed that >90% of *DPY19L2* NAHR events occurred within a 1.2 Kb region located in the centre of the 28 Kb LCR, suggesting the presence of a pro-recombination sequence within this hotspot. Myers et al. (2008) have characterized a degenerate 13 bp sequence motif that is present in approximately 40% of the identified human hotspots and which constitutes a *PRDM9* recognition signal [12–14]. *PRDM9* codes for a zinc finger array which catalyses the trimethylation of the lysine 4 of histone H3 (H3K4me3) [11]. This *PRDM9*-mediated post-translational histone modification likely initiates the recruitment of the recombination initiation complex, creating a favourable chromatin environment and allowing access of SPO11 to the DNA. SPO11 then initiates the formation of double-strand breaks (DSBs) which will be repaired by homologous recombination [35]. Here we identified a hotspot of NAHR located in the centre of a 28 Kb LCR. We showed that a *PRDM9* 13-mer recognition sequence is present at the epicentre of all the identified breakpoints. We however realised that the thymine, central to the 13-mer motif (CCNCCNTNNCCNC), was a T/C SNP, each nucleotide being found arbitrarily within LCR1 or LCR2. Following this observation one can wonder if recombination events at the *DPY19L2* hotspot occur preferentially in the presence of fully matching *PRDM9* 13-mer alleles. We measured the frequency of *de novo* recombination in sperm from three donors. As it happens, sequencing revealed that all three were homozygous for the C allele on both LCR1 and LCR2. This indicates that, at this locus, the presence of the 13-mer exact match is not necessary to initiate recombination. This observation is concordant with what was described previously at different loci and confirms that *PRDM9* tropism for the 13-mer recognition site might not be very strong and/or that other mechanisms also intervene in the choosing of double strand break localization [12,15]. One explanation can come from the extended sequence surrounding the 13-mer motif. Myers and colleagues (2008) [10] described a 39 bp pro-recombination sequence encompassing the 13-mer motif. We observe a greater than 50% sequence identity for the complete 39-mer sequence, indicating that a good match to the extended motif might be at least as important as a perfect match of the core 13-mer motif.

We identified a total of 5 distinct breakpoints (BP), all localized within a 1.2 Kb region located in the centre of the 28 Kb LCR. Others have described the localization of the deletions of globozoospermia patients [22]. They described a total of 9 separate BPs in the *DPY19L2* LCR. Looking at the precise localizations of the described BPs, we noticed that the nucleotides used to delimit BPs 1–6 in that study are in fact nucleotides that we identified as SNPs (markers 19–23 and 26), which strongly questions the validity of the BP localization in that study. Reanalyzing the presented data and using LCR-specific markers only, we conclude that Elinati et al. (2012) BPs 1, 2, 4, 5, 6 fall within the boundaries of “our” BP2 and that “their” BP3 corresponds to “our” BP3. This illustrates the difficulty in precisely identifying the localization of BPs and demonstrates that this can only be achieved with a high level of confidence after confirmation

that the markers used to define the BP positions are indeed locus-specific. From our reanalysis, Elinati et al. (2012) identified deletions in 27 globozoospermia patients, 23 had our BP2, one had BP3 and one had a BP that fell just outside of our studied region. These results thus confirm the importance of the recombination hotspot described here. Two additional BPs (BP8 and 9) were also identified in Elinati’s study which fell well outside of our recombination hotspot. This might constitute a second, less frequent recombination hotspot within the LCRs. We noticed that these two BPs are located 1200 bp telomeric from the 13-mer *PRDM9* site d (as indicated in Figure 3A). Thus this second putative hotspot is further away from a consensus 13-mer motif than our hotspot (the greatest distance of the BPs we identified from the 13-mer is 600 bp), but we can question again the accuracy of the positioning of these two breakpoints. Here, while analyzing the array CGH recombined patients we identified two recombined alleles which did not fall within our studied BP area. It is possible that these recombination events are also located within this second putative hotspot.

With the *DPY19L2* locus we believe that we have a good model to study the effect of the *PRDM9* recognition site on NAHR. We plan to accurately position the yet uncharacterized BPs in relation to other *PRDM9* sites. We are also currently screening an anonymized sperm bank to identify donors that are homozygous for the central 13-mer *PRDM9* recognition T allele and/or who present rarer *PRDM9* alleles to investigate how the recombination rate is affected by both the *PRDM9* genotype and the extended *PRDM9* recognition motif. We believe that although much work remains to be done, our study illustrates and consolidates the hotspot models described previously. In a moving environment we can imagine that the central region of the LCR will have the most opportunities to synapse with its paralogous sequence. The presence of an extended *PRDM9* recognition motif in the centre of the LCR then very likely contributes to DSB and NAHR. The combination of these parameters therefore probably explains why approximately 90% of the breakpoints occurred within a few hundred nucleotides from the most centrally located *PRDM9* recognition site.

Materials and Methods

Ethics statement

All patients, family members and anonymous DNA and sperm donors gave their written informed consent, and all national laws and regulations were respected. Ethical approval was obtained from Grenoble CHU review board.

Information on patients and control individuals

We previously reported that 15 out of 20 patients with globozoospermia had a homozygous deletion of the *DPY19L2* region [17]. These patients are included in this study. All patients are unrelated apart from two who are brothers. All patients originated from North Africa (Tunisia, n = 12; Morocco, n = 2 and Algeria, n = 1).

Array CGH data from a total of 1699 control anonymous individuals were re-analysed. These analyses had been carried out as a diagnosis for syndromic mental retardation either at Grenoble or Lyon’s hospital. As our aim was to identify *DPY19L2* centred CNVs in this cohort of patients and since there is no known link between *DPY19L2* and mental retardation, we believe that this cohort can serve as a control in this study. All individuals agreed to the anonymous use of their DNA in genetic studies and signed an informed consent. The fertility and ethnic origin of these individuals was not documented. All were French citizens. We

estimate that in excess of 90% of these individuals are of European origin and that the vast majority of the others are of North African origin.

There was no gender selection but this cohort contained approximately 2/3rd of males.

Array CGH results from these patients were scrutinized for the *DPY19L2* region.

Three hundred control individuals were analysed independently with recombinant *DPY19L2*-specific PCR (deleted and duplicated) to identify deleted and duplicated alleles. One hundred and fifty individuals originated from North Africa (Algeria, Morocco, and Tunisia) and 150 originated from Europe. All individuals gave their informed consent to constitute an anonymous DNA bank. Non-recombined LCR1 and 2 of twenty of these individuals were amplified and sequenced to identify LCR-specific SNPs. There was no gender selection and this cohort contained a similar number of males and females.

Lastly the *DPY19L2* CNV was also analysed from array CGH data available from web servers [29–33] for a total of 6575 control individuals, mainly from the Database of Genomic Variants (<http://projects.tcag.ca/variation/>). Most of these individuals originated from Europe (75%), Africa (18%) or Asia. Individual CNV could however not be linked to a particular individuals and its geographical origin. The location of the breakpoints of each CNV located in the *DPY19L2* locus was scrutinised to establish if they were located within the LCR and hence were caused by NAHR (Tables S2 and S3).

DNA extraction

Genomic DNA was extracted either from peripheral blood leucocytes using a guanidium chloride extraction procedure [36] or from saliva using Oragene DNA Self-Collection Kit (DNAgentech, Ottawa, Canada).

Sperm DNA was extracted from 2 ml of semen which were transferred to a 25 ml Falcon Tube (BD Biosciences). Ten ml of PBS was added, mixed gently and centrifuged at 3,000 rpm for 5 minutes. Supernatant was discarded and the pellet was resuspended again in 10 ml of PBS, mixed and centrifuged as before. Pellets were then resuspended in 1 ml digestion buffer (NTE buffer 0.5 mM NaCl, 10 mM Tris-HCl pH 7.5, 5 mM EDTA, pH 8 (100:10:1), 0.4% SDS), 25 µl of 10 mg/ml proteinase K solution (Sigma) were added and the mix was incubated overnight at 42°C with occasional mixing. Three hundred microliters of the contents of each Falcon tube were transferred into SafeLock tubes (Eppendorf). An equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) was added and mixed gently until emulsified. The tube was centrifuged at 3,000 rpm for 5 minutes. We repeated this process a second time, adding an equal volume of chloroform/isoamyl alcohol. The upper aqueous layer was transferred into a clean Eppendorf tube. The aqueous layers from the two phenol/chloroform extractions were combined and an ethanol precipitation was performed: 25 µl 3M sodium acetate pH 5.4 and 1 ml 100% ethanol were added to the aqueous phase, mixed gently and centrifuged as before. The pellets were washed twice with 70% ethanol and finally resuspended in 300 µl TE buffer (10 mM Tris-Cl pH 8.0, 0.1 mM EDTA, pH 8.0 (10:1)) by incubating overnight at 50°C with gentle shaking.

Information about the sperm donors

DNA was extracted from three fertile anonymous donors of European origin with normal sperm parameters and of similar age (between 30 and 35 years old). In each case a spermogram was realised according to WHO's 2010 guidelines [37]. Sperm

concentration ranged between 60–120 × 10⁶ spz/ml, with, in each case less than 1 × 10⁶ leucocytes/ml. We therefore considered that the presence of this small percentage of leucocytes had a negligible effect on the quantification of the sperm (only) DNA and on the ensuing calculations.

A molecular analysis was carried out to determine PRDM9 ZF array genotype. PCR amplification and sequencing of the PRDM9 ZF array were performed using primers and protocols as described previously ([12]. PCR and sequencing primer sequences are listed in the Table S1.

Amplification and sequencing of the LCRs

All primers were designed to have at least their 3' nucleotide specific to the LCR of interest (Table S1). PCR primers were designed to amplify specifically LCR 1 or LCR 2, in order to perform a sequence comparison of the two LCRs. For each recombined LCR locus (resulting from deletion or duplication), two sets of specific primers were designed (Figure 1B). The external primers (long primers) were used for sequencing analysis. They were also used as an outer primer for the digital PCR that was devised to measure the rate of *de novo* recombination in sperm. The short internal primers (SI) were used in duplex with *RYR2* primers that were used as a positive amplification control. These two sets of primers were used to detect the presence of recombined alleles in the 300 control individuals. They were also used as inner primers for the digital PCR.

PCR amplification was carried out on an Applied Biosystems genAmp 2700 thermocycler. Due to the high sequence homology between the two LCRs, the use of a precise annealing temperature was critical. The same thermocycler had to be used throughout the study as small variations in block temperature could introduce discrepancies in the amplification. Both the long and short PCR cycles were preceded by a 7 minutes denaturation at 95°C and followed by a 10 minutes elongation at 72°C. The specific annealing temperature of each primer set is indicated in Table S1. Thirty-five cycles were carried out for the long PCRs, with 30 seconds of denaturation at 95°C, 30 seconds of annealing and 2 minutes of elongation at 72°C. Forty-five cycles were carried out for the short PCRs with 30 seconds of denaturation at 95°C, 20 seconds of annealing and 2 minutes of elongation at 72°C.

We performed the long and short PCRs in 1× Takara Ex Taq buffer (Takara), 250 µM dNTPs (Takara dNTP mixture), 300 nM each primer, 1 unit Takara Ex Taq (Takara) with 200 ng of somatic DNA in a total volume of 25 µl.

All sequences (native LCR 1 and 2 and deleted and duplicated LCRs) were carried out with BigDye Terminator v3.1 (Applied Biosystems Courtaboeuf, France) on an ABI 3130XL (Applied Biosystems, Courtaboeuf, France).

Oligonucleotide array CGH was performed with the Agilent 105K or 180K Human Genome Microarray (Agilent Technologies, Santa Clara, CA, USA) (Hospices Civils de Lyon array CGH Platform and CHU Grenoble array CGH Platform). Extracted DNAs were labelled according to the instructions of the supplier and incubated overnight. The samples were purified and hybridised as described previously [17].

Graphical display and analysis of the data were performed with the Agilent DNA Analytics software version 4.0.81 (statistical algorithm: ADM-2, sensitivity threshold: 2.5, window: 0.5). A value of zero represents equal fluorescence intensities between sample and reference DNA. Copy-number losses shift the value to the left (≤ -1), and copy-number gains shift it to the right (≥ 0.58).

The design of the MLPA probes, MLPA reaction and data analysis were performed according to the recommendation of the

MRC-Holland synthetic protocol (www.mlpa.com) and as described in Coutton et al. (2012) [25].

Sperm assay design and digital PCR for sperm NAHR breakpoints mapping

We designed two nested LCR-specific PCRs as described in the PCR section. In addition, we designed a TaqMan dual labeled probe (Table S1) to allow the second step of the nested PCRs to be run on Biorad iCycler IQ real time PCR detection. We tested each recombinant-specific combination of primers for specificity and sensitivity on negative and positive (*DPY19L2* deleted and duplicated) control blood DNA (Figure 1B). Each of the two rearrangements was assayed on DNA extracted from three unrelated sperm donors. Each donor was confirmed to carry two copies of *DPY19L2* by MLPA analysis (data not shown). We note that our assay will not distinguish triplications of the *DPY19L2* locus, which are likely to occur at extremely low frequencies.

We performed the first LCR-specific PCR (long PCR) in 1× Takara Ex Taq buffer (Takara), 250 μM dNTPs (Takara dNTP mixture), 300 nM of each primer, 1 unit Takara Ex Taq (Takara), using sufficient copies of template DNA to give approximately 24 positive wells per 96-well plate (exact quantities determined empirically by successive dilutions) and 2.5 mM MgCl₂, in a total volume of 50 μl. Following thermal cycling we incubated 10 μl of the long PCR products with 5 μl of Exosap-IT PCR Clean-up Kit (GE Healthcare) for 15 min at 37°C to digest the long PCR primers followed by enzyme inactivation at 80°C for a further 15 min. Two μl of 10× diluted long PCR products was used as a template in the second PCR (short PCR). In the short PCR we used the same concentrations of buffer, dNTPs, primers and enzyme as in the Long PCR, but the total volume was 25 μl and we added a dual-labeled probe (final concentration 250 nM; Eurofins MWG Operon) (Table S1). To map the locations of breakpoints we re-amplified wells that we had previously identified as positive in the long PCR plate using the short primers and sequenced these amplicons.

The quantity of input sperm DNA was experimentally determined by serial dilutions to obtain approximately 24 positive breakpoint-specific amplifications per 96-well plate. The number of positive amplifications was then counted to estimate the number of recombinants in the input sperm. Each well contains a sample drawn from the input DNA without replacement, hence the number of recombinants in a given well is appropriately modeled using a hypergeometric distribution. We note that this hypergeometric distribution has often been approximated in the literature by Poisson (6, 22) or binomial (23) distributions, but although such approximations are acceptable we find no need for them in this study, as the direct calculation is simple. Indeed, using the hypergeometric distribution the probability that a well contains no recombinants is:

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

where N is the total number of copies of chromosome 12 in the input DNA (i.e. 1.6×10⁶ for the deletion assay and 3.2×10⁶ for the duplication assay, see Results section on digital PCR), W=N/96 is the number of copies per well, and R is the total number of recombinants. The value of R such that this probability is closest to the observed ratio of negative wells (i.e. one minus the fraction of wells that produced a positive amplification) is easily found by

tabulation. This leads to an estimation of the *de novo* recombination rate $\lambda=R/N$, and a 95% confidence interval is calculated by modeling the initial dilution to obtain the input DNA using the binomial distribution (with `binom.test` in the R stats package, <http://www.r-project.org>).

In order to evaluate the amplification efficiency of our duplication/deletion assays, we used as positive controls genomic DNA from one heterozygous duplicated individual and from one heterozygous deleted individuals. We believe that this type of control is more accurate than the use of cloned recombinant deleted and duplicated alleles as this reduces dilution factors. More importantly it reproduces faithfully the possible inhibitions due to the presence of the over majoritarian non-target genomic DNA or the potential amplification of homologous sequences that are present in the actual quantifying experiments.

The DNA concentration was measured by Nanodrop (Thermo-Scientific) and DNA quality was evaluated using an agarose gel electrophoresis (0.8%). No smear or fragments were observed. Considering that a human diploid genome represents 6 pg of DNA, we performed serial dilutions of the duplication and deletion controls to obtain a concentration of 1.5 pg/μl. One microliter of each solution was aliquoted in a 96-well plate, so that approximately 25% of the wells are expected to contain a recombinant allele (as we used heterozygous controls who carry only one copy of the deleted or duplicated alleles). The number of positive wells was then counted when amplifying deleted and duplicated DNA.

Calculations with the General Selection Model

Given a locus with two alleles (e.g. wild-type *DPY19L2* allele and deleted allele), and noting q the frequency of the minor (deleted) allele, the GSM predicts the change in allele frequency Δq at each generation given the relative fitness of each genotype. In our case the homozygous wild-type is used as a reference (fitness 1), and the homozygous deleted men are known to be 100% infertile while the deletion is considered to have no effect in women, hence the fitness of the homozygous deleted genotype is 0.5. Let W be the relative fitness of the heterozygous genotype, and p=1-q the frequency of the wild-type allele. Note that $q \approx 1.7 \times 10^{-3}$, hence $p \approx 998 \times 10^{-3}$. The GSM therefore simplifies to: $\Delta q \approx -q[W(q-p)+p-q]/2$.

In the first scenario, heterozygous individuals are assumed to have no phenotype, hence W=1 and the equation simplifies to: $\Delta q \approx -\frac{q^2}{2}$.

In the second scenario, we no longer assume W=1 and instead wish to calculate the value of W such that the GSM-predicted Δq exactly compensates the *de novo* rate of production of deleted alleles through NAHR, i.e. $\Delta q = -1.8 \times 10^{-5}$. Turning the previous

equation around, we obtain: $W \approx \frac{p + \frac{\Delta q}{2} - \frac{q}{2}}{p - q}$. Substituting the values of p, q and Δq, this yields W=0.99. Assuming that heterozygous women have no phenotype, we finally obtain a relative fitness of 98% for heterozygous males.

Supporting Information

Figure S1 Identification of a duplication of the *DPY19L2* locus by array CGH. Array-CGH analyses showed a 130 kb gain extending from base 63,947,732 to 64,078,229 in chromosome 12q14.2. Coordinates of variations or probes (y-axis) are based on the UCSC GRCh37/hg19 assembly. Graphical overview and analysis of the data were obtained with the Genomic Workbench

software, standard edition 6.5 (Agilent) with the following parameters: aberration algorithm ADM-2, threshold 6.0, fuzzy zero, centralisation and moving average window 0.5 Mb. The value of zero (x-axis) represents equal fluorescence intensity ratio between sample and reference DNA. Copy-number gains shift the ratio to the right (positive values). Three adjacent probes located at the *DPY19L2* locus are duplicated in the analyzed patient and the mean log₂ ratio was +0.53 according to the Alexa 5 deviation with a mirror image.

(GIF)

Table S1 Sequence of the PCR primers, position (Hg19), size of the amplified products (between brackets), hybridization temperature (Hyb.). The position of the primers is illustrated in Figure 1B. (DOC)

Table S2 Sequence of the recombination hotspot region of control subjects (10 Europeans and 10 North Africans) for the identification of LCR-specific markers and determination of the precise localisation of the breakpoints of 15 globozoospermia patients.

(DOC)

Table S3 Number and percentage (%) of recombined alleles with breakpoints located within (inside) or outside of the LCR. (DOC)

Table S4 Sequence of the recombination hotspot region of control subjects (10 Europeans and 10 North Africans) for the identification of LCR-specific markers and determination of the precise localisation of the breakpoints of 15 globozoospermia patients. The first column indicates the reference number of the identified variants as shown in Figure 2. Markers that are LCR-specific (i.e. homozygous and invariant within each LCR across all controls, but differing between LCR1 and LCR2) according to the results obtained from the 20 sequenced control individuals (columns 5–8) are indicated in larger bold lettering. Nucleotides that are not LCR-specific are considered as SNPs. The markers' sequence in each LCR according to the Hg19 reference sequence is indicated column 4. In patients, the presence of the Hg19

reference nucleotide is indicated by a cross. When a single nucleotide is detected, the patient is considered homozygous at that position. Because the rows are color-coded, with alternating grey and white rows corresponding to the Hg19 reference nucleotide for LCR1 or LCR2 respectively, and because bold crosses correspond to validated LCR-specific markers, the recombination breakpoints can be easily visualized. For each patient a vertical stretch of bold crosses in grey rows (displayed in orange rectangles) shows non-recombined genetic material coming from LCR1. This is followed by a stretch of bold crosses in white rows (displayed in yellow rectangles), which shows non-recombined DNA from LCR2. For each patient the breakpoint's localisation is inferred when his genotype shifts from LCR1- to LCR2-specific markers. Unboxed regions therefore correspond to breakpoint maximal regions. Patients 1–7 breakpoints are located between markers 17 and 18 (BP1). Patients 8–13 BPs are located between markers 18 and 24 (BP2). Patient 14 is the only heterozygous patient, with BP2 and a breakpoint between markers 25 and 28 (BP3). Patient 15 is homozygous for BP3. SNPs differing between patients with the same breakpoints are highlighted with a blue background. This indicates the presence of 3, 2 and 2 distinct haplotypes for BP1, BP2 and BP3 respectively. Overall this indicates the presence of 7 distinct haplotypes, so that the occurrence of at least 7 separate recombination events within our series of 15 patients can be inferred.

(XLS)

Acknowledgments

We thank Gaelle Vieville from Grenoble CHU and Audrey Labalme from the Hospices Civils de Lyon for their help with the array CGH analysis. We thank Daniel Turner from the Sanger institute for helpful discussions.

Author Contributions

Conceived and designed the experiments: PFR. Performed the experiments: CC FA TK. Analyzed the data: CC PFR VS JL P-SJ CA NT-M. Contributed reagents/materials/analysis tools: DS VS JL PFR. Wrote the paper: PFR CC NT-M.

References

- Gu W, Zhang F, Lupski JR (2008) Mechanisms for human genomic rearrangements. *Pathogenetics* 1: 4.
- Smith CE, Llorente B, Symington LS (2007) Template switching during break-induced replication. *Nature* 447: 102–5.
- Lupski JR, Stankiewicz P (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* 1: e49. doi:10.1371/journal.pgen.0010049
- Turner DJ, Miretti M, Rajan D, Fiegler H, Carter NP et al. (2008) Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat Genet* 40: 90–5.
- Molina O, Anton E, Vidal F, Blanco J (2011) Sperm rates of 7q11.23, 15q11q13 and 22q11.2 deletions and duplications: a FISH approach. *Hum Genet* 129: 35–44.
- Liu P, Lacaria M, Zhang F, Withers M, Hastings PJ et al. (2011) Frequency of nonallelic homologous recombination is correlated with length of homology: evidence that ectopic synapsis precedes ectopic crossing-over. *Am J Hum Genet* 89: 580–8.
- Arnheim N, Calabrese P, Tiemann-Boege I (2007) Mammalian meiotic recombination hot spots. *Annu Rev Genet* 41: 369–99.
- Paigen K, Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet* 11: 221–33.
- Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A et al. (2010) Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467: 1099–103.
- Myers S, Freeman C, Auton A, Donnelly P, McVean G (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat Genet* 40: 1124–9.
- Hayashi K, Yoshida K, Matsui Y (2005) A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. *Nature* 438: 374–8.
- Berg IL, Neumann R, Lam KW, Sarbajna S, Odenthal-Hesse L et al. (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat Genet* 42: 859–63.
- Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C et al. (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* 327: 876–9.
- Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C et al. (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327: 836–40.
- Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ et al. (2011) Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc Natl Acad Sci U S A* 108: 12378–83.
- Sasaki M, Lange J, Keeney S (2010) Genome destabilization by homologous recombination in the germ line. *Nat Rev Mol Cell Biol* 11: 182–95.
- Harbuz R, Zouari R, Pierre V, Ben Khelifa M, Kharouf M et al. (2011) A Recurrent Deletion of DPY19L2 Causes Infertility in Men by Blocking Sperm Head Elongation and Acrosome Formation. *Am J Hum Genet* 88: 351–61.
- Dam AH, Feenstra I, Westphal JR, Ramos L, van Golde RJ et al. (2007) Globozoospermia revisited. *Hum Reprod Update* 13: 63–75.
- Florke-Gerloff S, Topfer-Petersen E, Muller-Esterl W, Mansouri A, Schatz R et al. (1984) Biochemical and genetic investigation of round-headed spermatozoa in infertile men including two brothers and their father. *Andrologia* 16: 187–202.
- Kilani Z, Ismail R, Ghunaim S, Mohamed H, Hughes D et al. (2004) Evaluation and treatment of familial globozoospermia in five brothers. *Fertil Steril* 82: 1436–9.
- Nistal M, Herruzo A, Sanchez-Corral F (1978) [Absolute teratozoospermia in a family. Irregular microcephalic spermatozoa without acrosome]. *Andrologia* 10: 234–40.
- Dam AH, Koscienski I, Kremer JA, Moutou C, Jaeger AS et al. (2007) Homozygous mutation in SPATA16 is associated with male infertility in human globozoospermia. *Am J Hum Genet* 81: 813–20.
- Liu G, Shi QW, Lu GX (2010) A newly discovered mutation in PICK1 in a human with globozoospermia. *Asian J Androl* 12: 556–60.

