HAL
open science

Découverte de nouvelles protéines impliquées dans la spermatogenèse chez le rat

Sophie Chocu

To cite this version:

Sophie Chocu. Découverte de nouvelles protéines impliquées dans la spermatogenèse chez le rat. Biologie de la reproduction. Université de Rennes, 2014. Français. NNT : 2014REN1S064 . tel-01127389

HAL Id: tel-01127389

https://theses.hal.science/tel-01127389v1

Submitted on 7 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ANNÉE 2014

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie

Ecole doctorale (Vie Agro Santé)

Présentée par

Sophie CHOCU

préparée à IRSET- INSERM U1085

(UFR Sciences de la Vie et de l'Environnement)

Découverte de nouvelles protéines impliquées dans la spermatogenèse chez le rat

Thèse soutenue à Rennes
Le 30 septembre 2014
devant le jury composé de:

Thierry RABILLOUD
Directeur de recherche, iRTSV/CEA-Grenoble
Rapporteur

Pascal SOURDAINE
Professeur à l'Université de Caen
Rapporteur

Jean-Jacques LAREYRE
Directeur de recherche, LPGP/INRA-Rennes
Examinateur

Jean ARMENGAUD
Directeur de recherche, CEA-Marcoule
Examinateur

Denis MICHEL
Professeur à l'Université de Rennes 1
Président

Charles PINEAU
Directeur de recherche, INSERM
Directeur de thèse

"Nothing truly valuable can be achieved except by the unselfish cooperation of many individuals."

Albert Einstein (1879-1955).

REMERCIEMENTS

Tout d'abord, je tiens à remercier les membres du jury, les Docteurs Thierry Rabilloud, Pascal Sourdaine, Jean Jacques Lareyre, Jean Armengaud et Denis Michel d'avoir accepté d'évaluer ce travail et de s'être rendus si disponibles.

Je tiens à remercier mon directeur de thèse, le Docteur Charles Pineau pour m'avoir offert la possibilité de réaliser cette thèse dans d'excellentes conditions matérielles au sein de la plateforme Protéomique Biogenouest, ainsi que pour toutes nos discussions au sujet de la spermatogenèse. Je le remercie de m'avoir transmis son savoir-faire pour l'isolement de cellules germinales et de cellules de Sertoli de rat. Je le remercie également pour m'avoir permis d'assister à plusieurs formations et congrès internationaux qui m'ont permis de valoriser notre travail.

Je tiens à remercier Bernard Jégou, directeur de l'unité IRSET/INSERM U1085 au sein de laquelle a pu été réalisé ce travail de thèse, et à lui témoigner mon respect et ma reconnaissance. Si nous n'avons pas eu le temps de beaucoup échanger sur le plan scientifique, nos échanges humains positifs m'ont été très bénéfiques.

Merci à toute l'équipe de la plateforme : Blandine Charoy, Emmanuelle Com, Mélanie Lagarrique, Laëtitia Guilllot-Cloarec, et Régis Lavigne pour leur aide sur toute la partie Protéomique. Je remercie également Karine Rondel pour sa gentillesse et ses coups de mains, Sophie Guinard pour sa bonne humeur…et ses coups de mains. Et bien sûr, je remercie le Docteur Nathalie Melaine pour toutes nos discussions, et pour son énergie ! Merci à Loren Méar, que j'ai participé à encadrer pour son stage de Master 1, ainsi qu'à Stéhane Dinahet, stagiaire de Master 2.

Un immense merci à Frédéric Chalmel, Antoine Rolland, Bertrand Evrard, Florence Aubry et Aurélie Lardenois pour avoir contribué à donner du sens à cette histoire. Antoine, pour sa pédagogie, son calme, sa patience, sa bonne volonté, merci de m'avoir encadrée sur

l'utilisation d'outils d'analyses de données, et d'avoir préparé pour la première fois les spermatogonies de rat avec moi. Merci pour les tuyaux en biologie moléculaire, et d'avoir continué mon travail sur le clonage de gènes candidats et la production de protéines recombinantes. Merci à Florence Aubry pour m'avoir encadrée sur la partie production de protéines recombinantes. Merci à Aurélie Lardenois pour son aide précieuse sur les analyses statistiques des données d'ICPL, sa douceur et sa rigueur (c'est possible), ses conseils et son soutien pendant ma période de rédaction de ce manuscrit, merci pour ces moments sympas passés ensemble. Un grand merci à Bertrand pour son travail sur les validations biochimiques de TUTs. Merci à vous pour votre attitude directe et positive.