24. Carson AR, Cheung J, Scherer SW (2006) Duplication and relocation of the functional DPY19L2 gene within low copy repeats. *BMC Genomics* 7: 45.
25. Coutton C, Zouari R, Abada F, Ben Khelifa M, Merdassi G et al. (2012) MLPA and sequence analysis of DPY19L2 reveals point mutations causing globozoospermia. *Hum Reprod* 27: 2549–58.
26. Elinati E, Kuentz P, Redin C, Jaber S, Vanden Meerschaut F et al. (2012) Globozoospermia is mainly due to DPY19L2 deletion via non-allelic homologous recombination involving two recombination hotspots. *Hum Mol Genet* 21: 3695–702.
27. Koscienski I, Elinati E, Fossard C, Redin C, Muller J et al. (2011) DPY19L2 deletion as a major cause of globozoospermia. *Am J Hum Genet* 88: 344–50.
28. Pierre V, Martinez G, Coutton C, Delaroche J, Yassine S et al. (2012) Absence of Dpy19l2, a new inner nuclear membrane protein, causes globozoospermia in mice by preventing the anchoring of the acrosome to the nucleus. *Development* 139: 2955–65.
29. Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H et al. (2009) High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* 19: 1682–90.
30. Pinto D, Marshall C, Feuk L, Scherer SW (2007) Copy-number variation in control population cohorts. *Hum Mol Genet* 16 Spec No. 2: R168–73.
31. Itsara A, Cooper GM, Baker C, Girirajan S, Li J et al. (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet* 84: 148–61.
32. de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA et al. (2007) Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. *Hum Mol Genet* 16: 2783–94.
33. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O et al. (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464: 704–12.
34. Schimenti J (2000) Segregation distortion of mouse t haplotypes the molecular basis emerges. *Trends Genet* 16: 240–3.
35. Grey C, Barthes P, Chauveau-Le Fréec G, Langa F, Baudat F et al. (2011) Mouse PRDM9 DNA-binding specificity determines sites of histone H3 lysine 4 trimethylation for initiation of meiotic recombination. *PLoS Biol* 9: e1001176. doi:10.1371/journal.pbio.1001176
36. Jeanpierre M (1987) A rapid method for the purification of DNA from blood. *Nucleic Acids Res* 15: 9611.
37. World Health Organization (2010) WHO Laboratory Manual for the Examination and Processing of Human Semen. 5th ed. Geneva: WHO Press.

Principaux résultats

Contrairement à ce que prédit la théorie de formation des NAHRs, les résultats extraits des bases de données publiques indiquent un excès d'allèles dupliqués de *DPY19L2* dans la population générale. Nous avons donc cherché à déterminer les fréquences des duplications et délétions *de novo* de ce même locus. Ceci ayant pour but de déterminer si cet excès est dû à une sélection de l'allèle dupliqué ou au fait que celui-ci était effectivement produit plus fréquemment que l'allèle délété. Pour ce faire nous avons quantifié le taux d'apparition de ces événements génétiques à partir d'ADN spermatique. Les spermatozoïdes étant le produit direct de la méiose, ils sont donc les reflets d'haplotypes produits *de novo*. Pour cela, nous avons analysé par PCR digitale l'ADN spermatique de trois donneurs ainsi que l'ADN spermatique constitué d'un mix provenant de ces trois donneurs. Leur ADN a tout d'abord été dilué en série afin qu'environ 25% des 96 puits de la PCR contiennent un événement (délétion ou duplication). Ainsi, en acceptant qu'un génome haploïde humain représente 3pg, 50ng d'ADN spermatique furent déposés dans chaque puit pour la PCR spécifique à la délétion, et 100ng dans chaque puit spécifique à la duplication. Chaque puit contient donc une partie de cette charge d'ADN initiale. La distribution de cette charge d'ADN au sein des 96 puits peut donc s'apparier à un tirage sans remise, la probabilité qu'un puit soit positif pour un événement chromosomal (duplication ou délétion) peut donc être modélisé par une loi hypergéométrique (**Équation : (3.1)**). Nous permettant ainsi d'estimer la fréquence duplication / délétion λ pour chaque donneur (**Équation : (3.2)**).

$$\frac{\frac{(N-R)!}{W!(N-R-W)!}}{\frac{N!}{W!(N-W)!}} = \frac{(N-R)!(N-W)!}{N!(N-W-R)!} = \prod_{i=0}^{R-1} \frac{N-W-i}{N-i} \quad (3.1)$$

$$\lambda = \frac{R}{N} \quad (3.2)$$

Où :

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

L'intervalle de confiance (IC) à 95% est ensuite calculé grâce à une loi binomiale de sorte à modéliser la dilution initiale pour obtenir l'ADN d'*entrée*. Le puit contenant le *pool* des trois ADN spermatiques est donc celui ayant les résultats les plus robustes, l'IC étant le plus resserré et permet donc d'établir le taux de délétion *de novo* à 1.8×10^{-5} (IC 95% : 1.4×10^{-6} ; 2.2×10^{-6}) tandis que le taux de duplication *de novo* est estimé à 7.7×10^{-6} (IC 95% : 6.1×10^{-6} ; 9.7×10^{-6}) montrant un enrichissement environ deux fois supérieur des délétions *de novo* par rapport aux duplications sur le site de *DPY19L2*.

Ainsi nous avons observé qu'au locus *DPY19L2* les délétions *de novo* apparaissent, au cours de la méiose, deux fois plus fréquemment que les duplications alors que l'allèle dupliqué est trois fois plus fréquent que l'allèle délété dans la population générale. Cet effet pourrait en partie être dû aux effets de sélection naturelle. En effet, Bien qu'à notre connaissance, les femmes portant l'allèle délété à l'état homozygote ne soient caractérisées par aucun phénotype, les hommes, eux sont 100% infertiles tandis que l'allèle dupliqué ne subirait aucune sélection.

Cette étude a également été pour notre équipe l'occasion d'effectuer une étude plus approfondie des LCRs flanquant le locus de *DPY19L2*. Pour cela, nous avons génotypé 20 SNPs spécifiques des LCRs télomériques et centromériques. À partir de ces données, 5 points de cassures distincts (BP1-5) ont pu être identifiés sur les 185 allèles recombinés étudiés (108 délétés et 77 dupliqués). L'ensemble de ces points de cassures sont localisés dans une région d'environ 1150 pb localisée au centre des 28kb du LCR. L'analyse bioinformatique de cette région a permis de mettre en évidence, au centre de la région minimale de recombinaison un site de reconnaissance consensus de la protéine PRDM9 (CCNCCNTNNCCNC). Cette protéine à doigts de zinc est connue pour son rôle central dans les mécanismes de recombinaisons homologue au cours de la méiose chez l'humain et la souris en ciblant de manière spécifique la localisation des cassures doubles brins, préambule nécessaire à toute recombinaison [227, 228]. On peut donc penser que la présence de cette séquence consensus de PRDM9 est la raison pour laquelle toutes les points de cassures observés sont localisés dans cette région minimale de 1150 nucléotide.

3.3 Résultat 2 : La transcriptomique

3.3.1 Article n°7 :

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

Karaouzène T , El Atifi M, Issartel JP, Grepillat M, Charles Coutton C, Martinez D, Arnoult C and Ray PF

Basic and Clinical Andrology, 2013

Contexte et objectifs

Dans des études précédentes, notre équipe à réussi à démontrer que, chez la souris, la protéine *Dpy19l2* était localisée dans la membrane interne des noyaux des spermatides pendant la spermatogenèse et qu'elle était nécessaire pour fixer l'acrosome au noyau [218]. Dans cette même étude, nous avons pu montrer cette protéine colocalisait avec la protéine *Sun5* et que *Dpy19l2* pourrait être un partenaire de *Sun5* [218]. Chez la souris, la protéine *Sun1* est elle aussi nécessaire à la gamétogenèse et est connue pour permettre l'interaction entre le noyau et les télomères [229]. Dans cette étude nous avons donc cherché à savoir si l'absence de la protéine *Dpy19l2* pouvait entraîner des dérèglements transcriptionnels qui pourraient, entre autre, expliquer l'absence de la protéine *Plcz1* dans les spermatozoïdes globozoocéphales murins.

De plus, au cours de production des souris *Dpy19l2* KO au sein de notre laboratoire nous avons pu observer un excès de naissance de souris mâle lorsque l'on croisait deux souris *Dpy19l2^{+/−}*. Ainsi, en comparant les sexes des souris obtenues lors de 6 premières naissances (*Birth 1-6*) on observe un total de 28 souris mâles pour 16 souris femelles. La p-valeur obtenue en effectuant un test de χ^2 comparant ces deux effectifs était égale à 0.0486272 laissant supposer l'existence d'un enrichissement réel, bien que faible, en souris mâles.

C'est donc afin d'expliquer l'absence de la protéine *Plcz1* dans les spermatozoïdes des souris *Dpy19l2^{−/−}* ainsi que l'enrichissement en souris mâle dans les naissances issues d'accouplement de souris *Dpy19l2^{+/−}* que nous avons effectué une analyse comparative du transcriptome testiculaire de deux souris *Dpy19l2^{+/+}* (*S1⁺* et *S2⁺*) et deux souris *Dpy19l2^{−/−}* (*S1[−]* et *S2[−]*) ayant pour but de mettre en évidence d'éventuels dérèglements transcriptionnels chez la souris KO.

Dans cette étude j'ai pu effectuer l'intégralité des manipulations de biologie moléculaire telles que la mise en place du protocole de génotypage des souris, l'extraction de l'ARN testiculaire de souris et l'analyse sur puce, ainsi que l'intégralité de l'analyse bioinformatique des résultats.

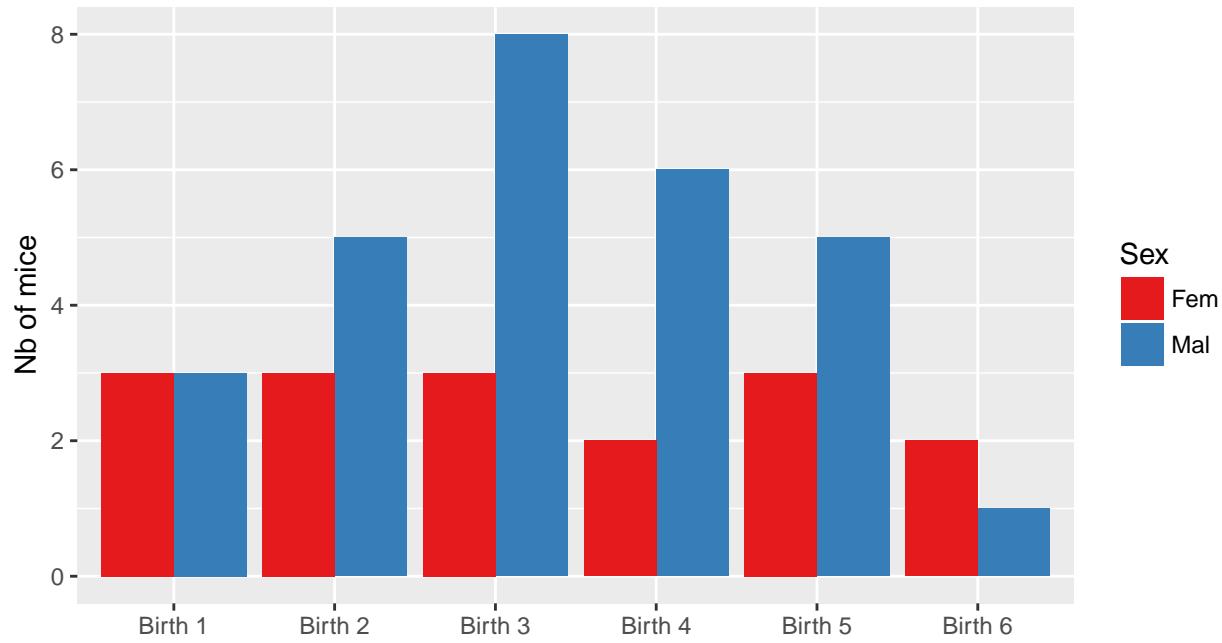


Figure 3.4 – Quantification des sexes des souris observés lors de chaque naissance issues d'un croisement de deux souris hétérozygotes Dpy19l2^{+/−} : Après 6 naissances, 28 souriceaux mâles sont nés pour seulement 16 femelles, ainsi le test du χ^2 comparant ces valeurs donne une p-valeur de 0.0486272.



RESEARCH ARTICLE

Open Access

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

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

Abstract

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

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

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

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

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

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

* Correspondence: pray@chu-grenoble.fr

¹Université Joseph Fourier, Grenoble F-38000, France

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

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

Résumé

Contexte: La globozoospermie est caractérisée par la présence dans l'éjaculat de près de 100% de spermatozoïdes ronds et dépourvus d'acrosome qui présentent un contenu chromosomique normal. L'injection intracytoplasmique (ICSI) de ces spermatozoïdes donne cependant un taux de fécondation et de développement embryonnaire particulièrement bas. Nous avons montré précédemment que la plupart des patients globozoospermes présentent une délétion homozygote de 200 Kb qui inclue la totalité de la séquence codante du gène *DPY19L2*. De plus nous avons montré que la protéine *DPY19L2* était localisée dans la membrane interne des noyaux des spermatides pendant la spermatogénèse et qu'elle est nécessaire pour fixer l'acrosome au noyau, réalisant ainsi une fonction similaire à celle des protéines Sun au sein du complexe LINC (Linker of Nucleoskeleton and Cytoskeleton). Il a par ailleurs été montré que SUN1 était nécessaire à la spermatogénèse et que cette protéine interagit avec les télomères chromosomiques. Il est donc possible que *Dpy19l2* interagisse également, directement ou indirectement avec l'ADN et module l'expression génique lors de la spermatogénèse. Dans cette étude nous avons donc comparé le transcriptome de testicules de souris invalidées (KO) pour le gène *Dpy19l2* à celui de souris sauvage afin d'identifier une éventuelle dérégulation génique qui pourrait expliquer le faible potentiel reproductif des spermatozoïdes globozoocéphales.

Méthode: L'ARN a été extrait de testicules de souris KO pour *Dpy19l2* et de souris sauvages. Le transcriptome a été réalisé en utilisant des puces d'expression ® Mouse Exon 1.0 ST Arrays. Les processus biologiques et les fonctions des gènes dérégulés ont été analysés en utilisant le logiciel PANTHER.

Résultats: Un total de 76 gènes a été identifié comme étant dérégulés, 70 gènes étaient surexprimés et 6 (incluant *Dpy19l2*) étaient sous-exprimés. Il s'agit de gènes principalement impliqués dans des interactions avec des acides nucléiques (ADN/ARN), ou ayant un rôle structural, dans le transport, ou présentant une activité catalytique.

Conclusions: Cette étude nous a permis d'identifier et de décrire un nombre important de gènes exprimés de manière différentielle chez les souris KO pour *Dpy19l2*. Ce travail peut permettre d'améliorer notre compréhension des fonctions de *Dpy19l2* et peut contribuer à obtenir une meilleure compréhension des mécanismes moléculaires nécessaire à la spermatogénèse.

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

Background

A recent study supported by the World Health Organization indicates than in 2010, an estimated 48.5 million couples worldwide were unable to have a child after five years [1]. Male factors are believed to be responsible for 30-50% of all infertility cases, but micro deletions of the Y chromosome are the only genetic defects altering human spermatogenesis, which are diagnosed routinely.

To be able to fertilize the oocyte, the spermatozoon needs to cross the zona pellucida (ZP), which is a glycoprotein layer surrounding the oocyte. The acrosomal reaction (AR), during which the acrosome (a giant vesicle of secretion) releases its content, plays an important role in the fertilization process. Enzymes released from the acrosome locally digest and soften the ZP so that the spermatozoon can penetrate deeper and fertilize the oocyte. The acrosome, a highly specialized organelle found only in sperm, is tightly bound to the nucleus via the acroplaxome (a network of proteins including keratin 5 and β-actin) [2].

Globozoospermia is a severe teratozoospermia characterized by the presence of 100% of round-headed spermatozoa devoid of acrosome. Men with globozoospermia have a primary infertility due to this absence of acrosome, which prevents their sperm from fertilizing the oocytes *in vivo* [3]. Spermatozoa from globozoospermic patients have near

normal levels of aneuploidy but give a poor fertilization rate and embryonic development even when performing Intra Cytoplasmic Sperm Injection (ICSI) [3]. Studies by immunocytochemistry showed that most round headed sperm lacked the phospholipase zeta protein (PLCzeta), a protein normally located around the sperm's head [4-7] and required to induce oocyte intracellular calcium oscillation and oocyte activation [8,9]. It has therefore been postulated that it is the absence of PLCzeta which might be responsible for the poor fertilization potential of round-headed spermatozoa [10]. In the course of this work we wanted to assess if the absence of PLCzeta in round-headed spermatozoa results from a transcriptional repression of the gene and if other transcriptional deregulations could also contribute to the poor fertilization potential of these gametes.

The syndrome of globozoospermia was first described in the seventies [7,11] and cases have been described regularly since [12-20]. Familial cases rapidly pointed to a genetic cause for this syndrome. In the recent years, *SPATA16* has been described to be involved in globozoospermia [21]. We demonstrated recently that *DPY19L2* was in fact the main locus associated with globozoospermia as 15 out of 20 analysed patients presented a 200 Kb homozygous deletion removing the entire gene [22]. We then identified

DPY19L2 point mutations and heterozygous deletions and demonstrated that 84% of the 31 globozoospermic patients analysed had a molecular alteration of DPY19L2 [23]. We finally confirmed that the recurrent deletion observed in a majority of men with globozoospermia was caused by non-allelic homologous recombination (NAHR), between two highly homologous sequences, or low-copy repeats (LCR), located on each side of DPY19L2 [24].

We previously characterized *Dpy19l2* Knockout mice (*Dpy19l2^{-/-}*) and showed that these mice present the same phenotype than men carrying mutations in *DPY19L2*, ie round-head spermatozoa without acrosome. It also permitted us to determine that i.) *DPY19L2* is located in the inner nuclear membrane of wild type mouse spermatids, ii.) *DPY19L2* is required for acrosome attachment to the nucleus and iii.) the detachment of the acrosome in *Dpy19l2^{-/-}* mice prevents correct anchoring of the manchette. Moreover we described that *SUN5* and *DPY19L2* partially colocalized in transfected HEK cells [25]. SUN-domain proteins are known to interact with chromosome-binding proteins and various KASH-domain partners to form SUN-domain-dependent 'bridges' across the inner and outer nuclear membranes. These bridges physically connect the nucleus to every major component of the cytoskeleton [26]. *SUN1*, one of the members of the family, was described to be necessary for gametogenesis and was shown to interact with the telomeres [27]. We can hypothesize that *Dpy19l2* could interact directly or indirectly with the DNA and thus have an effect on the regulation of transcription. It is thus possible that the absence of *Dpy19l2* could cause some modification in the germ cell transcription pattern.

The goal of this study was to determinate if *Dpy19l2* knock out mice present significant testis transcriptional modifications compared to wild type and in particular modifications that may explain the poor success rate encountered by globozoospemic patients following ICSI- IVF.

Methods

Ethical statement

Animal housing and sacrificing was in accordance with French guidelines on the use of animals in scientific investigations with the approval of the local Ethical Committee.

Animals

Dpy19l2 knock out mice were obtained from Mutant Mouse Regional Resource Center, University of California, Davis, CA. The mouse colony used in this study was initiated from two couples. The first one consisted of an heterozygous female and a wild type male. The second was composed of two heterozygous mice for the *Dpy19l2* deletion. Reproduction of these two couples achieved wild type, heterozygous and homozygous *Dpy19l2* deleted mice. Mice

were sacrificed at 2 months old, which means that they were pubescent and that their reproductive organs were fully established. A total of four animals were sacrificed. RNA was extracted from two homozygous WT and two homozygous KO animals.

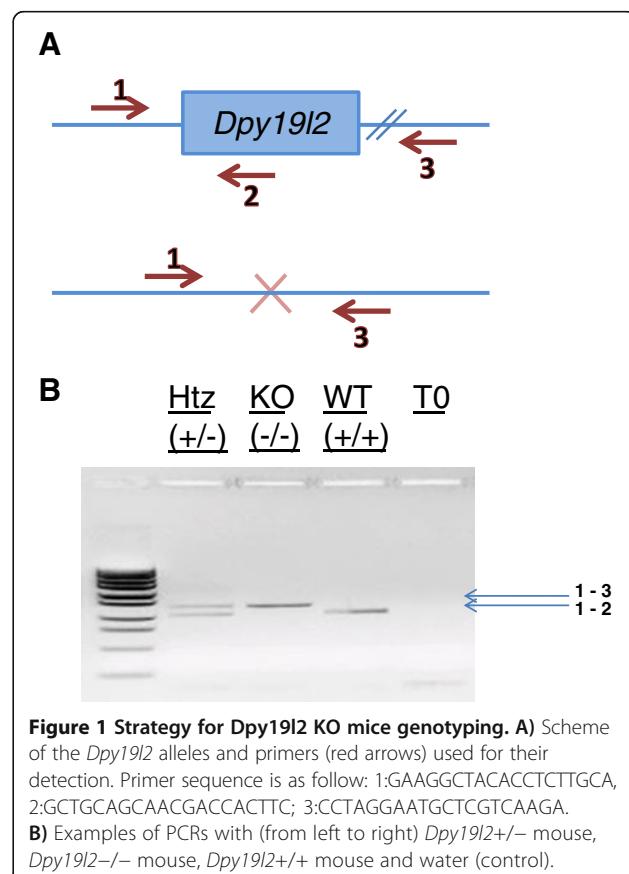
Genotyping PCRs

Genotyping was done on DNA isolated from tail biopsies. Tail biopsies (ca. 2 mm in length) were digested in 200 µl lysis Direct PCR Lysis Reagent (Tail) (Viagen Biotech inc, CA, USA) and 0,2 mg of proteinase K for 12–15 hours at 55°C and 1 hour at 85°C. The DNA was directly used for PCRs.

PCR was done for 35 cycles, with an annealing temperature of 57°C, and an elongation time of 60 seconds at 72°C. The primers used are described in Figure 1. PCRs products were separated on agarose gel electrophoresis. Genotypes were determined according to the migration pattern (Figure 1).

Tissue collection

Mice were sacrificed and testes were collected. Tissues were snap frozen in liquid nitrogen prior storage at -80°C. Two mice in each group were used for the micro-array analysis.



RNA extraction

Total RNA was extracted from tissues using mirVana isolation kit™ (Ambion, Applied Biosystems, Foster City, CA) as per the manufacturer's instructions. RNA purity and quantity was assessed using the NanoDrop c1000 (ThermoFisher Scientific, Waltham, USA). Quality was determined by both evaluation of the integrity of rRNA bands using RNA Nano 6000 kit (Bio-Analyser, Agilent Technologies, Palo Alto, CA) and absorption readings at 260 and 280 nm. For detail see Additional file 1: Table S1.

Array hybridization

For each group, two biological replicates were used. The replicates came from four separate RNA extractions: two from homozygous WT and two from homozygous KO animals. cDNA synthesis, amplification, enzymatic fragmentation and biotinylation were performed using the Ambion WT Expression Kit (Ambion, Austin, TX, USA). Samples were hybridized to Affymetrix GeneChip® Mouse Exon 1.0 ST Arrays as per the manufacturer's instructions. The Affymetrix Mouse Exon 1.0 ST array, contains probe sets for 35,557 genes. Briefly, 5 µg of fragmented biotinylated ssDNA was hybridised for 16 hrs at 45°C, 60 rpm to the array chip on a GeneChip® Hybridization Oven 640. After 16 hrs, GeneChips® were washed on a GeneChip® Fluidics station 450 using the washing script Prime 450 with buffers and stains supplied with the GeneChip Hybridisation Wash and Stain Kit from Affymetrix.

Data acquisition and analysis

Data was acquired on a GeneChip® Scanner 3000 7G and .CEL file generation performed using AGCC. Expression Console with Robust Multi-chip Average (RMA) was used initially to extract probe intensity data. RMA background correction was applied including pre-background adjustment for GC content and quantile normalization across all chips in the experiment. Probe data was log2 transformed.

Gene level expression analysis

Two separate experiments (experiment 1 and 2) were carried out, each time with one testis from homozygous wild type and homozygous KO mice. Hence for each gene a total of two values were obtained in WT (*Dpy* +/+ (1) and (2)) and KO mice (*Dpy* -/- (1) and (2)).

For each gene transcripts we calculated 4 ratios corresponding to the 4 possible combinations

$$R1 = \frac{Dpy-/- (1)}{Dpy +/+ (1)} \quad R3 = \frac{Dpy-/- (2)}{Dpy +/+ (1)}$$
$$R2 = \frac{Dpy-/- (1)}{Dpy +/+ (2)} \quad R4 = \frac{Dpy-/- (2)}{Dpy +/+ (2)}$$

For each gene, if at least three of these ratios appeared ≥ 1.7 fold up or down, the transcript was

considered to be significantly differentially expressed. These values and the log2 ratio for all deregulated genes are shown in Additional file 2: Table S2. The histogram of the log2 ratio of each deregulated gene is shown in Figure 2.

Gene ontology analysis

The lists of genes expressed differently in *Dpy19l2*-/- mice were imported into PANTHER (<http://www.pantherdb.org/>) to identify the biological process, molecular functions and gene networks significantly deregulated in *Dpy19l2*-/- testis compared to WT controls.

Results

Gene expression profile

Array hybridization was performed with the Affymetrix Mouse Exon 1.0 ST array, which contains probe sets for 35,557 genes. Of these, we identified that 76 genes had a level of testicular expression that was different between WT and *Dpy19l2*-/- mice (transcripts with an expression ratio ≥ 1.7 fold up- or down regulated). Among them, 6 genes were underexpressed and 70 genes were over-expressed (Figure 2 and Additional file 2: Table S2). As expected *Dpy19l2* was found part of the down-regulated genes, thus validating the experimental approach we used. Interestingly, we did not observe any difference in the expression level of PLCzeta in the testes from KO and WT mice.

Panther gene ontology analysis

The 76 genes that were differentially regulated were uploaded into the PANTHER software (Gene List Analysis). Among them 64 were recognized by the PANTHER software. The molecular functions and biological process predictions that are generated from PANTHER are based on the direction of expression of a number of downstream genes which have been previously shown to be associated with these functions. The list of each function associated to all deregulated genes is provided in Additional file 3: Table S3. Several molecular functions were found to be enriched in the testis of *Dpy19l2*-/- mice (Figure 3). Genes encoding proteins which are able to bind nucleic acids or proteins were most frequently deregulated (23 genes), especially those encoding for protein binding to the nucleic acids (12 genes), confirming that *Dpy19l2* could interact with DNA. Other functions such as catalytic activity, transcription regulator activity, structural functions were also deregulated in the KO mice testes. Because of its location in the inner nucleus membrane, *DPY19L2* could be a bridge between the nucleus and the cytoplasm. We observed that 5 genes encoding for transporters are deregulated in KO mice: among them, four are transmembrane transporters and one is a lipid transporter. Moreover globozoospermia is characterized by structural deficiency

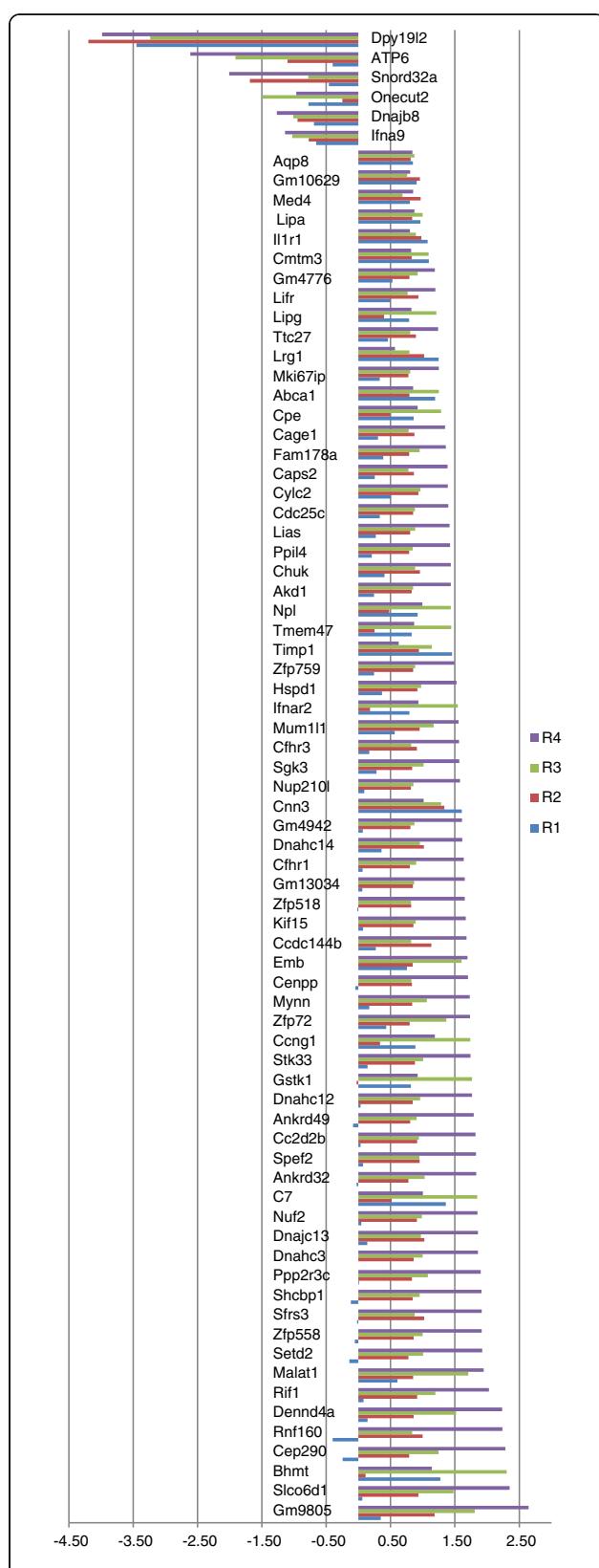


Figure 2 Histogram of the log₂ ratio of all deregulated transcripts. Genes that are upregulated in Dpy19l2 KO mice have positive values while genes that are downregulated have negative values. Values can be seen in Additional file 2: Table S2.

of spermatozoon head and we see that 6 deregulated genes encode for proteins with structural molecular function.

Numerous biological processes are also deregulated in the testis of *Dpy19l2*^{-/-} mice (Figure 4). Metabolic processes and cellular processes are most often deregulated. We see that 6 genes predicted to be involved in reproduction biological process separated deregulated. Among those, no genes were described to be involved in the acrosome formation but two genes encode for dyneins and one for a protein predicted to be involved in sperm motility.

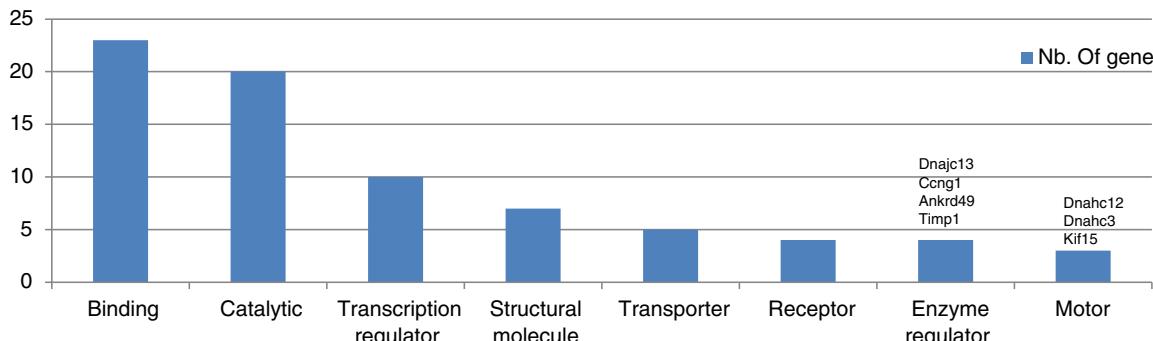
Discussion

Spermiogenesis is the final stage of spermatogenesis. During this step, the nucleus condenses, acquires its specific shape, and the flagellum and the acrosome are formed. The acrosome is essential for the spermatozoa to cross of the ZP and is thus necessary for *in vivo* fertilization. Globozoospermia is a teratozoospermia characterised by the formation of round-head spermatozoa without acrosome. This pathology has been described to be associated with the absence of the protein PLCzeta which is also known to be essential for fertilization and oocyte activation [4-7]. We previously demonstrated that this pathology is mostly due to a homozygous deletion of the testis-specific gene *DPY19L2* [22,23] and that DPY19L2 is expressed in spermatids and it is located only in a restricted zone of the nuclear membrane facing the acrosome.

This study revealed that 76 genes were deregulated in the testis of *Dpy19l2* KO mice. This result could be concordant with a very specific regulatory role of Dpy19l2 at the transcription level. On the other hand we note that the micro-array contains 35,557 probe-set for almost as many genes. It is therefore a small minority (0.2%) of genes that is deregulated in DPY19L2 KO mice. It is interesting to note that almost all of these genes appeared as up-regulated and that only 5 of them were down-regulated. If Dpy19l2 has a direct influence on gene regulation we can therefore say that it mainly act as a repressor of gene expression. We note that apart from Dpy19l2, which is obviously absent from the KO and is found (due to background fluorescence levels) to have a 4 fold decrease in expression compared to controls, the most down-regulated gene, ATP6, has a 2.6 fold decreased expression and the most up-regulated gene, Cepp, has a 2.2 fold increased expression. The observed level of transcription modifications is therefore moderate.

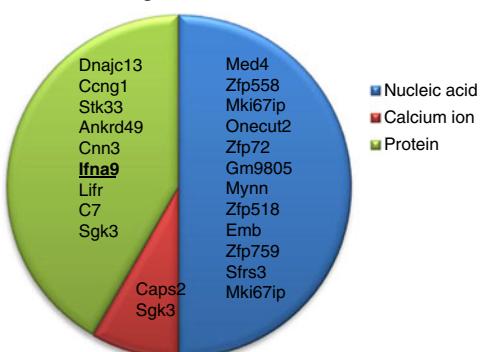
Dpy19l2 co-localises with SUN5 [28] and we hypothesized that SUN5 is a likely partner of Dpy19l2 [25]. In mouse, Sun1, another Sun protein, was also described to

I: Molecular functions

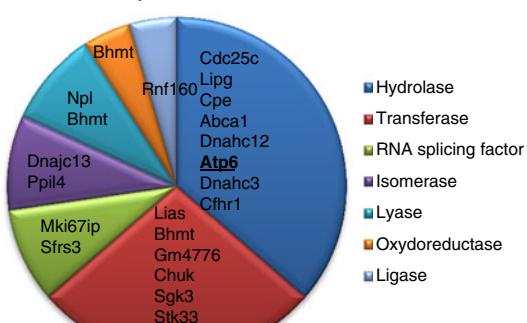


II: Molecular functions details

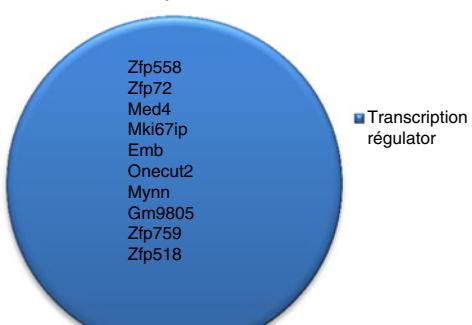
A: Binding



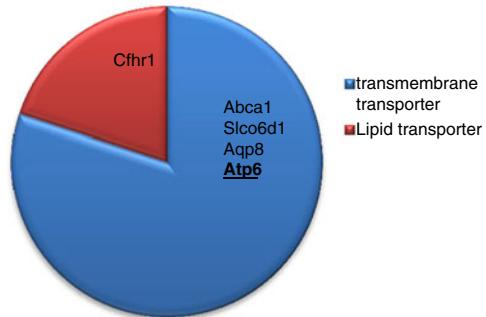
B: Catalytic



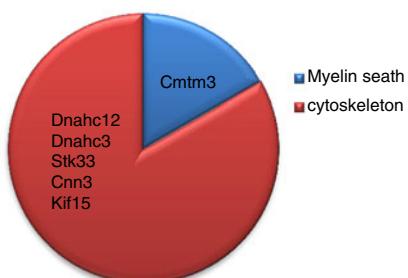
C: Transcription



D: Transporter



E: Structural molecule



F: Receptor

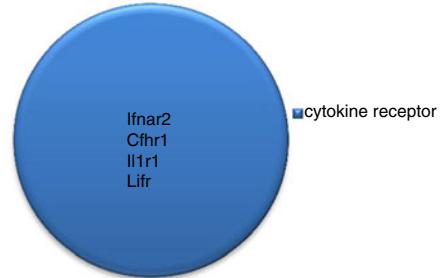
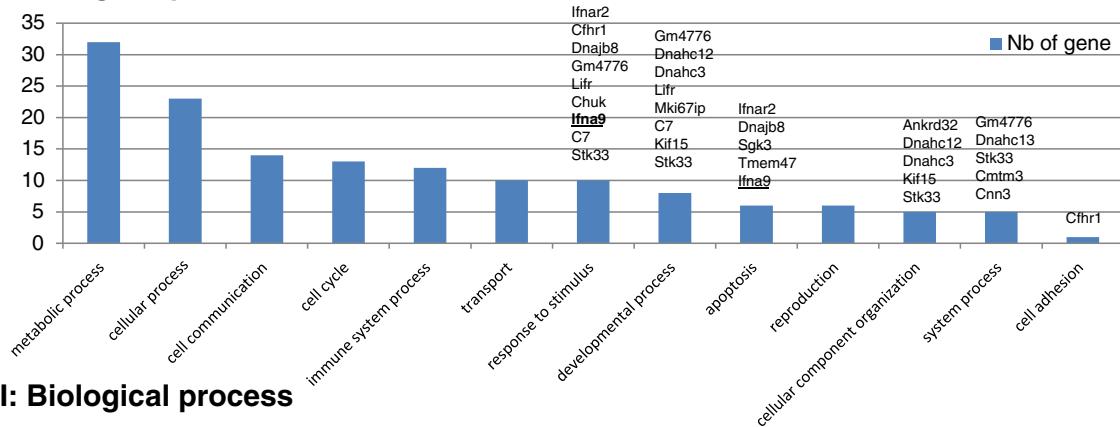


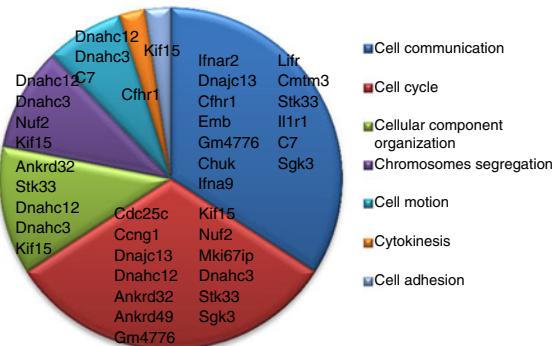
Figure 3 I. Histogram presenting all PANTHER molecular function of genes that are deregulated in *Dpy19l2^{-/-}* mice testes. **II, A-F.** Details of some of PANTHER molecular functions. Up-regulated genes are in bold and underlined.

I: Biological process

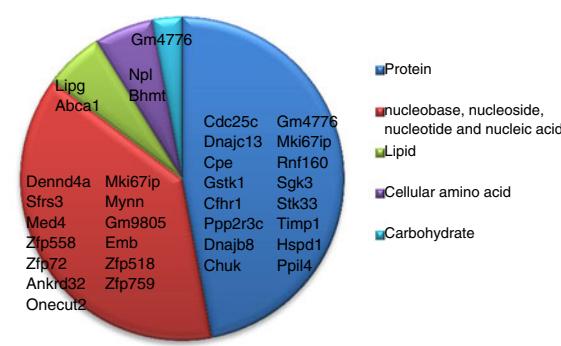


II: Biological process

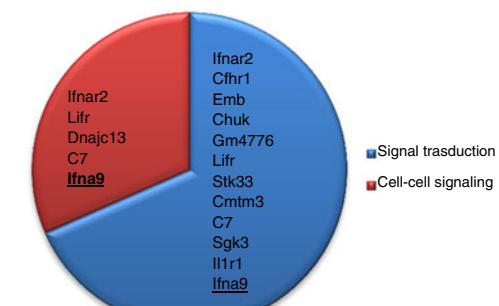
A: Cellular process



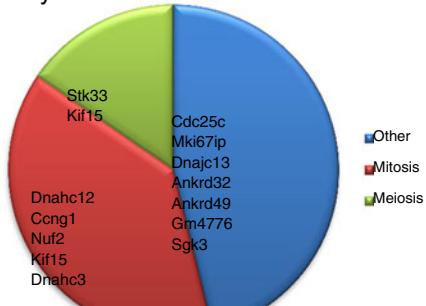
B: Metabolic process



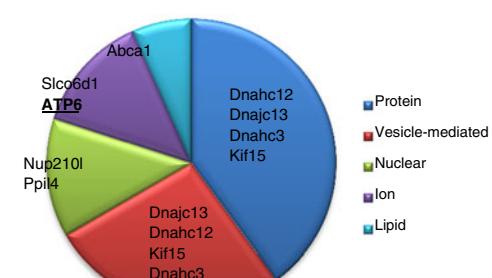
C: Cell communication



D: Cell cycle



E: Transport



F: Reproduction

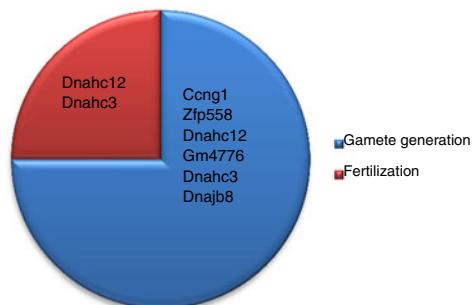


Figure 4 I. Histogram presenting all PANTHER biological process of genes deregulated in *Dpy19l2*^{-/-} mice testes. **II, A-F,** Details of some of PANTHER biological process. Up-regulated genes are in bold and underlined.

be necessary for gametogenesis and was shown to interact in the nucleus with the telomeres [27]. We observed that most of the deregulated genes (70/76) were up regulated in KO animal. We can hypothesize that Dpy19l2 could also interact, directly or indirectly, potentially via Sun5, with germ cell DNA and thus could have an effect on the regulation of transcription in spermatogenic cells. Heterochromatin is constituted by highly compact transcriptionally repressed DNA. It regroups down-regulated genes and is particularly abundant at the periphery of the nucleus where it interacts with factors located in the nuclear lamina. We can thus speculate that Dpy19l2 could intervene during spermiogenesis to include selected genes in heterochromatin repressive domains. In the absence of Dpy19l2, these genes would not be repressed and appear as up regulated. This regulations could be a limited to selected loci as electron microscopic observations of round spermatids nuclei from *Dpy19l2* KO animals do not show any obvious difference in the abundance of heterochromatin [25].

The PANTHER software allows a classification of genes according to their predicted molecular functions (Figure 3). We see that the most represented gene function that is deregulated in *Dpy19l2* mice is “binding” (23 genes). This group is divided in three sub categories: nucleic acid, protein and calcium ion (Additional file 2: Table S2). $[Ca^{2+}]_i$ is known to play an important role in male fertility. $[Ca^{2+}]_i$ signaling is the primary regulator of sperm flagellum beating and calcium intracellular rise is known to be essential for the acrosome reaction [29]. Indeed, solubilisation of the zona pelucida stimulates generation of IP₃ in mouse sperm [30] which is known to mobilize the acrosomal Ca²⁺ stored to permit acrosomal reaction [31,32]. The biochemical nature of the Ca²⁺-binding sites are globally unknown but recently a calcium-binding protein has been isolated from the acrosomal membrane of bovine spermatozoa [33]. We observe that in *Dpy19l2* KO mice two calcium binding proteins are up-regulated : Caps2 and Sgk3 (Figure 2 and 3). Ten of the deregulated genes are described to encode proteins with DNA binding abilities. Although we did not find direct evidence that these encoded proteins have transcriptional regulation activities, they might be involved in the regulation of gene expression and play a role in the up- and -down regulation of some of the other genes we found to be deregulated in this transcriptome analysis.

We did not observe a down-regulation of PLCzeta that could account for its absence from round-headed sperms. This suggests that in *Dpy19l2*−/− mice *PLCzeta* is normally expressed but that the absence of *Dpy19l2* and the abnormalities it induces on sperm morphology likely prevents the correct positioning of *PLCzeta*, which is likely to be eliminated in the residual body. This hypothesis is consolidated by the fact that several studies show that

treatment with a calcium ionophore improves ICSI success rates? results for men with globozoospermia [5]. We note however that fertilization and pregnancies can be achieved by ICSI on *DPY19L2* deleted men [34]. This can probably be explained by the fact that remains of misplaced *PLCzeta* often position near the manchette can be observed on a small proportion of round-headed sperm [6].

This study also reveals that several genes encoding for transporters were deregulated in *Dpy19l2* KO mice. Among them four are transmembrane transporters and one is a lipid transporter. We note the deregulation of the gene *Abca1*, which is expressed in mouse spermatozoa within the seminiferous tubules and the epididymis, and is a key regulator of cholesterol efflux. Depletion of the cholesterol from the cytoplasmic plasma membrane and modification of its lipid composition is one of the key events in the process of spermatozoa capacitation, which ultimately leads to the acrosome reaction and egg fertilization. Transporters and in particular those mediating cholesterol efflux, are thus particularly important. The deregulation of *Abca1* could therefore alter the physiological composition of mature sperm and contribute to the poor fertilization potential of *Dpy19l2* mutant sperm.

The analyze of biological process regulated in *Dpy19l2*−/− mice reveals 6 genes predicted to be involved in reproductive functions and particularly in gamete generation and fertilization. Surprisingly half of these genes code for dyneins, which are important constituents of the microtubules. The others are involved in the processes of sperm motility and cytoskeleton structure. These results can be linked to our previous observation that the absence of *Dpy19l2* leads to the destabilization of both the nuclear dense lamina and the junction between the acroplaxome and the nuclear envelope. This destabilization causes a failure of the linkage of the acrosome and the manchette to the acroplaxome, a cytoskeletal plate anchored to the nuclear envelope. The manchette is a transient microtubular structure necessary during spermatid elongation. Moreover, the manchette is necessary for protein trafficking and its defects could disturb the overall distribution of proteins in spermatids [35].

Conclusions

We showed that *Dpy19l2*−/− induced globozoospermia altered gene expression in mice testis but the overall modifications at the transcript level remained modest. We showed that *PLCzeta* was not down-regulated in KO mice indicating that the absence of the protein observed in the sperm of globozoospermic patient is not due to a transcriptional deregulation. This likely indicates that *PLCzeta* cannot reach its physiological localization on round-headed spermatozoa and that it is probably lost with the cytoplasmic elimination (residual body) during

spermiogenesis. We also observed that several genes encoding proteins involved in transports, and in particular Abca1, involved in the cholesterol efflux, were deregulated. This could also contribute to the poor fertilization potential of the round-headed spermatozoa. Secondary anomalies stemming from the morphological abnormalities of the sperm could also lead to a wide range of protein deregulation as exemplified by the absence of PLCzeta. A proteomic analysis of these deregulations could permit to have a functional view of the extent of the molecular anomalies present in *Dpy19l2* KO mice. Further work will permit a better comprehension of molecular mechanism involved in spermatogenesis and in the physiopathology of globozoospermia.

Additional files

Additional file 1: Table S1. RNA quantification.

Additional file 2: Table S2. Ratios of transcripts values measured in *Dpy19l2* WT and KO mice.

Additional file 3: Table S3. PANTHER output of all deregulated genes in *Dpy19l2* KO mice.

Abbreviations

AR: Acrosomal reaction; DNA: Deoxyribonucleic acid; dNTP: Deoxynucleotide triphosphates; FISH: Fluorescent in situ hybridization; ICSI: Intracytoplasmic sperm injection; IVF: In vitro fertilization; Kb: Kilobase (1000 nucleotides); KO: Knock-out; LCR: Low-copy repeats; LINC: Linker of nucleoskeleton and cytoskeleton; NAHR: Non-allelic homologous recombination; PCR: Polymerase chain reaction; PLC zeta: Phospholipase zeta; RNA: Ribonucleic acid; WT: Wild type; ZP: Zona pellucida.

Competing interests

The authors have no competing interests.

Authors' contributions

TK realised most of molecular work. MEA, MG, CC, MD and CA provided technical help. JPI and CA provided conceptual help. PFR conceived, designed the experiments and supervised the work. TK and PR wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the program GENOPAT 2009 from the French Research Agency (ANR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

¹Université Joseph Fourier, Grenoble F-38000, France. ²Laboratoire AGIM, CNRS FRE3405, Equipe "Génétique, Infertilité et Thérapeutiques", La Tronche F-38700, France. ³Team7 Nanomedicine and Brain, INSERM U836, Grenoble, France. ⁴Institut des Neurosciences, Université Joseph Fourier, Grenoble, France. ⁵Clinical Transcriptomics and Proteomics Platform, Centre Hospitalier Universitaire et Grenoble Institut des Neurosciences, Grenoble, CNRS, Grenoble, France. ⁶CHU de Grenoble, UF de Biochimie et Génétique Moléculaire, Grenoble cedex 9 F-38043, France. ⁷CHU de Grenoble, Département de Génétique et Procréation, Grenoble cedex 9 F-38043, France.

Received: 29 April 2013 Accepted: 22 July 2013

Published: 3 September 2013

References

1. Mascarenhas MN, Flaxman SR, Boerma T, Vanderpoel S, Stevens GA: National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 health surveys. *PLoS Med* 2012, 9:e1001356.
2. Kierszenbaum AL, Tres LL: The acrosome-acropaxosome-manchette complex and the shaping of the spermatid head. *Arch Histol Cytol* 2004, 67:271–84.
3. Dam AH, Feenstra I, Westphal JR, Ramos L, van Golde RJ, et al: Globozoospermia revisited. *Hum Reprod Update* 2007, 13:63–75.
4. Heytens E, Parrington J, Coward K, Young C, Lambrecht S, et al: Reduced amounts and abnormal forms of phospholipase C zeta (PLCzeta) in spermatozoa from infertile men. *Hum Reprod* 2009, 24:2417–28.
5. Taylor SL, Yoon SY, Morshed MS, Lacey DR, Jellerette T, et al: Complete globozoospermia associated with PLCzeta deficiency treated with calcium ionophore and ICSI results in pregnancy. *Reprod Biomed Online* 2010, 20:559–64.
6. Kashir J, Sermonade N, Sifer C, Oo SL, Jones C, et al: Motile sperm organelle morphology evaluation-selected globozoospermic human sperm with an acrosomal bud exhibits novel patterns and higher levels of phospholipase C zeta. *Hum Reprod* 2012, 27:3150–60.
7. Holstein AF, Schirren C, Schirren CG: Human spermatids and spermatozoa lacking acrosomes. *J Reprod Fertil* 1973, 35:489–91.
8. Swann K, Lai FA: PLCzeta and the initiation of Ca(2+) oscillations in fertilizing mammalian eggs. *Cell Calcium* 2013, 53:55–62.
9. Swann K, Larman MG, Saunders CM, Lai FA: The cytosolic sperm factor that triggers Ca2+ oscillations and egg activation in mammals is a novel phospholipase C: PLCzeta. *Reproduction* 2004, 127:431–9.
10. Yoon SY, Jellerette T, Salicioni AM, Lee HC, Yoo MS, et al: Human sperm devoid of PLC, zeta 1 fail to induce Ca(2+) release and are unable to initiate the first step of embryo development. *J Clin Invest* 2008, 118:3671–81.
11. Kullander S, Rausing A: On round-headed human spermatozoa. *Int J Fertil* 1975, 20:33–40.
12. Alvarez Sedo C, Rawe VY, Chemes HE: Acrosomal biogenesis in human globozoospermia: immunocytochemical, ultrastructural and proteomic studies. *Hum Reprod* 2012, 27:1912–21.
13. Dam AH, Ramos L, Dijkman HB, Woestenenk R, Robben H, et al: Morphology of partial globozoospermia. *J Androl* 2011, 32:199–206.
14. Escalier D: Failure of differentiation of the nuclear-perinuclear skeletal complex in the round-headed human spermatozoa. *Int J Dev Biol* 1990, 34:287–97.
15. Florke-Gerloff S, Topfer-Petersen E, Muller-Esterl W, Mansouri A, Schatz R, et al: Biochemical and genetic investigation of round-headed spermatozoa in infertile men including two brothers and their father. *Andrologia* 1984, 16:187–202.
16. Perrin A, Coat C, Nguyen MH, Talagas M, Morel F, et al: Molecular cytogenetic and genetic aspects of globozoospermia: a review. *Andrologia* 2013, 45:1–9.
17. Perrin A, Louanjli N, Ziane Y, Louanjli T, Le Roy C, et al: Study of aneuploidy and DNA fragmentation in gametes of patients with severe teratozoospermia. *Reprod Biomed Online* 2011, 22:148–54.
18. Perrin A, Morel F, Moy L, Colleu D, Amice V, et al: Study of aneuploidy in large-headed, multiple-tailed spermatozoa: case report and review of the literature. *Fertil Steril* 2008, 90:1201.e13–1201.e17.
19. Sermonade N, Hafhouf E, Dupont C, Bechoua S, Palacios C, et al: Successful childbirth after intracytoplasmic morphologically selected sperm injection without assisted oocyte activation in a patient with globozoospermia. *Hum Reprod* 2011, 26:2944–9.
20. Zhu F, Gong F, Lin G, Lu G: DPY19L2 gene mutations are a major cause of globozoospermia: identification of three novel point mutations. *Mol Hum Reprod* 2013, 19(6):395–404.
21. Dam AH, Kosciński I, Kremer JA, Moutou C, Jaeger AS, et al: Homozygous mutation in SPATA16 is associated with male infertility in human globozoospermia. *Am J Hum Genet* 2007, 81:813–20.
22. Harbuz R, Zouari R, Pierre V, Ben Khelifa M, Kharouf M, et al: A Recurrent Deletion of DPY19L2 Causes Infertility in Man by Blocking Sperm Head Elongation and Acrosome Formation. *Am J Hum Genet* 2011, 88:351–61.
23. Coutton C, Zouari R, Abada F, Ben Khelifa M, Merdassi G, et al: MLPA and sequence analysis of DPY19L2 reveals point mutations causing globozoospermia. *Hum Reprod* 2012, 27:2549–58.

24. Coutton C, Abada F, Karaouzene T, Sanlaville D, Satre V, et al: Fine Characterisation of a Recombination Hotspot at the DPY19L2 Locus and Resolution of the Paradoxical Excess of Duplications over Deletions in the General Population. *PLoS Genet* 2013, **9**:e1003363.
25. Pierre V, Martinez G, Coutton C, Delaroche J, Yassine S, et al: Absence of Dpy19l2, a new inner nuclear membrane protein, causes globozoospermia in mice by preventing the anchoring of the acrosome to the nucleus. *Development* 2012, **139**:2955–65.
26. Tzur YB, Wilson KL, Gruenbaum Y: SUN-domain proteins: 'Velcro' that links the nucleoskeleton to the cytoskeleton. *Nat Rev Mol Cell Biol* 2006, **7**:782–8.
27. Ding X, Xu R, Yu J, Xu T, Zhuang Y, et al: SUN1 is required for telomere attachment to nuclear envelope and gametogenesis in mice. *Dev Cell* 2007, **12**:863–72.
28. Frohnert C, Schweizer S, Hoyer-Fender S: SPAG4L/SPAG4L-2 are testis-specific SUN domain proteins restricted to the apical nuclear envelope of round spermatids facing the acrosome. *Mol Hum Reprod* 2011, **17**:207–18.
29. Darszon A, Nishigaki T, Beltran C, Trevino CL: Calcium channels in the development, maturation, and function of spermatozoa. *Physiol Rev* 2011, **91**:1305–55.
30. Strunko T, Goodwin N, Brenker C, Kashikar ND, Weyand I, et al: The CatSper channel mediates progesterone-induced Ca²⁺ influx in human sperm. *Nature* 2011, **471**:382–6.
31. O'Toole CM, Arnoult C, Darszon A, Steinhardt RA, Florman HM: Ca(2+) entry through store-operated channels in mouse sperm is initiated by egg ZP3 and drives the acrosome reaction. *Mol Biol Cell* 2000, **11**:1571–84.
32. Herrick SB, Schweissinger DL, Kim SW, Bayan KR, Mann S, et al: The acrosomal vesicle of mouse sperm is a calcium store. *J Cell Physiol* 2005, **202**:663–71.
33. Nagdas SK, Buchanan T, McCaskill S, Mackey J, Alvarez GE, et al: Isolation of a calcium-binding protein of the acrosomal membrane of bovine spermatozoa. *Int J Biochem Cell Biol* 2013, **45**:876–84.
34. Kuentz P, Vanden Meerschaut F, Elinati E, Nasr-Esfahani MH, Gurgan T, et al: Assisted oocyte activation overcomes fertilization failure in globozoospermic patients regardless of the DPY19L2 status. *Hum Reprod* 2013, **28**:1054–61.
35. Kierszenbaum AL, Rivkin E, Tres LL: Cytoskeletal track selection during cargo transport in spermatids is relevant to male fertility. *Spermatogenesis* 2011, **1**:221–230.

doi:10.1186/2051-4190-23-7

Cite this article as: Karaouzène et al.: Comparative testicular transcriptome of wild type and globozoospermic *Dpy19l2* knock out mice. *Basic and Clinical Andrology* 2013 **23**:7.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit



Principaux résultats :

Pour effectuer ces analyses, nous avons donc extrait l'ARN testiculaire des 4 souris que nous avons ensuite hybridé sur des puces à ADN Affymetrix GeneChip® Mouse Exon 1.0 contenant des sondes pour 35.557 gènes murins. Cette étape nous a permis d'obtenir pour chacune des 4 souris les valeurs d'expression testiculaire de l'ensemble de leurs gènes. Pour chacun de ces gènes, nous avons donc cherché à savoir s'ils étaient différemment exprimés chez les souris S1⁻ et S2⁻ lorsqu'on comparait leur expression avec celle des souris S1⁺ et S2⁺. Pour cela, nous avons calculé quatre ratios (R1, R2, R3 et R4) (**Équation** : (3.3)). Les gènes pour lesquels au moins 3 de leurs ratios étaient $\geq 1,7$ furent considérés comme sur-exprimés tandis que ceux pour lesquels 3 de leurs ratios étaient $\leq 0,58$ ($\frac{1}{1,7}$) furent considérés comme sous-exprimés.

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

$$\begin{aligned} R1_{gene} &= \frac{\exp_{gene}(S1^-)}{\exp_{gene}(S1^+)} & R2_{gene} &= \frac{\exp_{gene}(S2^-)}{\exp_{gene}(S1^+)} \\ R3_{gene} &= \frac{\exp_{gene}(S1^-)}{\exp_{gene}(S2^+)} & R4_{gene} &= \frac{\exp_{gene}(S2^-)}{\exp_{gene}(S2^+)} \end{aligned} \quad (3.3)$$

De cette manière cette étude a pu mettre en évidence la sous-expression de 6 gènes (incluant *Dpy19l2*) et la sur-expression de 70 gènes chez les souris *Dpy19l2*^{-/-}. *Plcz1* ne figurait pas parmi ces gènes indiquant que l'absence de cette protéine chez les spermatozoïdes globozoocéphales n'étaient pas due à un dysfonctionnement transcriptionnel.

Afin de prédire les fonctions moléculaires dans lesquels étaient impliqués ces gènes, nous nous sommes servis du logiciel PANTHER [182]. Ainsi, nous avons pu constater que 23 gènes codant pour des protéines de liaison étaient dérégulés (**Figure** : 3.5 - A), dont sont des protéines de liaison aux acides nucléiques (**Figure** : 3.5 - B) suggérant que *Dpy19l2* pourrait effectivement interagir avec l'ADN. D'autres fonctions moléculaires telles que l'activité catalytique, la régulation de la transcription et des protéines ayant des fonctions structurelles étaient également dérégulées chez les souris KO. Ces fonctions sont particulièrement intéressantes lorsque l'on sait que les spermatozoïdes globozoocéphales sont caractérisés par plusieurs défauts structurels.

Cette étude a pour nous été l'occasion de mieux caractériser la protéine *Dpy19l2* chez la souris. Nous avons ainsi pu montrer que les souris *Dpy19l2*^{-/-} présentaient des dérèglements transcriptionnels affectant plusieurs fonctions moléculaires pouvant potentiellement expliquer, du moins en partie, les nombreux défauts morphologiques caractérisant les spermatozoïdes globozoocéphales. De même, nous avons pu observer le dérèglement de nombreux gènes impliqués dans la liaison d'acide nucléique et de protéine pouvant ainsi expliquer les défauts d'ancre de l'acrosome au noyau chez les spermatozoïdes globozoocéphales.

Ces résultats ne nous ont cependant pas permis d'expliquer l'absence de la protéine *Plcz1* dans le spermatozoïde globozoocéphale murin l'expression du gène *Plcz1* n'ayant montré aucune dérégulation chez la souris *Dpy19l2^{-/-}*. De même, aucun des gènes retrouvés comme dérégulés ne nous a permis d'expliquer le biais de sexe que nous avions observés. Cela n'a pas été une surprise pour nous puisque après avoir entamé notre étude, une dernière portée issues d'un croisement de souris *Dpy19l2⁺⁻* a vu le jour. Celle-ci était composée de 4 souriceaux mâles et de 4 souriceaux femelles. Ainsi, avec un total de 32 souris mâles pour 20 souris femelles, la p-valeur de notre test du χ^2 à 0.0635765 laissant cette fois-ci supposer la non-existence d'un biais de sexe dans les naissances issues d'un croisement de souris *Dpy19l2^{-/-}*.

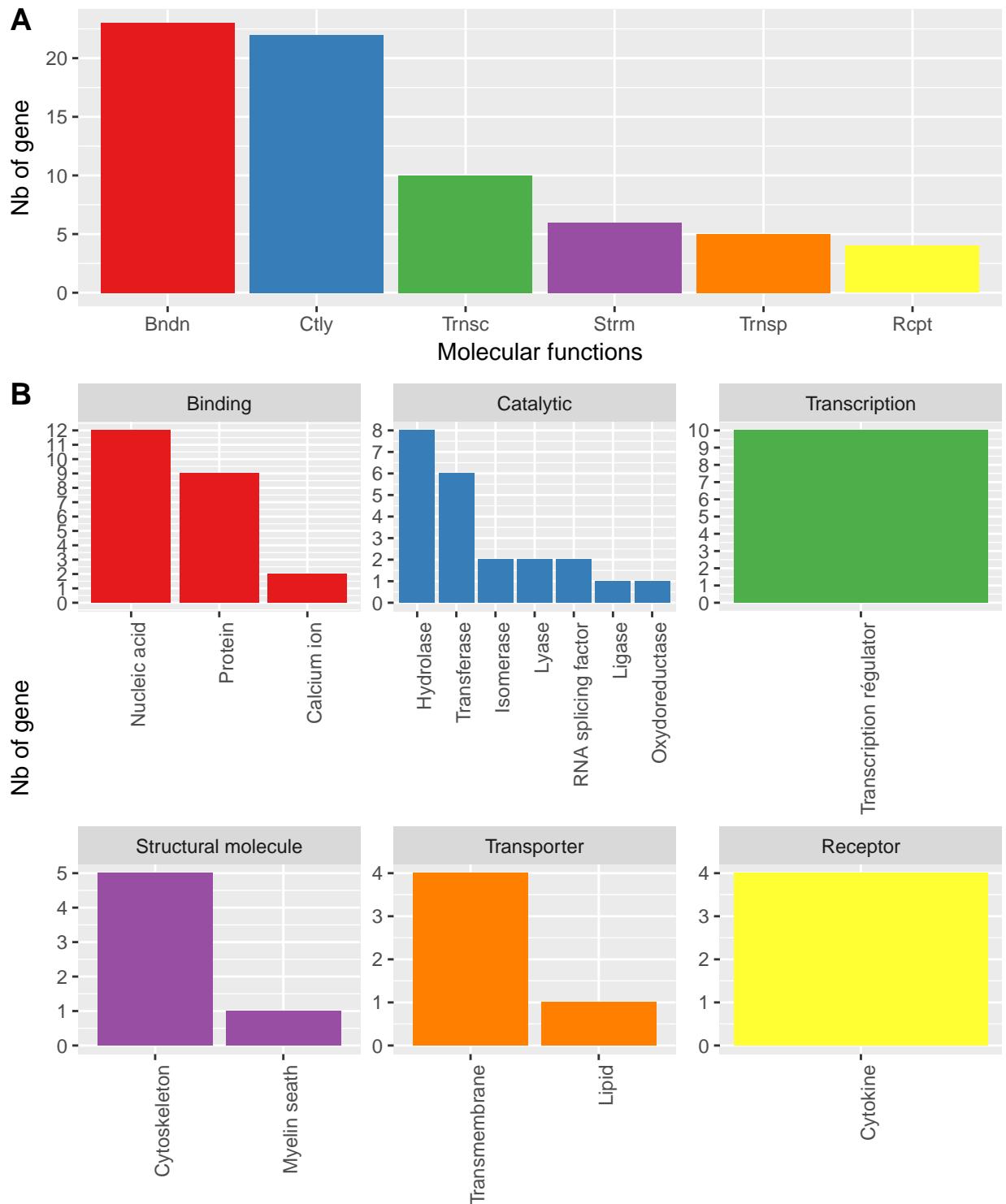


Figure 3.5 – Principales fonctions moléculaires affectées chez les souris Dpy19l2 KO : A : Liste des fonctions moléculaires affectées : Bndn = Binding, Ctly = Catalytic, Trnsc = Transcription, Strm = Structural molecule, Rcpt = Receptor. B : Détails des fonctions moléculaires affectées par les gènes dérégulés.

Conclusion et discussion

L'infertilité est une problématique qui concerne entre 10 et 15% des couples [53] faisant de cette pathologie un enjeu de santé publique. Bien que les causes de ce phénotype puissent être multifactorielles et acquises au cours de la vie de l'individu, notamment suite à des infections du système urogénital ou encore à des perturbations du système endocrinien, la composante génétique est extrêmement importante. À ce jour, malgré les efforts de nombreuses équipes, incluant la nôtre, seulement une poignée de gènes a pu être reliée à ce phénotype. De plus, pour nombre d'entre eux, les bases moléculaires reliant une mutation au phénotype d'infertilité restent inconnues. Ces quinze dernières années, les techniques d'investigation employées dans les analyses phénotypes-génotype, ont été bouleversées par l'émergence du séquençage haut-débit. En effet, ces technologies permettent aujourd'hui de séquencer l'ensemble du génome / exome d'un individu pour un coût et dans un temps raisonnable. Malheureusement, la masse de données générées par ces méthodes, bien qu'à l'origine du succès de celles-ci, deviennent aujourd'hui un frein dans l'analyse et la compréhension des processus biologiques étudiés.

Dans ce contexte, j'ai pu, au cours de ma thèse, mettre au point un pipeline permettant l'analyse des données issues de séquençage NGS. Ce pipeline a ainsi permis, entre 2012 et 2017, l'analyse des données de nombreux patients présentant tous un phénotype d'infertilité. Celle-ci ayant pour vocation première de mettre en évidence les variants responsables des phénotypes de ces patients. Contrairement à la plupart des pipelines d'analyse de données WES existant, celui-ci prend en charge l'ensemble des étapes de l'analyse allant de l'alignement des *short-reads* sur le génome de référence jusqu'à la priorisation des variants en passant par l'appel des variants et leur annotation. Les résultats de chacune de ces étapes pouvant être contrôlées et personnalisées grâce à des paramètres ajustables. L'alignement des *reads* est effectué par le logiciel MAGIC tandis que les variants et leur génotype sont appelés par un algorithme développé dans notre laboratoire spécifiquement conçu pour analyser les informations fournies par MAGIC et dont les paramètres sont ajustables en fonction de la distribution des pourcentages de *reads* variants observés dans les données analysées. Pour l'annotation nous avons utilisé plusieurs ressources extérieures tel que le logiciel *Variant Effect Predictor* qui va nous informer de l'effet d'un variant sur l'ensemble des transcrits qu'il chevauche. De même, les bases de données ExAC ESP6500 ou encore 1KG nous donne une indication de la fréquence des variants dans la population générale. Une fois ces étapes effectuées, nous avons mis en place plusieurs filtres successifs afin d'éliminer de nos listes les variants ayant le moins de chances d'être responsables du phénotype des différents patients. Ceux-ci s'appuient à la fois sur les critères qualité des résultats de séquençage, le génotype des variants, leur fréquence ou encore leur impact sur la protéine.

L'efficacité de ce pipeline a pu être démontré grâce à son utilisation sur des cas familiaux mais aussi sur des cohortes d'individus non apparentés. Ainsi, nous avons pu dans un premier temps confirmer l'importance de l'implication du gène *DNAH1* dans le syndrome MMAF en le retrouvant muté chez 9 de nos patients (4 au sein d'études familiales et 5 parmi notre cohorte d'individus non apparentés). Ensuite, dans un second temps, ce pipeline nous a permis de mettre en évidence un total de

5 nouveaux gènes dans des phénotypes d'infertilité masculine et féminine. Ainsi les gènes *CFPA43*, *CFAP44*, retrouvés respectivement mutés à chez 10 (9 homozygotes et 1 hétérozygote composite) et 6 (tous homozygotes) de nos patients ont pu être liés à leur syndrome MMAF. Aussi, une même mutation impactant le gène *PATL2* a pu être reliée au phénotype de déficience méiotique ovocytaire de cinq femmes. Pour finir, des mutations sur les gènes *SPINK2* et *PLCZ1* ont, elles aussi, pu être liées aux phénotypes d'azoospermie et d'échec de fécondation dont étaient atteint deux fratries.

Dans autre partie de mon travail de thèse j'ai pu prendre part à la caractérisation génétiques et moléculaires du gène *DPY19L2* impliqué dans le phénotype de globozoospermie. Ce phénotype, entraînant la production de 100% de spermatozoïdes à têtes rondes et dépourvus d'acrosomes est principalement causé, chez l'humain, par une délétion homozygote récurrente entraînant la perte de la totalité de la séquence du gène *DPY19L2*. Ainsi, dans deux études différentes, nous avons pu, dans un premier temps, mieux caractériser les mécanismes moléculaires responsables de cette délétion. Ainsi, nous avons pu mettre en évidence cinq points de cassures au niveau des LCRs flanquant la séquence de *DPY19L2* chez l'humain. Ceux-ci étant tous concentrés dans une région d'environ 1150 pb contenant en son centre un site de reconnaissance consensus de la protéine PRDM9 connue pour son implication dans la recombinaison chromosomique chez l'humain et la souris [227, 228]. Cette même étude a également permis de démontrer que les effets de la sélection naturelle étaient responsables du paradoxe observé dans la population générale : une fréquence plus élevée d'allèles dupliqués comparativement aux allèles délétés au locus *DPY19L2* tandis que *de novo*, l'allèle délété est produit, en théorie et en pratique, plus fréquemment que l'allèle dupliqué. L'étude de ce phénotype nous a par la suite poussée à étudier le modèle murin KO *Dpy19l2^{-/-}* présentant le même phénotype que l'humain. Afin d'expliquer l'absence de la protéine *PLCZ1* chez l'humain globozoosperme, nous avons effectué une analyse comparative des transcriptomes testiculaires de souris sauvages *Dpy19l2^{+/+}* et KO *Dpy19l2^{-/-}*. Bien qu'aucun dérèglement transcriptionnel n'ait pu être observé pour le gène *Plcz1* cette étude nous a permis de mettre en évidence un total de 75 gènes présentant des dérégulations transcriptionnelles pouvant expliquer en partie les anomalies physiologiques et morphologiques des spermatozoïdes des souris *Dpy19l2^{-/-}*.

Au cours des différents travaux réalisés au cours de ma thèse, nous avons pu constater la puissance des technologies de séquençage haut-débit. En effet, en seulement 5 ans, celles-ci ont permis l'identification de 5 nouveaux gènes impliqués dans des phénotypes d'infertilité au sein de notre laboratoire. Ces résultats sont cependant à relativiser puisque qu'aucun candidat n'a pu être identifié pour 68% des patients analysés. Plusieurs raisons peuvent expliquer cela. Tout d'abord, au cours des analyses décrites dans ces manuscrits nous nous concentrons uniquement sur les SNPs et les indels. Cependant de nombreux logiciels tel que ExomeDepth [230], CoNIFER [231] ou encore ExomeCNV [232] permettent de détecter des CNVs à partir de données WES et / ou WGS. Les stratégies de prédictions de ces logiciels pouvant être extrêmement différents (**Figure : 3.6**), le profil des CNVs détectés ou non le sera aussi [233, 234].

Ainsi, dans des analyses non décrites dans ce manuscrit, j'ai pu chercher à identifier des CNVs à partir de nos données d'exome à l'aide du logiciel ExomeDepth [230]. Cette approche a été extrêmement concluante puisqu'elle a permis d'identifier une délétion homozygote sur le gène *WDR66* chez 7 de nos patients pour lesquels aucun candidat n'avait été alors identifié. Ces délétions ont ensuite pu être confirmées par PCR et la caractérisation de ce gène est actuellement en cours au sein de notre équipe. Au vu de cette réussite, il est désormais prévu d'intégrer ce genre d'analyse de manière automatique et systématique au sein de notre pipeline.

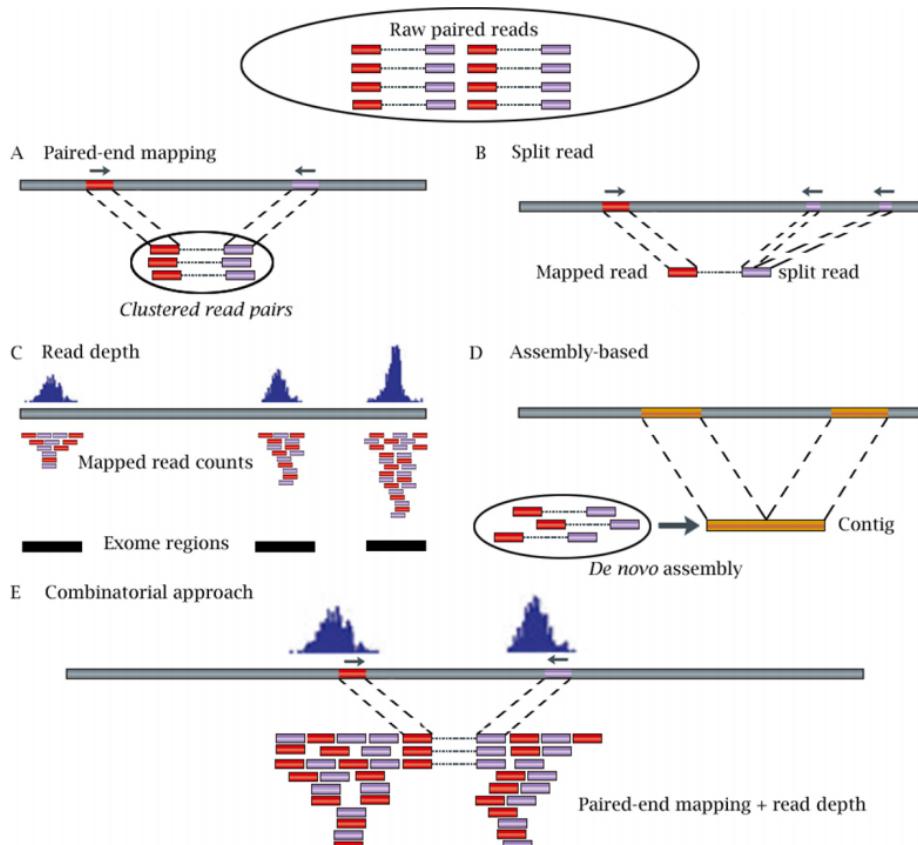


Figure 3.6 – Présentation de cinq approches permettant la détection de CNVs à partir de données NGS d'après [233] : **A** : Cette stratégie, permet de prédire des CNVs à partir des alignements discordants des deux *ends* d'un même *read*, c'est à dire en répertoriant les *reads* pour lesquels la distance séparant les deux *ends* après l'alignement est significativement supérieure à la taille moyenne de l'insert. **B** : La méthode *split-read* se base sur les *reads* s'alignant de manière partielle sur plusieurs régions génomiques. **C** : L'approche *read depth* compare la couverture observée sur plusieurs régions génomiques pour prédire des CNVs. **D** : Cette méthode effectue un assemblage *de novo* (sans utilisé de génome de référence) ; les résultats de l'assemblage appelés *contigs* sont comparés au génome de référence *a posteriori* pour détecter les CNVs. **E** : Cette méthode combine les approches **A** et **C**.

Pour les patients n'ayant eu aucun candidat identifié, il est possible que le choix de la stratégie du séquençage exomique plutôt que du génome entier ait masqué la cause génétique du phénotype de certains de nos patients. En effet, dans ces analyses, nous nous sommes concentrés sur l'analyse des variants situés dans les parties codantes **uniquement**. Ainsi les variants situés par exemple dans les microARN n'ont pu être observés. Or, les microARN jouent un rôle important dans la régulation génique principalement en influant sur la stabilité d'ARNm cibles et sont présent en grande quantité au sein des cellules germinales et leur importance dans la spermatogenèse a déjà été démontrée chez la souris [235] ainsi que plus récemment chez d'autres mammifères dont l'humain [236] laissant penser que des défauts altérants ces microARN pourraient entraîner des dysfonctionnements de la spermatogenèse. Aussi, il faut noter que les analyses WES et WGS ne permettent pas d'observer les défauts épigénétiques, or, ceux-ci représentent une part croissante des causes impliquées dans les cas d'infertilité masculine [237–239]. Aussi, au vu du grand nombre de gènes impliqués dans la spermatogénèse il est très possible que les causes génétiques responsables d'un même phénotype puissent être très hétérogènes. Par exemple, dans le cas de l'analyse de la cohorte de patients MMAF, 3630 variants subsistaient après avoir appliqué l'ensemble des filtres. Ces variants impactaient 2780 gènes différents parmi lesquels 1684 étaient retrouvés mutés chez uniquement un seul des 78 patients de la cohorte. Au vu de ce nombre important de gènes, il est très compliqué d'effectuer des analyses poussées sur l'ensemble d'entre eux. Dès lors, il est possible que la cause génétique responsable du phénotype d'un patient soit "noyée" parmi les nombreux variants restant mettant ainsi en évidence la nécessité de créer de nouveaux filtres afin de pouvoir réduire encore cette liste.

C'est dans ce but que notre équipe travaille actuellement au développement du score MutaScript. Ce score a pour but de classer l'ensemble des transcrit codant en fonction de leur charge mutationnelle avec l'idée sous-jacente que les transcrits les plus mutés dans la population générale ne sont probablement pas impliqués dans des pathologies sévères à transmission Mendélienne, et *a contrario* ceux retrouvés comme n'étant pas / peu mutés le sont probablement. Pour ce faire, le score MutaScript repose sur trois informations principales. La première étant le jeu de transcrits fourni par Ensembl [199]. Afin de connaître la charge mutationnelle de ces transcrits, nous nous sommes basées sur les variants mis à disposition par ExAC [171] qui réunit les données d'exome de 60.706 individus non apparentés que nous avons ensuite annoté grâce au logiciel *Variant Effect Predictor* [172] afin de prédire l'impact de chaque variant sur l'ensemble des transcrits qu'ils chevauchent de sorte à ce que les variants ayant un impact prédict comme étant délétère aient une plus grosse contribution au score MutaScript que ceux ayant un impact faible. Ces dernières années, des scores tel que le *residual variance intolerance score* (RVIS) [184] ou encore *the Probability of loss-of-function Incoherency* (pLI) [171] ont vu le jour. MutaScript se présente comme une alternative à ces derniers scores et, bien que sa fonction soit similaire, il diffère de ceux-ci sur de nombreux points. Tout d'abord, MutaScript donne un score à l'ensemble des transcrits codant pour une protéine là où pLI donne un score seulement au transcrit consensus de chaque gène et RVIS qui agrège les séquences codantes

de l'ensemble des transcrits d'un même gène créant ainsi un transcrit “chimérique”. Ce procédé facilite l'interprétation du score mais engendre une perte d'information puisque l'on se retrouve avec un seul score par gène et non par transcrits. De plus, dans la conception de leur score, RVIS et pLI ne considèrent que les variants dit *loss-of-function* (LoF), c'est à dire les variants impactant l'épissage, engendrant un codon stop ou un décalage du cadre de lecture. Cependant, ces variants ne représentent qu'une faible proportion des variants fournis par la base de données ExAC. C'est pourquoi, MutaScript prend en compte l'ensemble des variants, quelque que soit leur impact sur les différents transcrits qu'ils chevauchent, et leur attribue un poids en fonction de cet impact de sorte à ce que les variants considérés comme étant les plus délétères contribuent plus au score d'un transcrits que les autres. Aussi, l'étude des scores RVIS et pLI nous a permis de mettre en évidence une corrélation forte entre le score qu'ils attribuent à un gène et la taille de la séquence codante (CDS) de ce même gène. Cette corrélation étant due à un biais causé par leur manière de calculer leur score et non à une réalité biologique, MutaScript est construit de sorte à éviter cette corrélation qui peut mener à des erreurs d'interprétations. Le développement de ce score est en cours de finalisation.

Après avoir fait ses preuves dans la recherche, le séquençage NGS joue un rôle de plus en plus important dans le domaine du diagnostic clinique et on peut logiquement penser que ces technologies remplaceront bientôt la plupart des techniques diagnostiques actuelles. Il est néanmoins légitime de se demander dès aujourd'hui quelle sera son efficacité. En effet, en se basant sur les données de nos 78 patients non apparentés souffrant du syndrome MMAF on peut s'attendre, à obtenir entre 747 et 2499 variants par patients avec un coefficient de variation de 29% (avec $C_v = \frac{\sigma}{\mu}$ où σ est l'écart-type et μ la moyenne). Ces chiffres sont obtenus en filtrant de notre ensemble de variants ceux ayant une fréquence ≥ 0.01 dans les bases de données publiques ainsi qu'en retirant les variants introniques, synonymes et ceux impactant les séquences UTRs. On constate alors que parmi l'ensemble de ces variants 26815 d'entre eux (26%) sont “individuels”, c'est à dire porté uniquement par un seul des patients (**Figure : 3.7 - A**). De même on peut observer que chacun de nos patients est porteur d'environ 153 variants tronquants parmi lesquels 78 le sont à l'état homozygote (**Figure : 3.7 - B**). La priorisation de certains gènes par des outils tel que MutaScript permettra alors d'orienter les analyses vers les gènes les plus prometteurs. De même la recherche de variants parmi des panels de gènes permettra également de cibler les recherches sur les gènes déjà connus comme étant lié au phénotype en question. Ainsi, au vu de nos résultats dans le cadre de patients atteints du syndrome MMAF, la recherche de variants dans les gènes *DNAH1*, *CFAP43*, *CFAP44* et *WDR66* permettrait d'obtenir un diagnostic positif dans environ 36% des cas. On peut dès lors s'attendre que les recherches futures vont permettre d'agrandir cette liste de gènes cibles améliorant ainsi l'efficacité du diagnostic.

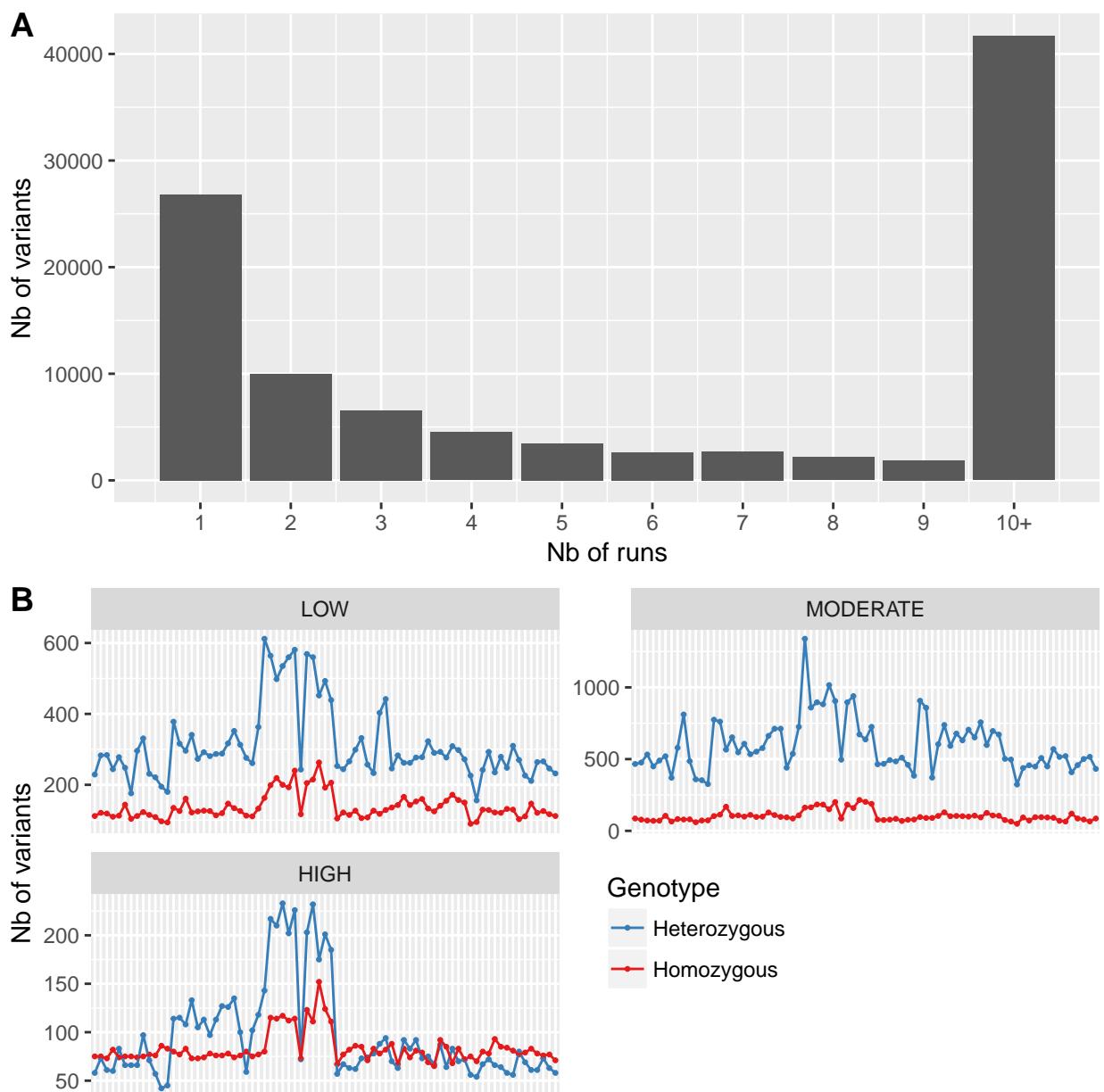


Figure 3.7 – Analyse des variants restant sur chacun des 78 patients après filtrage : A : Répartition du nombre de patients portant un même variant. Les variants portés par 10 patients ou plus sont regroupés sous l'étiquette 10+. On voit ici clairement qu'une grande majorité des variants sont spécifiques à un seul patient. B : Sur cette figure chaque point représente un patient. Les variants homozygote sont représentés en rouge, les variants hétérozygotes en bleu. L'impact *LOW* correspond aux variants ayant un impact peu délétère sur le gène tel que les variants introniques situés dans les zones d'épissage lointaines. Les variants *MODERATE* tel que les faux-sens ont un impact modéré sur le gène. Les *HIGH* représentent les variant tronquants (décalage du cadre de lecture, codon stop...).

Pendant de nombreuses années, la Science et la Technique étaient considérées comme des disciplines distinctes. Elles étaient pratiquées, dans la grande majorité des cas, de manière indépendante l'une de l'autre et surtout par des personnes différentes n'entretenant que peu d'interactions. Bien que la distinction entre Science et Technique soit réelle, la première pouvant être définie comme la quête de la connaissance et de la compréhension du monde tandis que la seconde met en œuvre un ensemble de moyen afin de modifier celui-ci d'une manière déterminée à l'avance, l'interdépendance liant ces deux notions n'a jamais été aussi forte qu'à notre époque tant et si bien qu'elles sont souvent confondues. En effet, il est courant d'entendre parler de progrès scientifique pour présenter une innovation technologique et *vice versa*. Ainsi, si la Science n'est pas la Technique, elle est dans de nombreux cas dépendante de celle-ci. En effet, comme nous avons pu le voir, l'étude et la connaissance du génome ont dû attendre les progrès techniques permettant notamment le séquençage de l'ADN. La Technique, elle, n'a pas nécessairement besoin de savoirs scientifiques pour être conçue : des savoirs empiriquement acquis suffisent à l'application d'une technique. Par exemple, bien qu'ils n'aient eu aucune conscience des mécanismes scientifiques sous-jacents, les premiers hommes ont su maîtriser plusieurs techniques de production et d'entretien du feu. De la même manière les agriculteurs n'ont pas eu besoin d'attendre et de comprendre les travaux sur la génétique et l'hérédité pour observer que la mise en reproduction des bêtes les plus productives permettait de maximiser les chances que la descendance soit elle aussi très productive. Cependant la Technique utilise de plus en plus des connaissances scientifiques et a ainsi finit par beaucoup dépendre d'elle en utilisant et appliquant des savoirs scientifiques. Ainsi est née la Technologie. Les travaux décrits dans ce manuscrit illustrent parfaitement cette relation d'interdépendance entre la Science et le Technique / Technologie. En effet, la connaissance du génome a été permise par l'émergence des différentes technologies de séquençage qui s'appuient elles aussi sur de nombreuses connaissances scientifiques. On peut dès lors s'attendre à ce que Science et Techniques / Technologie continuent d'évoluer de manière concomitante en s'entre alimentant. Dès lors, on peut prédire que les prochains progrès technologiques seront à l'origine de découvertes scientifiques qui serviront elles même à la fois de socle mais aussi de guide aux évolutions technologiques futures.

CHAPITRE A

Article annexe

Mutations in DNAH1, which Encodes an Inner Arm Heavy Chain Dynein, Lead to Male Infertility from Multiple Morphological Abnormalities of the Sperm Flagella

Ben Khelifa M*, Coutton C*, Zouari Raoudha, **Karaouzène T**, Rendu J, Bidart M, Yassine S, Pierre V, Delaroche J, Hennebicq S, Grunwald D, Escalier D, Pernet-Gallay K, Jouk PS, Thierry-Mieg N, Touré A, Arnoult C, Ray PF

American Journal of Human Genetics, Janvier 2014

* Co-premiers auteurs

Mutations in *DNAH1*, which Encodes an Inner Arm Heavy Chain Dynein, Lead to Male Infertility from Multiple Morphological Abnormalities of the Sperm Flagella

Mariem Ben Khelifa,^{1,2,3,15} Charles Coutton,^{1,2,4,15} Raoudha Zouari,⁵ Thomas Karaouzène,^{1,2} John Rendu,^{1,6,7} Marie Bidart,^{1,7} Sandra Yassine,^{1,2} Virginie Pierre,^{1,2} Julie Delaroche,^{1,7} Sylviane Hennebicq,^{1,2,8} Didier Grunwald,^{1,7} Denise Escalier,⁹ Karine Pernet-Gallay,^{1,7} Pierre-Simon Jouk,^{10,11} Nicolas Thierry-Mieg,¹⁰ Aminata Touré,^{12,13,14} Christophe Arnoult,^{1,2,16} and Pierre F. Ray^{1,2,6,16,*}

Ten to fifteen percent of couples are confronted with infertility and a male factor is involved in approximately half the cases. A genetic etiology is likely in most cases yet only few genes have been formally correlated with male infertility. Homozygosity mapping was carried out on a cohort of 20 North African individuals, including 18 index cases, presenting with primary infertility resulting from impaired sperm motility caused by a mosaic of multiple morphological abnormalities of the flagella (MMAF) including absent, short, coiled, bent, and irregular flagella. Five unrelated subjects out of 18 (28%) carried a homozygous variant in *DNAH1*, which encodes an inner dynein heavy chain and is expressed in testis. RT-PCR, immunostaining, and electronic microscopy were carried out on samples from one of the subjects with a mutation located on a donor splice site. Neither the transcript nor the protein was observed in this individual, confirming the pathogenicity of this variant. A general axonemal disorganization including mislocalization of the microtubule doublets and loss of the inner dynein arms was observed. Although *DNAH1* is also expressed in other ciliated cells, infertility was the only symptom of primary ciliary dyskinesia observed in affected subjects, suggesting that *DNAH1* function in cilium is not as critical as in sperm flagellum.

Male infertility affects more than 20 million men worldwide and represents a real health concern.¹ It is a typical multifactorial disorder with a strong genetic basis and additional etiological factors such as urogenital infections, immunological or endocrine diseases, attack from reactive oxygen species (ROS), or perturbations from endocrine disruptors. To date, despite substantial efforts made to identify genes specifically involved in male infertility by many teams including ours,^{2,3} only a handful of genes have been formally correlated with human sperm defects. Male infertility caused by impaired sperm motility (asthenozoospermia) is also often observed in men with primary ciliary dyskinesia (PCD), a group of mainly autosomal-recessive disorders caused by dysfunctions of motile cilia leading primarily to respiratory infections and often to situs inversus. Recent research on PCD has been extremely prolific and allowed the identification and characterization of numerous proteins necessary for adequate axonemal molecular structure and assembly (Table S1 available online). The axoneme is a highly evolutionarily conserved structure found in motile cilia and in sperm flagella,

mainly composed of an intricate network of microtubules and dyneins. Sperm parameters have not been systematically explored and are often only scarcely described in manuscripts investigating PCD-affected individuals. Although sperm flagella and motile cilia have a similar axonemal structure based on the presence of nine peripheral microtubule doublets plus two central ones, they present several differences that might explain why PCDs are not always associated with asthenozoospermia.⁴ We note that no mutations in axonemal genes have been described as being involved exclusively in infertility without also inducing PCD.

In the present study, we analyzed 20 subjects presenting with asthenozoospermia resulting from a combination of five morphological defects of the sperm flagella (absent, short, bent, and coiled flagella and flagella of irregular width) without any of the other PCD-associated symptoms. Similar phenotypes have been previously described and named "dysplasia of the fibrous sheath," "short tails," or "stump tails."^{5–15} We propose to call this syndrome "multiple morphological anomalies of the flagella

¹Université Joseph Fourier, Grenoble 38000, France; ²Laboratoire AGIM, CNRS FRE3405, Equipe "Andrologie et Génétique," La Tronche 38700, France; ³Laboratoire de génomique Biomédicale et Oncogénétique, Institut Pasteur de Tunis, 1002 Tunis, Tunisie; ⁴CHU de Grenoble, Hôpital Couple Enfant, Département de Génétique et Procréation, Laboratoire de Génétique Chromosomique, Grenoble 38000, France; ⁵Clinique des Jasmins, 23, Av. Louis BRAILLE, 1002 Tunis, Tunisia; ⁶CHU de Grenoble, Institut de Biologie et Pathologie, Département de Biochimie, Toxicologie et Pharmacologie (DBTP), UF de Biochimie et Génétique Moléculaire, Grenoble 38000, France; ⁷INSERM, U836, Grenoble Institute of Neuroscience, La Tronche 38700, France; ⁸CHU de Grenoble, Hôpital Couple Enfant, Département de Génétique et Procréation, Laboratoire d'Aide à la Procréation – CECOS, Grenoble 38000, France; ⁹INSERM UMR_S933, Université Pierre et Marie Curie (Paris 6), Paris 75012, France; ¹⁰Université Joseph Fourier-Grenoble 1 / CNRS / TIMC-IMAG UMR 5525, Grenoble 38041, France; ¹¹CHU de Grenoble, Hôpital Couple Enfant, Département de Génétique et Procréation, Service de Génétique Clinique, Grenoble 38000, France; ¹²INSERM, U1016, Institut Cochin, Paris 75014, France; ¹³CNRS, UMR8104, Paris 75014, France; ¹⁴Université Paris Descartes, Sorbonne Paris Cité, Faculté de Médecine, Paris 75014, France

¹⁵These authors contributed equally to this work

¹⁶These authors contributed equally to this work and are co-senior authors

*Correspondence: pray@chu-grenoble.fr

<http://dx.doi.org/10.1016/j.ajhg.2013.11.017>. ©2014 by The American Society of Human Genetics. All rights reserved.

(MMAF)," a name that provides a more accurate description of this phenotype. We carried out a SNP whole-genome scan on 20 individuals presenting with severe MMAF. The study was approved by our local ethics committee; all individuals gave their signed informed consent and national laws and regulations were respected. All individuals originated from North Africa (11 Tunisians, 7 Algerians, and 2 Libyans) and were treated in Tunis (Clinique des Jasmins, Tunis, Tunisia) for primary infertility. Twelve of the subjects were born from related parents, usually first cousins. None of the subjects were related to one another apart from three individuals (P1–P3) who were brothers. All subjects had normal somatic karyotypes. All sperm analyses were performed at least twice, in accordance with the World Health Organization recommendations.¹⁶ Subjects were recruited on the basis of the identification of >5% of at least four of the aforementioned flagellar morphological abnormalities (absent, short, coiled, bent, and irregular flagella) (Table 1). All subjects presented with severe asthenozoospermia: 11 out of 20 subjects had no (0%) motility, 8 had sperm motility <10%, and one (P6) had 35% motility. Saliva was obtained from all participants via Oragene DNA Self-Collection Kit (DNAgenotech) but only one subject (P3) agreed to donate sperm and blood samples for research use. During their medical consultation for infertility, all subjects answered a health questionnaire focused on PCD manifestations, and none indicated suffering from any of the other symptoms encountered in PCD.

Homozygosity mapping was carried out with 250K Sty1 SNP mapping arrays (Affymetrix) on DNA extracted from the 20 studied subjects' saliva samples. Common regions of homozygosity were identified with the homoSNP software. After exclusion of the centromeric regions, we identified two regions located on chromosomes 3 and 20 with a region of homozygosity > 1 Mb common to 10/20 analyzed individuals (Figure S1). In addition, 4 and 9 subjects presented with a stretch of homozygosity > 15 Mb overlapping chromosomes 20 and 3 regions, respectively. All three brothers (P1–P3) were homozygous at the chromosome 3 region, although only two of them were homozygous at the chromosome 20 region. We excluded all other regions of homozygosity because they did not fulfil the following criteria: (1) more than eight individuals including at least two of the brothers sharing a region of homozygosity > 1 Mb and (2) presence of a potential candidate gene in the region according to its expression profile and/or presumed function. Finer analysis of the chromosome 3 region showed that 15 individuals were homozygous for two smaller subregions located at chr3: 46,745,396–47,606,570 and chr3: 52,111,974–53,028,375 (UCSC Genome Brower human reference genome build hg17, Figure S1). Sixteen genes are annotated in the first subregion (Table S2), among which only one gene (*KIF9* [MIM 607910]) appeared as a good candidate; indeed, studies in the protist *Trypanosoma brucei* showed that *kif9A* (the mouse ortholog of human *KIF9*) is located in

the axoneme and that its depletion alters motility.¹⁷ The second subregion in chromosome 3 includes 28 genes (Table S2). The dynein heavy chain 1 gene (*DNAH1* [MIM 603332]) appeared as the best candidate gene because it codes for an axonemal dynein heavy chain and is expressed in various tissues including testis.¹⁸ Furthermore, asthenozoospermia was described in mice lacking *Dnahc1*, the *DNAH1* mouse ortholog (previously named *Mdhc7*).¹⁹ Finally, among the ten genes located in the selected region of chromosome 20 (chr20: 33,572,687–34,070,415), only *SPAG4* (MIM 603038) appeared as a good candidate: it was described in rat to be associated with the axoneme in elongating spermatids and epididymal sperm.²⁰ We therefore decided to sequence *KIF9* (RefSeq accession number NM_001134878.1), *DNAH1* (RefSeq NM_015512.4), and *SPAG4* (RefSeq NM_003116.1).

We sequenced the 12 exons and the intron boundaries of *SPAG4* in the 13 individuals homozygous at this locus, and the 19 exons and intron boundaries of *KIF9* in the 15 relevant individuals. We did not identify any likely pathogenic variants in these two genes. We then sequenced the 78 exons and intron boundaries of *DNAH1* in P3 (primer sequences available in Table S3). We identified one homozygous splicing mutation (c.11788–1G>A) in intron 73. The same homozygous mutation was identified in the two other brothers (Figure S2). We then sequenced *DNAH1* for the 17 remaining subjects. The same homozygous mutation (c.11788–1G>A) was identified in one additional individual (P17). We identified three other homozygous variants: another splicing mutation (c.5094+1G>A) in individual P9, a homozygous no-stop mutation disrupting the stop codon in exon 78 (c.12796T>C [p.4266Glnext*21]) in individual P8, and a homozygous missense variant in exon 23 in individual P6 (c.3877G>A [p.Asp1293Asn]). The localization of the *DNAH1* mutations is presented in Figure 1. If we consider only index cases, we identified 5 homozygous variants in 18 unrelated individuals (28%). None of these variants were detected in our control cohort of 100 individuals of North African origin. We note that the parents of the subjects could not be analyzed to confirm the transmission of the variants. We therefore cannot formally exclude the possibility that some of the identified variants may be hemizygous with a deletion on the other allele. However, depending on its size, its position, and its effect on the reading frame, a deleted allele would be at least as deleterious as the identified variants.

To evaluate the association of the variants with the pathology, we compared their frequency in our cohort with that in the Exome Variant Server (EVS) database. At the four genomic positions of interest, the EVS data are of sufficient coverage to provide genotype calls for at least 6,200 individuals, corresponding to 12,400 alleles. There were no variant nucleotides identified at positions c.5094+1, c.11788–1, or c.12796, and only one A allele was identified out of 12,460 alleles at position c.3877.

Table 1. Semen Parameters of the 20 Subjects and the 7 Subjects Carrying *DNAH1* Homozygous Variants

Semen Parameters	Average of 20 Subjects ^a	P1	P2	P3	P6	P8	P9	P17
DNAH1 mutations	c.11788–1G>A (p.Gly3930Alafs*120)	c.11788–1G>A (p.Gly3930Alafs*120)	c.11788–1G>A (p.Gly3930Alafs*120)	c.3877G>A (p.Asp1293Asn)	c.12796 T>C (p.Leu1700Serfs72)	c.5094+1G>A (p.Leu1700Serfs72)	c.11788–1G>A (p.Gly3930Alafs*120)	
Consanguinity	yes	yes	yes	no	yes	yes	yes	no
Origin of the subject	Tunisia	Tunisia	Tunisia	Algeria	Algeria	Algeria	Algeria	Tunisia
Sperm volume (ml)	3.2 (1–5.5)	5	2.5	2	5	4.5	3.5	2.5
Sperm concentration 10 ⁶ /ml	22 (0–59)	45	0	2.8	57	11	53	31
Motility (A+B) 1 hr	2.5 (0–35)	0	NA	2	35	0	0.5	0
Vitality	44 (6–73)	22	NA	NA	73	61	48	NA
Normal spermatozoa	0.35 (0–6)	0	NA	0	6	0	0	0
Absent flagella	30 (8–46)	34	+	34	+	+	16	30
Short flagella	44 (16–70)	38	+	44	+	+	70	20
Coiled flagella	13 (2–32)	14	NA	14	+	+	12	32
Angulation	12 (2–19)	4	NA	2	+	+	8	6
Flagella of irregular caliber	55 (16–92)	48	NA	50	+	+	54	16
Multiple anomalies index	2.9 (1.9–3.9)	3.1	NA	2.4	NA	NA	3	2.6

Values are expressed in percents, unless specified otherwise. Abbreviations are as follows: NA, not available; plus sign, anomalies reported (>5%) but not accurately quantified.

^aValues are expressed as the mean with the lower and higher values in parentheses.

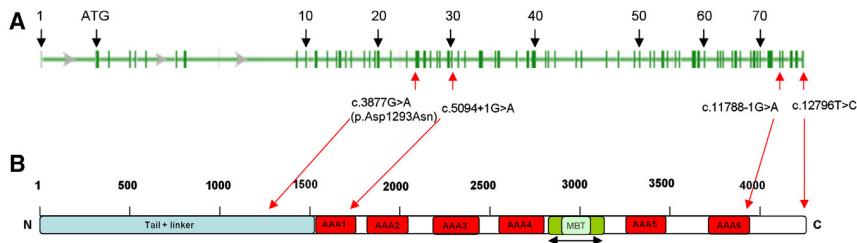


Figure 1. Location of *DNAH1* Mutations in the Intron-Exon Structure and in the Protein Representation of *DNAH1*

(A) *DNAH1* genomic structure.
(B) *DNAH1* domain map showing the location of the four identified mutations. The red boxes indicate the six known AAA-ATPase domains (AAA 1 to 6) as detected by homology (Uniprot server). The microtubule-binding domain (MBT) lies between AAA4 and AAA5. The N-terminal part of the protein binds to the intermediate, light-intermediate dynein chains. The position of the stalk and the microtubule-binding domain (MTB) are indicated.

We performed Fisher exact tests (with `fisher.test` in R) to evaluate whether each observed SNV was statistically overrepresented in our cohort of 18 unrelated individuals, compared to EVS. All four individual SNVs are significantly enriched (p values: 5.9×10^{-11} for $c.11788-1G>A$, 8.1×10^{-6} for $c.5094+1G>A$ and $c.12796T>C$, and 2.4×10^{-5} for $c.3877G>A$). We then investigated whether *DNAH1* as a whole was significantly enriched in damaging SNVs in our cohort. Overall, EVS contains five nonsense and two splice-site SNVs, all observed heterozygously in a total of ten individuals. By using the coverage data available on the EVS website, we find that 20,737 positions covering the *DNAH1* exons and intron boundaries had sufficient sequence coverage to be genotyped in 6,189 individuals on average. By contrast, counting only the two splice-site mutations as damaging, we observe 6 damaging alleles among 36 in our cohort: this represents a highly significant enrichment (Fisher exact test, p value: 3×10^{-12}). Furthermore, we note that there were no homozygous damaging variants observed in the 6,189 EVS individuals compared to 3 in our cohort of 18 (Fisher exact test, p value: 3×10^{-8}). Altogether we believe that these genetic results convincingly demonstrate that mutations in *DNAH1* are associated with MMAF.

The $c.5094+1G>A$ variant found in individual P9 affects *DNAH1* intron 31 consensus donor splice site. The abnormal splicing is predicted to cause the prolongation of exon 31 until the introduction of a nonsense codon (p.Leu1700Serfs72). The position of the next donor site was predicted by "Splice Site Prediction by Neural Network." Unfortunately we could not obtain any leukocytes from this subject to validate this prediction and observe whether this variant also led to nonsense-mediated mRNA decay (NMD). Because of the location of the variant on a consensus splice site and the unambiguous predictions of splice prediction software, we did not synthesize a minigene to verify the effect of this variant in vitro. A missense change, p.Asp1293Asn, was identified in P6. Interestingly, the Asp1293 amino acid is well conserved across species (Figure S3). This missense change is also predicted to be possibly damaging by SIFT and Poly-Phen-2, two prediction softwares for nonsynonymous SNPs. It affects the N-ter of the protein (Figure 1B), known to be important for the structure of dynein arms.²¹ Variant

$p.4266Glnext*21$ found in individual P8 abolishes the stop codon in exon 78, leading to the addition of 21 codons at the 3' end of the coding sequence. The role of the C-terminal domain is uncertain, but based on the *D. discoideum* structures, it may participate in long-range allosteric communication between microtubule-binding and ATPase regions.^{22,23} The addition of 21 extra amino acids to this region is likely to disrupt these interactions.

The $c.11788-1G>A$ variant identified in four subjects (P1–P3 and P17) affects the final G nucleotide of *DNAH1* intron 73, one of the consensus splice acceptor nucleotides. The resulting abnormal splicing is predicted to recognize a new CG acceptor site located just one nucleotide further, thus shifting the reading frame and inducing a premature stop codon (p.Gly3930Alafs*120). As could be expected, P1–P3 share a common haplotype (Table S4). P17 also shares a common haplotype of 30 SNPs with P1–P3, suggesting a founder effect for this mutation. To assess the functional impact of the *DNAH1* splice acceptor site mutation $c.11788-1G>A$, we studied mRNA products isolated from control and P3 lymphocytes (primer sequences available in Table S5). RT-PCR of P3's samples yielded no product despite repeated attempts, whereas the three amplification attempts from control lymphocytes yielded the expected product (Figure 2A). RT-PCR targeting *GAPDH* (MIM 138400) and *RPLPO* (MIM 180510) confirmed the integrity of P3's RNA (Figure 2B). This suggests a specific degradation of the mutant *DNAH1* transcripts by NMD. To further validate the pathogenicity of this variant, we analyzed *DNAH1* localization in sperm from P3 by immunofluorescence and the ultrastructure of the flagella by electron microscopy. In control individuals, *DNAH1* antisera decorated the full length of the sperm flagellum (Figure 2C), suggesting a putative role in the tethering of the inner dynein arms along the entire axoneme. In contrast, in sperm from individual P3 carrying the $c.11788-1G>A$ mutation, *DNAH1* immunostaining was absent, confirming that the splicing defect results in the degradation of the transcripts by NMD (Figure 2D). We next tested the integrity of the outer and inner dynein arms by using antibodies directed against *DNALI1* and *DNAI2*, two well-established diagnostic markers of the inner and outer dynein arms, respectively. Staining with *DNALI1* was strongly reduced in the sperm of individual

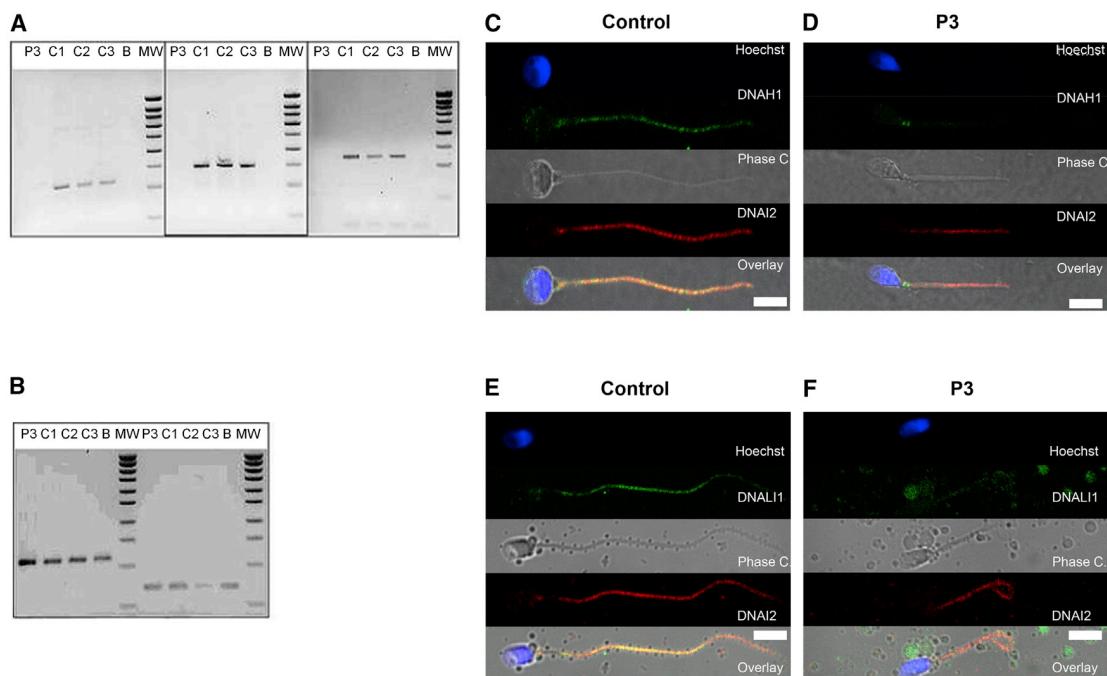


Figure 2. Analysis of P3 Carrying the c.11788–1G>A Variant Evidencing DNAH1 mRNA Decay by RT-PCR and the Absence of DNAH1 in Sperm by Immunolocalization

(A) RT-PCR analyses of subject P3 (c.11788–1G>A homozygote) and control individuals from the general population (C1–C3). Electrophoresis showing the RT-PCR amplification of *DNAH1* exons 30–32, 64–66, 73–75. C1, C2, and C3 yield a normal fragment of 228, 293, and 241 bp, whereas subject P3 shows no amplification. There is no amplification from the RT-negative blank control (column B).

(B) Electrophoresis showing the amplification of the same cDNAs with *GAPDH* and *RPL0* primers. Bands of equivalent intensity are obtained from all samples including P3. Reverse transcription was carried out with 500 ng of extracted RNA and oligo dT priming. Two microliters of the obtained cDNA mix was used for the subsequent PCR. PCR amplification was carried out with three couples of primers located in exons 30–32, 64–66, and 73–75 of *DNAH1* at an elongation temperature of 57°C (40 cycles), in parallel to amplification of the same samples with the control housekeeping *GAPDH* and *RPL0*, respectively, at an elongation temperature of 60°C (35 cycles). RT-PCR primers are listed in Table S4.

(C and D) Immunofluorescence staining of human spermatozoa with DNAH1 antibodies (green) and DNAI2 (red). DNAH1 is observed throughout the flagellum in control sperm, whereas it is absent from P3's sperm. In both control and P3 sperm, ODA is present as witnessed by the immunostaining of DNAI2.

(E and F) Immunofluorescence staining of human spermatozoa with DNAL1 antibodies (green) and DNAI2 (red). DNAL1, a marker of IDA, is localized throughout flagella in control sperm, whereas it is strongly reduced in sperm from P3. No difference is noticed in the coimmunostaining of DNAI2. Sperm were counterstained with Hoechst 33342 (blue) as nuclei marker. White scale bars represent 5 μm. Sperm cells were washed in phosphate-buffered saline (PBS), fixed in 4% PFA for 2 min at room temperature (RT), and washed twice in PBS. Fixed spermatozoa were allowed to air-dry on poly-L-lysine coated slides followed by permeabilization with 0.5% Triton X-100. Samples were then blocked with (PBS)/1% bovine serum albumin (BSA)/2% normal goat serum (NGS) for 30 min at RT. Slides were incubated with the primary antibodies 2 hr followed by an incubation with the secondary antibodies for 45 min at RT and mounting in Dako mounting medium (Dako). Appropriate controls were performed, omitting the primary antibodies. Polyclonal mouse DNAL1 and monoclonal mouse DNAI2 were purchased from Abcam (UK) and Abnova Corporation (Taiwan), respectively. Polyclonal DNAH1 antibodies were purchased from Prestige Antibodies (Sigma-Aldrich). Monoclonal mouse anti-acetylated-α-tubulin were purchased from Sigma-Aldrich. Highly cross-adsorbed secondary antibodies (Alexa Fluor 488 and Alexa Fluor 546) were obtained from Molecular Probes (Invitrogen).

P3, suggesting that inner arms were mostly absent in this individual (Figure 2F). On the other hand, the antibodies directed against DNAI2 stained the sperm flagella in both control and individual P3, suggesting that the outer dynein arms were not affected by the absence of DNAH1 (Figures 2E and 2F). In order to confirm that the inner arms were disorganized, we studied the ultrastructure of individual P3's sperm by transmission electron microscopy (TEM) (Figure 3). We could observe 40 doublets of microtubules in cross sections presenting a sufficient quality to observe the dynein arms. Fifteen outer dynein arms (ODA) and only 4 inner dynein arms (IDA) were observed,

confirming the complete disorganization of the IDA. Moreover, approximately one third of the microtubule doublets were malformed or absent in the observed sections. Furthermore, the central singlet of microtubules was missing (9+0) in 47% of these sections. The fibrous sheath was also strongly disorganized in 90% of the sections (Figure 3).

After complete DNA sequencing of *DNAH1*, we identified two variants altering a consensus splice site, highly likely to have a damaging effect, in 3 out of 18 unrelated individuals. The predicted effect of the other two identified variants is not as clear but the addition of 21 residues

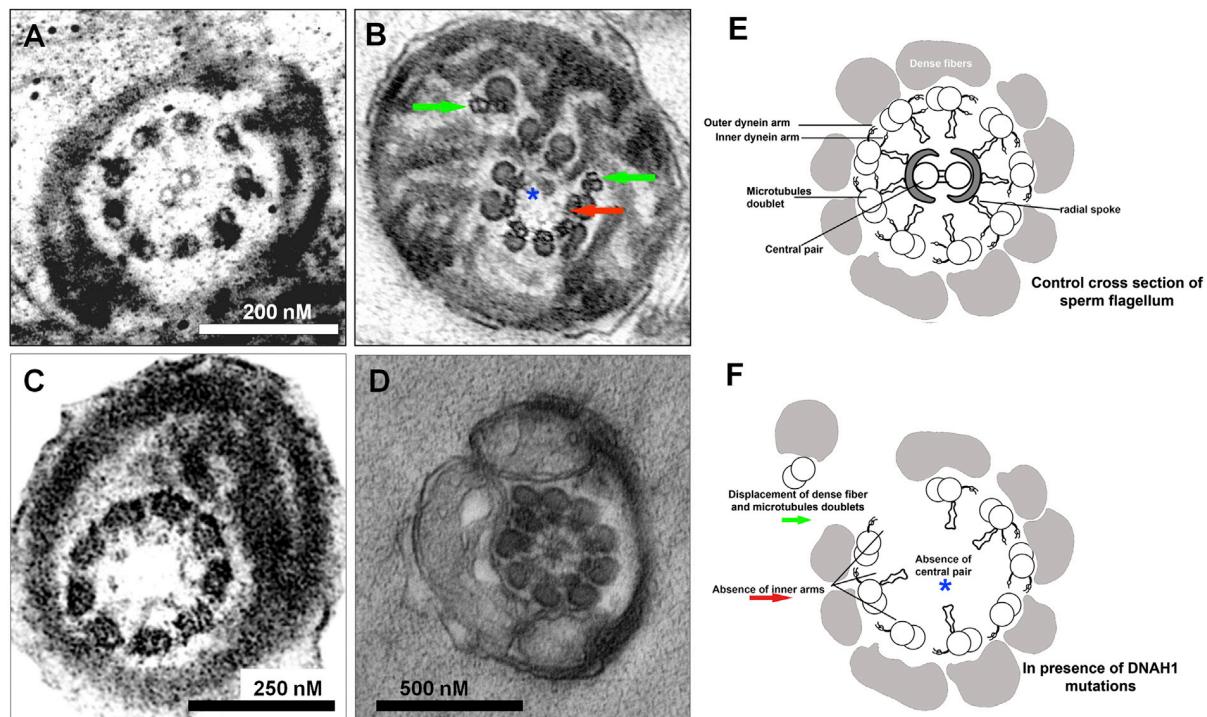


Figure 3. Electron Microscopy Analysis of Spermatozoa from P3, Carrying the c.11788–1G>A Variant, Reveals Numerous Ultrastructural Defects

(A) EM cross-section of a flagellum from a control individual sperm sample.
 (B) Cross-section from the individual P3 sample showing numerous defects: lack of IDA (red arrow) and axonemal disorganization with mislocalized peripheral doublets (green arrows) associated with a displacement of the central pair (blue asterisk).
 (C) Cross-section from individual P3 sperm flagellum showing a complete absence of the central pair.
 (D) Cross-section from individual P3 sperm flagellum showing supernumerary dense fibers with absence of mitochondrion on the right side of the mid-piece.
 (E and F) Drawings describing the normal sperm axoneme ultrastructure with their different components (E) and different defects observed in *DNAH1* mutated subjects (F).

Sperm cells were fixed with 2.5% glutaraldehyde in 0.1 M cacodylate buffer (pH 7.4) during 2 hr at room temperature. Details of transmission electron microscopy technique were detailed previously.³

caused by the stop-loss variant is probably pathogenic. Interestingly, P1, P2, P3, P8, P9, and P17—who have inherited the variants predicted to have most severe effects—present 0% morphologically normal spermatozoa with a motility <2%, in contrast to P6, carrying the p.Asp1293Asn variant, who presents a milder phenotype with 35% motility and 6% morphologically normal spermatozoa (Table 1). These data therefore suggest that the c.3877G>A variant might be a hypomorphic allele, which is consistent with a single amino acid substitution in a large protein. Unfortunately we could not obtain any additional biological material from the other mutated subjects and in particular from P6 and could not assess the effects of the other variants on protein expression/localization and on the ultrastructure of the flagella. In addition, no mutations were identified in *DNAH1* in 13 subjects, suggesting that MMAF is genetically heterogeneous. We are currently sequencing the exomes of these 13 subjects in order to identify other genes involved in MMAF. We note that with the exception of P6, who carried a missense mutation and presented a milder form of the pathology, we included here only individuals with the most severe phenotypes.

We can therefore expect that individuals with intermediate asthenozoospermia and low levels of morphological anomalies could also harbor homozygous or compound heterozygous *DNAH1* mutations of moderate severity.

The data we present here are consistent with the phenotype described for *Dnahc1* knockout (KO) mice (the ortholog of *DNAH1*), which display asthenozoospermia and male infertility.¹⁹ In this model, however, no structural defects of the axoneme were observed either by optical or by transmission electronic microscopy.¹⁹ This contrasts with the strong axonemal disorganization we observed in sperm from P3 carrying the homozygous c.11788–1G>A mutation, where the inner dynein arms and the central pair of microtubules were mostly absent. In the *Dnahc1* KO, however, the authors describe that the targeted deletion did not lead to a complete disruption of the gene and resulted in a truncated protein with a preserved N terminus.¹⁹ Because the N-terminal part of the DyHCs plays a crucial role in the assembly and stabilization of the inner dynein arms, as shown in *Chlamydomonas* mutants,²⁴ it is likely that the formation of the base of the inner dynein arm is preserved and that the described

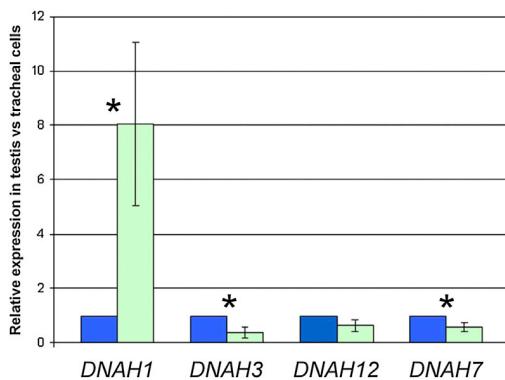


Figure 4. Relative mRNA Expression of DNAH1, DNAH3, DNAH7, and DNAH12 in Testis and in Tracheal Cells

Expression of *DNAH1* mRNA (in green) in testis is significantly higher (8-fold) than in tracheal cells (in blue). Tracheal cell expression is set to 1. Expression of *DNAH3* and *DNAH7* mRNA are significantly lower in testis than in tracheal cells, whereas *DNAH12* expression is not significantly different. Data are presented as mean \pm standard deviation of three independent quantitative real-time PCR experiments. Statistical tests (paired t test) with a two-tailed p value ≤ 0.05 were considered as significant (*). Human testis and trachea cDNA were obtained from Amsbio (Abingdon). mRNA expression were assessed by qPCR with a Biorad CFX9 (Biorad). PCR primers used to amplify *DNAH1* and three other inner dynein arm heavy chain genes (*DNAH3*, *DNAH7*, and *DNAH12*) and the reference gene *ACTB* are listed in Table S5. The PCR cycle was as follows: 10 min 95°C, 1 cycle; 10 s 95°C, 30 s 58°C + fluorescence acquisition, 55 cycles. Analysis was performed with Biorad software CFX Manager v.3.0, with advanced relative quantification mode. Values for each gene were normalized to expression level of beta-actin gene (*ACTB*) via the 2- $\Delta\Delta CT$ method.²⁵ The 2- $\Delta\Delta CT$ value was set at 0 in tracheal cells, resulting in an arbitrary expression of 1.

KO animals could ensure correct axonemal biogenesis and organization. Alternatively, it is possible that the *DNAH1* role in axonemal structure is not as central in mouse as it is in human.

Apart from infertility, none of the 20 individuals declared suffering from any of the principal PCD symptoms such as an impairment of the respiratory functions. This suggests that *DNAH1* function in cilia is probably compensated by other HC dyneins. Previous phylogenetic studies indicate that *DNAH3* (MIM 603334), *DNAH7* (MIM 610061), and *DNAH12* (MIM 603340) are close paralogs to *DNAH1*, *DNAH12* being the closest.¹⁸ We therefore measured the relative expression of these four IDA heavy chains to assess whether the expression of these proteins could compensate the absence of *DNAH1* in other ciliated tissues. By using qPCR (primer sequences available in Table S6), we showed that *DNAH1* is expressed at a much higher level (8-fold) in the testis compared to the trachea (Figure 4). Conversely, the other IDA heavy chains are expressed at higher levels (*DNAH3*, *DNAH7*) or at a similar level (*DNAH12*) in control tracheal cells as in testis (Figure 4). Data available from public expression databases show a similar expression pattern to what was observed here. Moreover, these data show that whereas *DNAH1* and *DNAH7* expression is restricted to ciliated cells,

DNAH1 and *DNAH12* expression is rather atypical, because it is almost ubiquitous (EST profile viewer and GeneHub-GEPIS). We therefore note that not only are *DNAH1* and *DNAH12* closest to one another from a phylogenetic point of view, but they also share a broad expression pattern. We therefore believe that *DNAH12* is the most likely candidate for a potential functional compensation of *DNAH1* in ciliated cells. In addition and/or alternatively to this compensation, we cannot exclude the possibility that some of the affected individuals might retain expression of *DNAH1* in the ciliated cells of the trachea, although this is not expected for the most severe variants identified. Alternatively, we cannot exclude a reduction of ciliary beats, which could lead to a small decrease of cilia function in respiratory epithelium or in other ciliated tissues without pathological consequences, or at least none that have been noticed by the affected men themselves. We could not obtain nasal brushings or curette biopsies from affected individuals and therefore cannot formally exclude this possibility. Future work on *DNAH1* mutated subjects should include a thorough analysis of PCD symptoms including nasal nitric oxide measurements, video microscopy, and transcription electron microscopy. This would provide valuable information regarding the role of *DNAH1* in ciliated cells as well as indicate whether mutated men might be at risk of developing PCD symptoms, perhaps as late onset.

Inner dynein arms are organized in seven molecular complexes, viewed in electron microscopy as globular heads arranged in 3-2-2 groups and corresponding to three different types of inner arms (IDA1 to IDA3, see Figure 5). In *Dnahc1* KO mice, electron microscopy studies indicated that one head of the IDA3 was missing, leading to a 3-2-1 globular head arrangement, suggesting that *DNAH1* is a component of IDA3. Radial spokes are present on microtubule doublets and interact with the inner arms. They allow a connection between external doublets of the microtubules and the two central microtubules. They are multiprotein complexes of more than 20 proteins. In mammals, there are three different radial spokes (RS1, RS2, and RS3) binding tightly to the inner arm bases of different IDAs. Among the different proteins involved in axonemal formation and organization, only mutations in *CCDC39* (MIM 613798) and *CCDC40* (MIM 613799) lead to a disorganization of the axonemal structure, a phenotype similar to what we observe in subjects with *DNAH1* mutations. *CCDC39* and *CCDC40* control the assembly of the dynein regulatory complex (DRC), a major regulatory node interacting with numerous axonemal structures.²⁷⁻²⁹ In the absence of DRC, RS2 anchoring is weakened, leading to the displacement or the absence of the central pair and the mislocalization of the peripheral doublets. Interestingly, in *T. thermophila*, the RS3 stalk is directly connected to the dynein d/a tail through an arc-like structure (Figure 5).²⁶ It can therefore be speculated that the absence of *DNAH1* removes the anchoring site of the radial spoke 3. As a consequence, the attachment of the two central

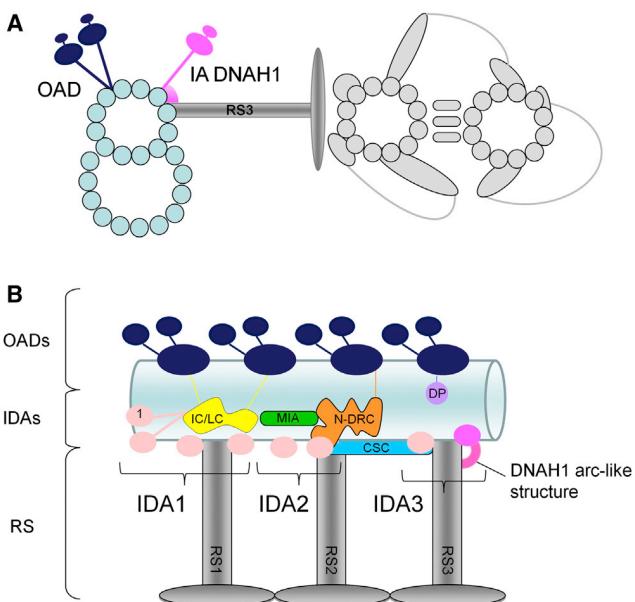


Figure 5. Proposed Schematic Model for the Location and the Function of the Inner Arm Heavy Chain DNAH1 in the Axoneme of Human Sperm Flagellum

(A) Simplified representation showing a cross-sectional view of one microtubule doublet of an axoneme surrounding the central pair complex; the viewing is from the flagellar base. Light gray, central pair complex; light blue, outer doublets; dark blue, outer arm dyneins (OAD); dark pink, inner arm dynein heavy chain DNAH1; dark gray, radial spoke 3.

(B) Longitudinal view illustrating the approximate localization on the outer doublet A-tubule of the various dyneins and regulatory structures within a single 96 nm axonemal repeat. RS3 stalk is directly connected to the DNAH1 tail through an arc-like structure. DNAH1 may therefore stabilize the RS3.²⁶ Light pink, inner arm dyneins (IAD); yellow, IC/LC, intermediate chain/light chain; orange, nexin-dynein regulatory complex (N-DRC); green, modifier of inner arms (MIA) complex; blue, calmodulin- and spoke-associated complex (CSC); purple, distal protrusion (DP).

singlet microtubules should be weakened. This scheme is in agreement with our data that show 47% of 9+0 axonemes (Figure 3).

As illustrated in Table S1, not all PCDs are associated with an infertility phenotype. For instance, subjects with mutations in *CCDC114* (MIM 615038)^{30,31} and *DNAH11* (MIM 603339)⁴ are fully fertile and could procreate spontaneously whereas subjects with *DNAAF2* (MIM 612517), *DNAH5* (MIM 603335), *DNAI1* (MIM 604366), or *HYDIN* (MIM 610812)^{4,32–34} present complete sperm immobility. Also, mutations in genes involved in the preassembly of the dynein arms *DNAAF1* (MIM 611390), *DNAAF2* (MIM 612517), *DNAAF3* (MIM 614566),^{32,35–37} and *LRRC6* (MIM 614930),³⁸ described to induce IDA loss, did not present axonemal microtubule disorganization. Consistently, IDAs are not or are only partially affected in deficient *Chlamydomonas* mutants for *ODA7* (*DNAAF1*), *ktu/pf13* (*DNAAF2*), and *pf22* (*DNAAF3*): in *ODA7* mutants, inner arms are not affected, in *ktu/pf13* mutants only IDA dynein c is missing, and in *pf22* mutants IDA dynein b and c are absent.^{27,29,36,37,39} Most importantly, IDA dynein f and

p28 remain located in the flagella. It can thus be speculated that there are several pathways for the preassembly and/or the targeting of the different mammalian IDAs, thus explaining the absence of axonemal microtubule disorganization in subjects presenting with mutations in *DNAAF1*, *DNAAF2*, *DNAAF3*, and *LRRC6*. Here we observed that DNALI1 immunostaining was strongly reduced along the whole flagellum (Figure 2E), suggesting that DNALI1 may be located mainly in IDA3.⁴⁰ In agreement with this result, p28, the *Chlamydomonas* ortholog of *DNALI1*, is associated with the inner dynein arm located in IDA3. Interestingly, *DNALI1* has an expression pattern similar to *DNAH1*: it presents a predominant testis expression and also a remarkable expression in nonciliated cells.⁴¹ Altogether, these facts suggest a close molecular partnership between DNAH1 and DNALI1.

Mutations affecting axonemal components and/or axoneme assembly often result in PCD, which frequently includes a male infertility phenotype. In this study we describe that mutations in *DNAH1*, which codes for an axonemal component, leads to male infertility only with no other apparent PCD-associated syndromes. Our data indicate that *DNAH1* is required in spermatozoa for the formation of the inner dynein arms and that its absence is deleterious for the organization and biogenesis of the axoneme. Overall our data confirm that despite close structural similarities, sperm flagella and cilia present important divergences in axonemal organization and biogenesis.

Supplemental Data

Supplemental Data include three figures and six tables and can be found with this article online at <http://www.cell.com/AJHG/>.

Acknowledgments

This work was supported by the research grant ICG2I funded by the program GENOPAT 2009 from the French Research Agency (ANR). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are grateful to Frédéric Plewniak, IGBMC, Strasbourg, who developed the homoSNP software used in this study (the software is available on request to plewniak@igbmc.u-strasbg.fr).

Received: July 19, 2013

Accepted: November 18, 2013

Published: December 19, 2013

Web Resources

The URLs for data presented herein are as follows:

1000 Genomes, <http://browser.1000genomes.org>

Berkeley Drosophila Genome Project NNSplice 0.9, http://www.fruitfly.org/seq_tools/splice.html

dbSNP v.137, http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi

GeneHub-GEPIS, <http://research-public.gene.com/Research/genentech/genehub-gepis/>

NHLBI Exome Sequencing Project (ESP) Exome Variant Server, <http://evs.gs.washington.edu/EVS/>
 Online Mendelian Inheritance in Man (OMIM), <http://www.omim.org/>
 PolyPhen-2, <http://www.genetics.bwh.harvard.edu/pph2/>
 RefSeq, <http://www.ncbi.nlm.nih.gov/RefSeq>
 SIFT, http://sift.jcvi.org/www/SIFT_chr_coords_submit.html
 UniGene, <http://www.ncbi.nlm.nih.gov/unigene>
 UCSC Genome Browser, <http://genome.ucsc.edu>

References

- Boivin, J., Bunting, L., Collins, J.A., and Nygren, K.G. (2007). International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care. *Hum. Reprod.* **22**, 1506–1512.
- Dieterich, K., Soto Rifo, R., Faure, A.K., Hennebicq, S., Ben Amar, B., Zahi, M., Perrin, J., Martinez, D., Sèle, B., Jouk, P.S., et al. (2007). Homozygous mutation of AURKC yields large-headed polyploid spermatozoa and causes male infertility. *Nat. Genet.* **39**, 661–665.
- Harbuz, R., Zouari, R., Pierre, V., Ben Khelifa, M., Kharouf, M., Coutton, C., Merdassi, G., Abada, F., Escoffier, J., Nikas, Y., et al. (2011). A recurrent deletion of DPY19L2 causes infertility in man by blocking sperm head elongation and acrosome formation. *Am. J. Hum. Genet.* **88**, 351–361.
- Schwabe, G.C., Hoffmann, K., Loges, N.T., Birker, D., Rossier, C., de Santi, M.M., Olbrich, H., Fliegauf, M., Failly, M., Liebers, U., et al. (2008). Primary ciliary dyskinesia associated with normal axoneme ultrastructure is caused by DNAH11 mutations. *Hum. Mutat.* **29**, 289–298.
- Rawe, V.Y., Galaverna, G.D., Acosta, A.A., Olmedo, S.B., and Chemes, H.E. (2001). Incidence of tail structure distortions associated with dysplasia of the fibrous sheath in human spermatozoa. *Hum. Reprod.* **16**, 879–886.
- Olmedo, S.B., Rawe, V.Y., Nodar, F.N., Galaverna, G.D., Acosta, A.A., and Chemes, H.E. (2000). Pregnancies established through intracytoplasmic sperm injection (ICSI) using spermatozoa with dysplasia of fibrous sheath. *Asian J. Androl.* **2**, 125–130.
- Chemes, H.E., Brugo, S., Zanchetti, F., Carrere, C., and Lavieri, J.C. (1987). Dysplasia of the fibrous sheath: an ultrastructural defect of human spermatozoa associated with sperm immotility and primary sterility. *Fertil. Steril.* **48**, 664–669.
- Neugebauer, D.C., Neuwinger, J., Jockenhövel, F., and Nieschlag, E. (1990). '9 + 0' axoneme in spermatozoa and some nasal cilia of a patient with totally immotile spermatozoa associated with thickened sheath and short midpiece. *Hum. Reprod.* **5**, 981–986.
- Escalier, D., and Albert, M. (2006). New fibrous sheath anomaly in spermatozoa of men with consanguinity. *Fertil. Steril.* **86**, e1–e9.
- Escalier, D. (2006). Arrest of flagellum morphogenesis with fibrous sheath immaturity of human spermatozoa. *Andrologia* **38**, 54–60.
- Escalier, D., Gallo, J.M., and Schrével, J. (1997). Immunohistochemical characterization of a human sperm fibrous sheath protein, its developmental expression pattern, and morphogenetic relationships with actin. *J. Histochem. Cytochem.* **45**, 909–922.
- David, G., Feneux, D., Serres, C., Escalier, D., and Jouannet, P. (1993). [A new entity of sperm pathology: peri-axonemal flagellar dyskinesia]. *Bull. Acad. Natl. Med.* **177**, 263–271, discussion 272–275.
- Stalf, T., Sánchez, R., Köhn, F.M., Schalles, U., Kleinstein, J., Hinz, V., Tielsch, J., Khanaga, O., Turley, H., Gips, H., et al. (1995). Pregnancy and birth after intracytoplasmic sperm injection with spermatozoa from a patient with tail stump syndrome. *Hum. Reprod.* **10**, 2112–2114.
- Moretti, E., Geminiani, M., Terzuoli, G., Renieri, T., Pascarelli, N., and Collodel, G. (2011). Two cases of sperm immotility: a mosaic of flagellar alterations related to dysplasia of the fibrous sheath and abnormalities of head-neck attachment. *Fertil. Steril.* **95**, e19–e23.
- Marmor, D., and Grob-Menendez, F. (1991). Male infertility due to asthenozoospermia and flagellar anomaly: detection in routine semen analysis. *Int. J. Androl.* **14**, 108–116.
- World Health Organization (2010). WHO Laboratory Manual for the Examination and Processing of Human Semen, fifth ed.
- Demonchy, R., Blisnick, T., Deprez, C., Toutirais, G., Loussert, C., Marande, W., Grellier, P., Bastin, P., and Kohl, L. (2009). Kinesin 9 family members perform separate functions in the trypanosome flagellum. *J. Cell Biol.* **187**, 615–622.
- Maiti, A.K., Mattéi, M.G., Jorissen, M., Volz, A., Zeigler, A., and Bouvagnet, P. (2000). Identification, tissue specific expression, and chromosomal localisation of several human dynein heavy chain genes. *Eur. J. Hum. Genet.* **8**, 923–932.
- Neesen, J., Kirschner, R., Ochs, M., Schmiedl, A., Habermann, B., Mueller, C., Holstein, A.F., Nuesslein, T., Adham, I., and Engel, W. (2001). Disruption of an inner arm dynein heavy chain gene results in asthenozoospermia and reduced ciliary beat frequency. *Hum. Mol. Genet.* **10**, 1117–1128.
- Shao, X., Tarnasky, H.A., Lee, J.P., Oko, R., and van der Hoorn, F.A. (1999). Spag4, a novel sperm protein, binds outer dense-fiber protein Odf1 and localizes to microtubules of manchette and axoneme. *Dev. Biol.* **211**, 109–123.
- Habura, A., Tikhonenko, I., Chisholm, R.L., and Koonce, M.P. (1999). Interaction mapping of a dynein heavy chain. Identification of dimerization and intermediate-chain binding domains. *J. Biol. Chem.* **274**, 15447–15453.
- Moore, D.J., Onoufriadiis, A., Shoemark, A., Simpson, M.A., Zur Lage, P.I., de Castro, S.C., Bartoloni, L., Gallone, G., Petridi, S., Woppard, W.J., et al. (2013). Mutations in ZMYND10, a gene essential for proper axonemal assembly of inner and outer dynein arms in humans and flies, cause primary ciliary dyskinesia. *Am. J. Hum. Genet.* **93**, 346–356.
- Höök, P., and Vallee, R. (2012). Dynein dynamics. *Nat. Struct. Mol. Biol.* **19**, 467–469.
- Myster, S.H., Knott, J.A., Wysocki, K.M., O'Toole, E., and Porter, M.E. (1999). Domains in the 1alpha dynein heavy chain required for inner arm assembly and flagellar motility in *Chlamydomonas*. *J. Cell Biol.* **146**, 801–818.
- Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402–408.
- King, S.M. (2013). A solid-state control system for dynein-based ciliary/flagellar motility. *J. Cell Biol.* **201**, 173–175.
- Antony, D., Becker-Heck, A., Zariwala, M.A., Schmidts, M., Onoufriadiis, A., Forouhan, M., Wilson, R., Taylor-Cox, T., Dewar, A., Jackson, C., et al.; Uk10k (2013). Mutations in CCDC39 and CCDC40 are the major cause of primary ciliary dyskinesia with axonemal disorganization and absent inner dynein arms. *Hum. Mutat.* **34**, 462–472.

28. Becker-Heck, A., Zohn, I.E., Okabe, N., Pollock, A., Lenhart, K.B., Sullivan-Brown, J., McSheene, J., Loges, N.T., Olbrich, H., Häffner, K., et al. (2011). The coiled-coil domain containing protein CCDC40 is essential for motile cilia function and left-right axis formation. *Nat. Genet.* **43**, 79–84.
29. Merveille, A.C., Davis, E.E., Becker-Heck, A., Legendre, M., Amirav, I., Bataille, G., Belmont, J., Beydon, N., Billen, F., Clément, A., et al. (2011). CCDC39 is required for assembly of inner dynein arms and the dynein regulatory complex and for normal ciliary motility in humans and dogs. *Nat. Genet.* **43**, 72–78.
30. Onoufriadis, A., Paff, T., Antony, D., Shoemark, A., Micha, D., Kuyt, B., Schmidts, M., Petridi, S., Dankert-Roelse, J.E., Haarmann, E.G., et al.; UK10K (2013). Splice-site mutations in the axonemal outer dynein arm docking complex gene CCDC114 cause primary ciliary dyskinesia. *Am. J. Hum. Genet.* **92**, 88–98.
31. Knowles, M.R., Leigh, M.W., Ostrowski, L.E., Huang, L., Carson, J.L., Hazucha, M.J., Yin, W., Berg, J.S., Davis, S.D., Dell, S.D., et al.; Genetic Disorders of Mucociliary Clearance Consortium (2013). Exome sequencing identifies mutations in CCDC114 as a cause of primary ciliary dyskinesia. *Am. J. Hum. Genet.* **92**, 99–106.
32. Omran, H., Kobayashi, D., Olbrich, H., Tsukahara, T., Loges, N.T., Hagiwara, H., Zhang, Q., Leblond, G., O'Toole, E., Hara, C., et al. (2008). Ktu/PF13 is required for cytoplasmic pre-assembly of axonemal dyneins. *Nature* **456**, 611–616.
33. Fliegauf, M., Olbrich, H., Horvath, J., Wildhaber, J.H., Zariwala, M.A., Kennedy, M., Knowles, M.R., and Omran, H. (2005). Mislocalization of DNAH5 and DNAH9 in respiratory cells from patients with primary ciliary dyskinesia. *Am. J. Respir. Crit. Care Med.* **171**, 1343–1349.
34. Olbrich, H., Schmidts, M., Werner, C., Onoufriadis, A., Loges, N.T., Raidt, J., Banki, N.F., Shoemark, A., Burgoyne, T., Al Turki, S., et al.; UK10K Consortium (2012). Recessive HYDIN mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry. *Am. J. Hum. Genet.* **91**, 672–684.
35. Loges, N.T., Olbrich, H., Becker-Heck, A., Häffner, K., Heer, A., Reinhard, C., Schmidts, M., Kispert, A., Zariwala, M.A., Leigh, M.W., et al. (2009). Deletions and point mutations of LRRC50 cause primary ciliary dyskinesia due to dynein arm defects. *Am. J. Hum. Genet.* **85**, 883–889.
36. Duquesnoy, P., Escudier, E., Vincensini, L., Freshour, J., Bridoux, A.M., Coste, A., Deschildre, A., de Blic, J., Legendre, M., Montantin, G., et al. (2009). Loss-of-function mutations in the human ortholog of *Chlamydomonas reinhardtii* ODA7 disrupt dynein arm assembly and cause primary ciliary dyskinesia. *Am. J. Hum. Genet.* **85**, 890–896.
37. Mitchison, H.M., Schmidts, M., Loges, N.T., Freshour, J., Dritsoula, A., Hirst, R.A., O'Callaghan, C., Blau, H., Al Dabbagh, M., Olbrich, H., et al. (2012). Mutations in axonemal dynein assembly factor DNAAF3 cause primary ciliary dyskinesia. *Nat. Genet.* **44**, 381–389, S1–S2.
38. Kott, E., Duquesnoy, P., Copin, B., Legendre, M., Dastot-Le Moal, F., Montantin, G., Jeanson, L., Tamalet, A., Papon, J.F., Siffroi, J.P., et al. (2012). Loss-of-function mutations in LRRC6, a gene essential for proper axonemal assembly of inner and outer dynein arms, cause primary ciliary dyskinesia. *Am. J. Hum. Genet.* **91**, 958–964.
39. Freshour, J., Yokoyama, R., and Mitchell, D.R. (2007). *Chlamydomonas* flagellar outer row dynein assembly protein ODA7 interacts with both outer row and I1 inner row dyneins. *J. Biol. Chem.* **282**, 5404–5412.
40. Mazor, M., Alkrinawi, S., Chalifa-Caspi, V., Manor, E., Sheffield, V.C., Aviram, M., and Parvari, R. (2011). Primary ciliary dyskinesia caused by homozygous mutation in DNAL1, encoding dynein light chain 1. *Am. J. Hum. Genet.* **88**, 599–607.
41. Rashid, S., Breckle, R., Hupe, M., Geisler, S., Doerwald, N., and Neesen, J. (2006). The murine Dnali1 gene encodes a flagellar protein that interacts with the cytoplasmic dynein heavy chain 1. *Mol. Reprod. Dev.* **73**, 784–794.

CHAPITRE B

Tables des variants restant après application des filtres pour les cas familiaux

Table B.1 – Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P1 et P2 de la famille MMAF1

SYMBOL	Variant impact	
	HGVSc, HGVSp	Consequence
PLA2G4B	c.1710-6delA ; .	splice region

Table B.2 – Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P3 et P4 de la famille MMAF2

SYMBOL	Variant impact	
	HGVSc, HGVSp	Consequence
SEMA5B	c.658C>A ; p.Glu220Lys	missense
CCDC37	c.1234_1236delTGG ; p.Glu412del	inframe deletion
DAPK1	c.2431T>A ; p.Val811Met	missense
MTSS1L	. ; p.Arg609Trp	missense
HYDIN	c.14857>T ; p.Arg4953Trp	missense
TMEM231	. ; p.Ala71Val	missense
ZNF469	. ; p.Arg1864Lys	missense
TGIF2	c.496C>A ; p.Leu166Met	missense
MMP9	c.820G>A ; p.Glu274Lys	missense

Table B.3 – Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P5 et P6 de la famille MMAF3

SYMBOL	Variant impact	
	HGVSc, HGVSp	Consequence
MYH11	c.4625G>A ; p.Arg1542Gln	missense
DNAH1	. ; .	splice acceptor

Table B.4 – Liste des variants ayant passé l'ensemble des filtres pour les deux sœurs P8 et P9 de la famille MMAF4

SYMBOL	Variant impact	
	HGVSc, HGVSp	Consequence
WEE2	. ; p.Pro92Leu	missense
GBP2	c.412T>A ; p.Ala138Thr	missense
ZFYVE28	c.1729C>A ; p.Val577Met	missense
FCGR3A	c.133T>C ; p.Ala45Pro	missense

Table B.5 – Liste des variants ayant passé l'ensemble des filtres pour le patient P10 de la famille MMAF5

SYMBOL	Variant impact	
	HGVSc, HGVSp	Consequence
ALOX15	c.268T>C ; p.Asp90His	missense
NFATC4	. ; p.Glu577Gly	missense
MYH7	c.77G>T ; p.Ala26Val	missense
LDHAL6A	c.334A>T ; p.Arg112Ter	stop gained
GBF1	c.1588G>T ; p.Arg530Cys	missense
TRBV6-6	c.5A>T ; p.Ser2Ile	missense
MROH2A	. ; p.Ile180Thr	missense
FSIP2	. ; p.Ala86Val	missense
ZSWIM2	c.676G>A ; p.Leu226Met	missense
CPZ	c.1236A>T ; p.Lys412Asn	missense
CCDC73	c.851A>C ; p.Phe284Ser	missense
NXPE2	c.1595>T ; p.Thr532Met	missense
NKX2-1	c.843G>T ; p.Gln281His	missense
CASP8	c.74A>G ; p.Pro25Arg	missense
OR6S1	. ; p.Leu116Met	missense
PDZD7	. ; p.Ser324Asn	missense
COL27A1	c.5063G>A ; p.Arg1688Gln	missense
SPEF2	c.3240delG ; p.Phe1080LeufsTer2	frameshift
ACAP1	c.1597A>T ; p.Arg533Trp	missense
C6	c.2312C>A ; p.Gly771Asp	missense
SLIT1	. ; .	splice region
ARHGAP19-SLIT1	c.*1721-7A>T ; .	splice region
DNAH2	. ; p.Arg2538Gln	missense

CHAPITRE C

Table des variants retrouvés sur le
gène *PATL2*

Table C.1 – Table des variants retrouvés sur le gène *PATL2*

Patient	Geno	Variant impact	
		HGVSc, HGVSp	Consequence
Ghs103	Homo	c.478G>T ; p.Arg160Ter	stop gained
Ghs104	Homo	c.478G>T ; p.Arg160Ter	stop gained
Ghs114	Homo	c.478G>T ; p.Arg160Ter	stop gained
Ghs5	Homo	c.478G>T ; p.Arg160Ter	stop gained
Ghs98	Homo	c.478G>T ; p.Arg160Ter	stop gained

CHAPITRE D

Variants retrouvés au sein de notre cohorte de patients MMAF

Table D.1 – Variantes homozygotes retrouvées sur le gène *CFAP43*

Patient	Geno	HGVS_C, HGVS_P	Variant impact		Consequence	SIFT; PolyPhen	Variant frequency	ExAC; ESP; 1KG
Ghs162	Homo	c.1240_1241delGC; p.Val414LeufsTer46		frameshift	.; .	.; .	.; .	.; .
Ghs164	Homo	c.1240_1241delGC; p.Val414IleufsTer46		frameshift	.; .	.; .	.; .	.; .
Ghs25	Homo	c.2141+5T>A ; p.Lys714Val*11		splice region	.; .	.; .	.; .	.; .
Ghs17	Homo	c.2658G>A ; p.Trp886Ter		stop gained	.; .	.; .	.; .	.; .
Ghs41	Homo	c.2680C>T ; p.Arg894Ter		stop gained	.; .	.; .	.; .	.; .
Ghs126	Homo	c.3352C>T ; p.Arg1118Ter		stop gained	.; .	.; .	.; .	.; .
Ghs105	Homo	c.3541-2A>C ; p.Ser1181Lysfs*4		splice acceptor	.; .	.; .	.; .	.; .
Ghs160	Homo	c.3541-2A>C ; p.Ser1181Lysfs*4		splice acceptor	.; .	.; .	.; .	.; .
Ghs102	Homo	c.3882delA ; p.Glu1294AspfsTer47		frameshift	.; .	.; .	.; .	.; .
Ghs132	Hete	c.1040G>C ; p.Val347Ala		missense	tolerated ; possibly damaging		7.41e-05 ; 2e-04 ; .	.; .
Ghs132	Hete	c.1300_1301insT ; p.Leu435SerfsTer26		frameshift	.; .	.; .	.; .	.; .

Table D.2 – Variants homozygotes retrouvés sur le gène *CFAP44*

Patient	HGVSc, HGVSp	Variant impact	Variant frequency
		Consequence	ExAC ; ESP ; 1KG
Ghs155	c.1387G>T ; p.Glu463Ter	stop gained	. ; . ; .
Ghs177	c.1387G>T ; p.Glu463Ter	stop gained	. ; . ; .
Ghs34	c.1890+1G>A ; p.Pro631Ile*22	splice donor	. ; . ; .
Ghs168	c.2818_2819insG ; p.Glu940GlyfsTer19	frameshift	8.24e-06 ; . ; .
Ghs22	c.3175C>T ; p.Arg1059Ter	stop gained	. ; . ; .
Ghs181	c.4767delC ; p.Ile1589MetfsTer6	frameshift	. ; . ; .

References

1. E. Tosti and Y. Ménézo : “Gamete activation : basic knowledge and clinical applications.” *Human Reproduction Update*. vol. 22, no. 4, pp. 420–439, 2016.
2. G. Hartshorne, S. Montgomery, and L. Krentzeris : “A case of failed oocyte maturation in vivo and in vitro.” *Fertility and sterility*. vol. 71, no. 3, pp. 567–70, 1999.
3. 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.
4. 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.
5. A. Hourvitz, E. Maman, M. Brengauz, R. Machtiner, 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.
6. 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</i> and Human Oocyte Meiotic Arrest.” *New England Journal of Medicine*. vol. 374, no. 3, pp. 223–232, 2016.
7. O. Hertwig : “Beitr{ä}ge zur Kenntniss der Bildung, Befruchtung und Theilung des thierischen Eies.” *W. Engelmann*, 1875.
8. 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.
9. 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.
10. S. Miyazaki, H. Shirakawa, K. Nakada, and Y. Honda : “Essential role of the inositol 1,4,5-trisphosphate receptor/Ca²⁺ release channel in Ca²⁺ waves and Ca²⁺ oscillations at fertilization of mammalian eggs.” <http://www.sciencedirect.com/science/article/pii/S0012160683711681>, (1993).
11. K. Swann : “A cytosolic sperm factor stimulates repetitive calcium increases and mimics fertilization in hamster eggs.” *Development*. vol. 110, no. 4, 1990.
12. 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.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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).
22. 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.
23. 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.
24. Y. Clermont : “The cycle of the seminiferous epithelium in man.” *American Journal*

- of Anatomy.* vol. 112, no. 1, pp. 35–51, 1963.
25. Y. Clermont : “Renewal of spermatogonia in man.” *American Journal of Anatomy.* vol. 118, no. 2, pp. 509–524, 1966.
26. E. Goossens and H. Tournaye : “Adult stem cells in the human testis.” *Seminars in Reproductive Medicine.* vol. 31, no. 1, pp. 39–48, 2013.
27. 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.
28. 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.
29. 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.
30. L.H. Yves Clermont, Richard Oko : “Cell and molecular biology of the testis.” *Oxford University Press,* 1993.
31. 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.
32. G.M.H. Hamilton, D. W., Waites : “Cellular and Molecular Events in Spermiogenesis.” *Cambridge University Press,* 1990.
33. 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.
34. U. Schenck and W.B. Schill : “Cytology of the human seminiferous epithelium.” *Acta cytologica.* vol. 32, no. 5, pp. 689–96,
35. M.M. Adelman and E.M. Cahill : “Atlas of sperm morphology.” *ASCP Press,* 1989.
36. World Health Organization : “WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction.” *Cambridge University Press,* 1992.
37. a Ogura, J. Matsuda, and R. Yanagimachi : “Birth of normal young after electrofusion of mouse oocytes with round spermatids.” *Proceedings of the National Academy of Sciences of the United States of America.* vol. 91, no. 16, pp. 7460–7462, 1994.
38. 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.
39. 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.
40. 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.
41. B. Asimakopoulos : “Is There a Place for Round and Elongated Spermatids Injection in.” vol. 1, no. 1, pp. 1–6, 2003.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. 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.
48. 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.
49. R.E. Braun : “Packaging paternal chromosomes with protamine.” *Nature genetics*. vol. 28, no. 1, pp. 10–12, 2001.
50. K. Inaba : “Molecular Architecture of the Sperm Flagella : Molecules for Motility

- and Signaling.” *Zoological Science*. vol. 20, no. 9, pp. 1043–1056, 2003.
51. E.M. Eddy : “The scaffold role of the fibrous sheath.” *Society of Reproduction and Fertility supplement*. vol. 65, pp. 45–62, 2007.
52. 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.
53. 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.
54. J.G.(.G. Grudzinskas and J. Yovich : “Gametes : the spermatozoon.” *Cambridge University Press*, 1995.
55. M. Michael and K. Joel : “Zellformen in normalen und pathologischen Ejakulaten und ihre klinische Bedeutung.” *Schweiz. Med. Wsch.* 1937.
56. 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.
57. J. MacLeod : “The Significance of Deviations in Human Sperm Morphology.” Presented at the (1970).
58. 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.
59. 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].” *Urologia i nefrologiya*. no. 2, pp. 45–7, 1993.
60. 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.
61. 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.
62. 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.
63. 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.
64. H.S. Levin : “Testicular biopsy in the study of male infertility.” *Human Pathology*. vol. 10, no. 5, pp. 569–584, 1979.
65. 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.
66. 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.
67. 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.
68. 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.
69. C. Lindholmer : “The importance of seminal plasma for human sperm motility.” *Biology of reproduction*. vol. 10, no. 5, pp. 533–42, 1974.
70. L. Björndahl : “The usefulness and significance of assessing rapidly progressive spermatozoa.” *Asian journal of andrology*. vol. 12, no. 1, pp. 33–5, 2010.
71. 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.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. A. Bojesen and C.H. Gravholt : "Morbidity and mortality in Klinefelter syndrome (47,XXY)." *Acta Paediatrica*. vol. 100, no. 6, pp. 807–813, 2011.
78. 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 Cyto-geneticiens 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.
79. D.J. Elliott and H.J. Cooke : "The molecular genetics of male infertility." *BioEssays*. vol. 19, no. 9, pp. 801–809, 1997.
80. C. Krausz and G. Forti : "Clinical aspects of male infertility." *Results and problems in cell differentiation*. vol. 28, pp. 1–21, 2000.
81. 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.
82. 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.
83. G. Minase, T. Miyamoto, Y. Miyagawa, M. Iijima, H. Ueda, Y. Saijo, M. Namiki, and K. Sengoku : "Single-nucleotide polymorphisms in the human *RAD21L* 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.
84. 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. Tütelmann : "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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. 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.
99. 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.
100. 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.
101. 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.
102. 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.
103. 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.
104. 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.
105. A. Amiri-Yekta, C. Coutton, Z.-E. Kherraf, T. Karaouzène, P. Le Tanno, M.H. Sannati, M. Sabbaghian, N. Almadani, M.A. Sadighi Gilani, S.H. Hosseini, S. Bahrami, A. Daneshipour, 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. T. Miyamoto, S. Hasuike, L. Yoge, 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.
113. A.N. Yatsenko, A. Roy, R. Chen, L. Ma, L.J. Murthy, W. Yan, D.J. Lamb, and M.M. Matzuk : “Non-invasive genetic diagnosis of male infertility using spermatozoal RNA : KLHL10mutations in oligozoospermic patients impair homodimerization.” *Human Molecular Genetics*. vol. 15, no. 23, pp. 3411–3419, 2006.
114. 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.
115. 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.
116. 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.
117. 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.
118. 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.
119. 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.
120. 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.
121. 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.
122. 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.
123. R. Bumgarner : “Overview of DNA microarrays : types, applications, and their future.” *Current protocols in molecular biology.* vol. Chapter 22, pp. Unit 22.1., 2013.
124. P.O. Brown, J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, 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.
125. 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.
126. M.L. Metzker : “Sequencing technologies - the next generation.” *Nature reviews. Genetics.* vol. 11, no. 1, pp. 31–46, 2010.
127. 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.
128. 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.
129. 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 Massively Parallel Sequencing of twelve human exomes.” *Nature.* vol. 461, no. 7261, pp. 272–276, 2010.
130. 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.
131. 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.
132. 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.
133. 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.
134. 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.
135. B. Wold and R.M. Myers : “Sequence census methods for functional genomics.” *Nature Methods.* vol. 5, no. 1, pp. 19–21, 2007.
136. 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.
137. 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.
138. 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.
139. 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.
140. 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.
141. 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.
142. 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.
143. 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.
144. 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.
145. D.S. Horner, G. Pavese, 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.
146. E.R. Mardis : “The impact of next-generation sequencing technology on genetics.” *Trends in Genetics.* vol. 24, no. 3, pp. 133–141, 2008.
147. D.R. Bentley : “Whole-genome re-sequencing.” *Current Opinion in Genetics and Development.* vol. 16, no. 6, pp. 545–552, 2006.
148. 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.
149. 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.
150. 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.
151. 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.
152. 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.
153. B. Langmead and S.L. Salzberg : “Fast gapped-read alignment with Bowtie 2.” *Nature Methods.* vol. 9, no. 4, pp. 357–359, 2012.
154. 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.
155. 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.
156. H. Li and R. Durbin : “Fast and accurate short read alignment with Burrows-Wheeler transform.” *Bioinformatics.* vol. 25, no. 14, pp. 1754–1760, 2009.
157. 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. Devanarayan, 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. Samboorg, 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.
158. 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.
159. S. Thankaswamy-Kosalai, P. Sen, and I. Nookae : “Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics.” *Genomics*. 2017.
160. 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.
161. 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.
162. 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.
163. 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.
164. 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.
165. 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.
166. 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. Gredler : “Evaluation of variant identification methods for whole genome sequencing data in dairy cattle.” *BMC genomics*. vol. 15, no. 1, pp. 948, 2014.
167. 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.
168. 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.
169. 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.
170. 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.
171. 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.
172. W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R.S. Ritchie, A. Thormann, P. Flliceck, and F. Cunningham : "The Ensembl Variant Effect Predictor." *Genome biology*. vol. 17, no. 1, pp. 122, 2016.
173. 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.
174. 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.
175. 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.
176. 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.
177. D. Salgado, M.I. Bellgard, J.P. Desvignes, and C. B??roud : "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.
178. 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.
179. 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.
180. 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.
181. 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.
182. 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.
183. 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.
184. 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.
185. 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.
186. 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.
187. 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.
188. 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.
189. J.D. McPherson : “Next-generation gap.” *Nature Methods.* vol. 6, no. 11s, pp. S2–S5, 2009.
190. 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.
191. 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.
192. RStudio Team : “RStudio : Integrated Development Environment for R.” *RStudio, Inc.*, Boston, MA, 2015.
193. S.B. Ng, K.J. Buckingham, C. Lee, A.W. Bigham, H.K. Tabor, K.M. Dent, C.D. Huff, P.T. Shannon, E.W. Jabs, D.A. Nickerson, J. Shendure, and M.J. Bamshad : “Exome sequencing identifies the cause of a Mendelian disorder.”
194. K. Pelak, K.V. Shianna, D. Ge, J.M. Maia, M. Zhu, J.P. Smith, E.T. Cirulli, J. Fellay, S.P. Dickson, C.E. Gumbs, E.L. Heinzen, A.C. Need, E.K. Ruzzo, A.

- Singh, C.R. Campbell, L.K. Hong, K.A. Lornsen, A.M. McKenzie, N.L.M. Sobreira, J.E. Hoover-Fong, J.D. Milner, R. Ottman, B.F. Haynes, J.J. Goedert, and D.B. Goldstein : “The characterization of twenty sequenced human genomes.” *PLoS genetics*. vol. 6, no. 9, pp. e1001111, 2010.
195. P.N. Robinson, S. Köhler, A. Oellrich, S.M.G. Sanger Mouse Genetics Project, K. Wang, C.J. Mungall, S.E. Lewis, N. Washington, S. Bauer, D. Seelow, P. Krawitz, C. Gilissen, M. Haendel, and D. Smedley : “Improved exome prioritization of disease genes through cross-species phenotype comparison.” *Genome research*. vol. 24, no. 2, pp. 340–8, 2014.
196. R Core Team : “R : A Language and Environment for Statistical Computing.” *R Foundation for Statistical Computing*, Vienna, Austria, 2017.
197. H. Wickham : “ggplot2 : Elegant Graphics for Data Analysis.” *Springer-Verlag New York*, 2009.
198. J.T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E.S. Lander, G. Getz, J.P. Mesirov, L. Pachter, J. Aach, E. Leproust, K. Eggan, G. Church, H. Li, Y. Lu, X. Fang, H. Liang, Z. Du, D. Li, Y. Zhao, Y. Hu, Z. Yang, H. Zheng, I. Hellmann, M. Inouye, J. Pool, X. Yi, J. Zhao, J. Duan, Y. Zhou, and J. Qin : “Integrative genomics viewer.” *Nature Biotechnology*. vol. 29, no. 1, pp. 24–26, 2011.
199. B.L. Aken, P. Achuthan, W. Akanni, M.R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C.G. Girón, L. Gordon, T. Hourlier, S.E. Hunt, S.H. Janacek, T. Juettemann, S. Keenan, M.R. Laird, I. Lavidas, T. Maurel, W. McLaren, B. Moore, D.N. Murphy, R. Nag, V. Newman, M. Nuhn, C.K. Ong, A. Parker, M. Patricio, H.S. Riat, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, S.P. Wilder, A. Zadissa, M. Kostadima, F.J. Martin, M. Muffato, E. Perry, M. Ruffier, D.M. Staines, S.J. Trevanion, F. Cunningham, A. Yates, D.R. Zerbino, and P. Flicek : “Ensembl 2017.” *Nucleic acids research*. vol. 45, no. D1, pp. D635–D642, 2017.
200. Y.-F. Chang, J.S. Imam, and M.F. Wilkinson : “The Nonsense-Mediated Decay RNA Surveillance Pathway.” *Annual Review of Biochemistry*. vol. 76, no. 1, pp. 51–74, 2007.
201. K.E. Baker and R. Parker : “Nonsense-mediated mRNA decay : terminating erroneous gene expression.” *Current opinion in cell biology*. vol. 16, no. 3, pp. 293–9, 2004.
202. B. Lee, I. Park, S. Jin, H. Choi, J.T. Kwon, J. Kim, J. Jeong, B.-N. Cho, E.M. Eddy, and C. Cho : “Impaired spermatogenesis and fertility in mice carrying a mutation in the Spink2 gene expressed predominantly in testes.” *The Journal of biological chemistry*. vol. 286, no. 33, pp. 29108–17, 2011.
203. M. Aarabi, H. Balakier, S. Bashar, S.I. Moskovtsev, P. Sutovsky, C.L. Librach, and R. Oko : “Sperm content of postacrosomal WW binding protein is related to fertilization outcomes in patients undergoing assisted reproductive technology.”

- Fertility and Sterility.* vol. 102, no. 2, pp. 440–447, 2014.
204. M. Aarabi, H. Balakier, S. Bashar, S.I. Moskvtsev, P. Sutovsky, C.L. Librach, and R. Oko : “Sperm-derived WW domain-binding protein, PAWP, elicits calcium oscillations and oocyte activation in humans and mice.” *FASEB journal : official publication of the Federation of American Societies for Experimental Biology.* vol. 28, no. 10, pp. 4434–40, 2014.
205. M.S. Lehti, F.-P. Zhang, N. Kotaja, and A. Sironen : “SPEF2 functions in microtubule-mediated transport in elongating spermatids to ensure proper male germ cell differentiation.” *Development.* vol. 144, no. 14, 2017.
206. A. Marnef, M. Maldonado, A. Bugaut, S. Balasubramanian, M. Kress, D. Weil, and N. Standart : “Distinct functions of maternal and somatic Pat1 protein paralogs.” *RNA.* vol. 16, no. 11, pp. 2094–2107, 2010.
207. Y. Nakamura, K.J. Tanaka, M. Miyauchi, L. Huang, M. Tsujimoto, and K. Matsumoto : “Translational repression by the oocyte-specific protein P100 in *Xenopus*.” *Developmental biology.* vol. 344, no. 1, pp. 272–83, 2010.
208. A.E. Ivliev, P.A.C. ’t Hoen, W.M.C. van Roon-Mom, D.J.M. Peters, and M.G. Sergeeva : “Exploring the Transcriptome of Ciliated Cells Using In Silico Dissection of Human Tissues.” *PLoS ONE.* vol. 7, no. 4, pp. e35618, 2012.
209. T.F. Smith : “Diversity of WD-Repeat proteins.” The coronin family of proteins. pp. 20–30. *Springer New York*, New York, NY (2008).
210. R. Broadhead, H.R. Dawe, H. Farr, S. Griffiths, S.R. Hart, N. Portman, M.K. Shaw, M.L. Ginger, S.J. Gaskell, P.G. McKean, and K. Gull : “Flagellar motility is required for the viability of the bloodstream trypanosome.” *Nature.* vol. 440, no. 7081, pp. 224–227, 2006.
211. I. Subota, D. Julkowska, L. Vincensini, N. Reeg, J. Buisson, T. Blisnick, D. Huet, S. Perrot, J. Santi-Rocca, M. Duchateau, V. Hourdel, J.-C. Rousselle, N. Cayet, A. Namane, J. Chamot-Rooke, and P. Bastin : “Proteomic Analysis of Intact Flagella of Procytic <i>Trypanosoma brucei</i> Cells Identifies Novel Flagellar Proteins with Unique Sub-localization and Dynamics.” *Molecular & Cellular Proteomics.* vol. 13, no. 7, pp. 1769–1786, 2014.
212. G. Singh : “Ultrastructural features of round-headed human spermatozoa.” *International journal of fertility.* vol. 37, no. 2, pp. 99–102,
213. H. Pedersen and H. Rebbe : “Fine structure of round-headed human spermatozoa.” *Journal of reproduction and fertility.* vol. 37, no. 1, pp. 51–4, 1974.
214. S. Taylor, S. Yoon, M. Morshedi, D. Lacey, T. Jellerette, R. Fissore, and S. Oehninger : “Complete globozoospermia associated with PLC ζ deficiency treated with calcium ionophore and ICSI results in pregnancy.” *Reproductive BioMedicine*

- Online.* vol. 20, no. 4, pp. 559–564, 2010.
215. S.-Y. Yoon, T. Jellerette, A.M. Salicioni, H.C. Lee, M.-S. Yoo, K. Coward, J. Parrington, D. Grow, J.B. Cibelli, P.E. Visconti, J. Mager, and R.A. Fissore : “Human sperm devoid of PLC ζ fail to induce Ca(2+) release and are unable to initiate the first step of embryo development.” *The Journal of clinical investigation.* vol. 118, no. 11, pp. 3671–81, 2008.
216. A.H.D.M. Dam, I. Koscinski, J.A.M. 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.” *American journal of human genetics.* vol. 81, no. 4, pp. 813–20, 2007.
217. P.F. Ray and C. Arnoult : “La délétion homozygote du gène <i>DPY19L2</i> est responsable de la majorité des cas de globozoospermie.” *médecine/sciences.* vol. 27, no. 8-9, pp. 692–693, 2011.
218. V. Pierre, G. Martinez, C. Coutton, J. Delaroche, S. Yassine, C. Novella, K. Pernet-Gallay, S. Hennebicq, P.F. Ray, and C. Arnoult : “Absence of Dpy19l2, a new inner nuclear membrane protein, causes globozoospermia in mice by preventing the anchoring of the acrosome to the nucleus.” *Development.* vol. 139, no. 16, pp. 2955–2965, 2012.
219. A.R. Carson, J. Cheung, and S.W. Scherer : “Duplication and relocation of the functional DPY19L2 gene within low copy repeats.” *BMC genomics.* vol. 7, pp. 45, 2006.
220. J. Cheung, X. Estivill, R. Khaja, J.R. MacDonald, K. Lau, L.-C. Tsui, and S.W. Scherer : “Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence.” *Genome biology.* vol. 4, no. 4, pp. R25, 2003.
221. J.A. Bailey, Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler : “Recent Segmental Duplications in the Human Genome.” *Science.* vol. 297, no. 5583, pp. 1003–1007, 2002.
222. B. Walsh : “Population-genetic models of the fates of duplicate genes.” *Genetica.* vol. 118, no. 2-3, pp. 279–94, 2003.
223. S. Ohno : “Evolution by Gene Duplication.” *Springer Berlin Heidelberg,* Berlin, Heidelberg, 1970.
224. 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.
225. P. Liu, C.M. Carvalho, P. Hastings, and J.R. Lupski : “Mechanisms for recurrent and complex human genomic rearrangements.” *Current Opinion in Genetics & Development.* vol. 22, no. 3, pp. 211–220, 2012.
226. J.R. MacDonald, R. Ziman, R.K.C. Yuen, L. Feuk, and S.W. Scherer : “The Database of Genomic Variants : a curated collection of structural variation in the human genome.” *Nucleic acids research.* vol. 42, no. Database issue, pp. D986–92, 2014.
227. E.D. Parvanov, P.M. Petkov, and K. Paigen : “Prdm9 controls activation of mammalian recombination hotspots.” *Science (New York, N.Y.).* vol. 327, no. 5967, pp. 835, 2010.
228. F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy : “PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice.” *Science (New York, N.Y.).* vol. 327, no. 5967, pp. 836–40, 2010.
229. X. Ding, R. Xu, J. Yu, T. Xu, Y. Zhuang, and M. Han : “SUN1 Is Required for Telomere Attachment to Nuclear Envelope and Gametogenesis in Mice.” *Developmental Cell.* vol. 12, no. 6, pp. 863–872, 2007.
230. V. Plagnol, J. Curtis, M. Epstein, K.Y. Mok, E. Stebbings, S. Grigoriadou, N.W. Wood, S. Hambleton, S.O. Burns, A.J. Thrasher, D. Kumararatne, R. Doffinger, and S. Nejentsev : “A robust model for read count data in exome sequencing experiments and implications for copy number variant calling.” *Bioinformatics (Oxford, England).* vol. 28, no. 21, pp. 2747–54, 2012.
231. N. Krumm, P.H. Sudmant, A. Ko, B.J. O’Roak, M. Malig, B.P. Coe, N.E.S. NHLBI Exome Sequencing Project, A.R. Quinlan, D.A. Nickerson, and E.E. Eichler : “Copy number variation detection and genotyping from exome sequence data.” *Genome research.* vol. 22, no. 8, pp. 1525–32, 2012.
232. J.F. Sathirapongsasuti, H. Lee, B.A.J. Horst, G. Brunner, A.J. Cochran, S. Binder, J. Quackenbush, and S.F. Nelson : “Exome sequencing-based copy-number variation and loss of heterozygosity detection : ExomeCNV.” *Bioinformatics (Oxford, England).* vol. 27, no. 19, pp. 2648–54, 2011.
233. M. Zhao, Q. Wang, Q. Wang, P. Jia, and Z. Zhao : “Computational tools for copy number variation (CNV) detection using next-generation sequencing data : features and perspectives.” *BMC Bioinformatics.* vol. 14, pp. S1, 2013.
234. Y. Guo, Q. Sheng, D.C. Samuels, B. Lehmann, J.A. Bauer, J. Pietenpol, and Y. Shyr : “Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control.” *BioMed research international.*

- vol. 2013, pp. 915636, 2013.
235. S. Comazzetto, M. Di Giacomo, K.D. Rasmussen, C. Much, C. Azzi, E. Perlas, M. Morgan, and D. O'Carroll : “Oligoasthenoteratozoospermia and Infertility in Mice Deficient for miR-34b/c and miR-449 Loci.” *PLoS Genetics*. vol. 10, no. 10, pp. e1004597, 2014.
236. X. Chen, X. Li, J. Guo, P. Zhang, and W. Zeng : “The roles of microRNAs in regulation of mammalian spermatogenesis.” *Journal of animal science and biotechnology*. vol. 8, pp. 35, 2017.
237. D.T. Carrell and K.I. Aston : “The search for SNPs, CNVs, and epigenetic variants associated with the complex disease of male infertility.” *Systems Biology in Reproductive Medicine*. vol. 57, no. 1-2, pp. 17–26, 2011.
238. R. Dada, M. Shamsi, and K. Kumar : “Genetic and epigenetic factors : Role in male infertility.” *Indian Journal of Urology*. vol. 27, no. 1, pp. 110, 2011.
239. R. Dada, M. Kumar, R. Jesudasan, J.L. Fernández, J. Gosálvez, and A. Agarwal : “Epigenetics and its role in male infertility.” *Journal of Assisted Reproduction and Genetics*. vol. 29, no. 3, pp. 213–223, 2012.