Je remercie tout particulièrement Frédéric Chalmel pour la supervision d'une partie importante de cette thèse, travail qui s'est concrétisé par l'article principal que je présente. Frédéric n'a pas hésité à me donner un sérieux coup de pouce à un moment de doute durant cette thèse. Pour ses conseils et pour le partage de ses idées, et le travail qu'il a réalisé, celui que nous avons réalisé ensemble, je suis vraiment reconnaissante. Il a aussi présenté notre travail au 18ème « European Testis Workshop » au Danemark en mai 2014. Frédéric Chalmel est un chercheur qui donne généreusement de son temps et de ses compétences sans compter, et qui sera je l'espère, récompensé à la hauteur de ses efforts incessants pour la science. J'apprécie particulièrement sa vision des choses quant à la mutualisation des données de transcriptomique et de protéomique pour les scientifiques de la reproduction. Je pense aussi que c'est par la coopération que l'on peut faire avancer la connaissance.

Je remerci d'ailleurs Ollivier Sallou de L'IRISA et Laëtitia Guillot pour avoir fait figurer visuellement en ligne sur RGV les données de protéomique que j'ai générées.

Merci au Dr.Amos Bairoch d'avoir pris le temps de regarder en détail mon « top candidat » qui correspond à la « putative » YP032 chez l'humain.

Merci aussi à Christine Kervarrec; Isabelle Coiffec, Christelle Desdoit pour leur conseils et leur aide technique. Merci à Fatima Smagulova pour toute sa positivité et ses encouragements, et à Igor Stuparevic pour ses conseils.

Un grand merci à Clémentine Chalmey pour nos échanges de rongeurs, nos bavardages et surtout, son soutien, particulièrement à la fin de cette thèse. Merci aussi aux autres doctorantes de l'Unité U1085. Millissia Ben Maamar (*Keep CALM and don't slice too much…*), Lauriane Sèdes, Céline Camus. Merci à toutes !

Merci à Fanny Jumeau, avec qui ce fut un plaisir de collaborer même brièvement, merci pour ce soutien et nos échanges fort sympathiques.

Merci à Laure Tonini pour nos discussions protéomiques et culinaires, notre randonnée Tyrolienne fut fort appréciable.

Merci aux filles de la promo du master SCMV : Claire Schirmer, Géraldine David, Sarah Tessier, Christine Saffray, Maëna le Corvec, Nadia Saïdi pour toutes nos soirées. Courage dans votre nouvelle vie de docteur ou pharmacien, et de maman pour certaines.

Merci aux yogis et yoginis rencontrés sur ma route.

De si loin, merci à Amandine Etchessahar et Solenn Lamprière, mes amies pour la vie.

Un grand merci à ma Maman qui m'a fait découvrir très très tôt les sciences du vivant et la géologie. Elle a su s'occuper sans compter de ma petite Cléo pendant mes années d'études. Je connais peu de personnes aussi généreuses et fiables.

Enfin merci à mon très cher Tristan Auvray, talentueux violoniste et tabliste, et qui a su tout adoucir et supporter avec humour, calme et intelligence. Merci à nos petites filles, surtout la gentille et sage Cléo pour m'avoir si souvent ramenée à l'essentiel, *(«…Et bien, vous êtes les dernières parce que vous êtes les plus importantes ! »).* Peut être appendrez-vous, qu'une des choses les plus stables dans la vie est finalement le changement, la mutation, et que les réponses se trouvent souvent à l'intérieur de nous mêmes.

CONTEXTE GÉNÉRAL

Ce travail a été réalisé au sein de l'IRSET (Institut de Recherche sur la Santé, l'Environnement et le Travail) - INSERM U1085, sur la plateforme protéomique Biogenouest dirigée par Charles Pineau. La plateforme axe ses activités de recherche principalement autour de la biologie de la reproduction. Aujourd'hui, ces recherches se placent dans un contexte environnemental dans lequel la santé reproductive mâle est de plus en plus fragilisée. Un certain nombre des pathologies de la reproduction concernent les cellules testiculaires, germinales ou somatiques et aboutissent à une altération de la production de spermatozoïdes. Pour comprendre les mécanismes qui sous-tendent la spermatogenèse normale et pathologique, l'unité travaille sur des modèles mammifères : rat, souris ainsi que chez l'homme. Depuis plus de trente ans, les travaux entrepris dans ce domaine au sein de l'équipe de Charles Pineau se sont focalisés sur le devenir des cellules germinales et le dialogue qu'elles établissent avec les cellules de Sertoli dans les tubules séminifères.

Pour étudier les interactions entre les cellules germinales et les cellules de Sertoli, deux axes de travail ont été utilisés : *in vivo* et *in vitro*, par un certain nombre de travaux qui ont permis de mettre en évidence le concept de germines. Les études *in vivo* (Pinon-Lataillade et al., 1988; Pineau et al., 1989; Pinon-Lataillade et al., 1991; Jégou et al., 1993), ont consisté à irradier des rats pendant différentes périodes pour éliminer les cellules germinales en division (les spermatogonies et les spermatocytes préleptotène) et obtenir une déplétion progressive et séquentielle du nombre de spermatocytes pachytènes, spermatides rondes et spermatides âgées dans l'épithélium séminifère, alors que le nombre de cellules de Sertoli restait inchangé. Les niveaux de la protéine de liaison des androgènes, ABP (Androgen binding protein) produite par les cellules de Sertoli, et les niveaux de gonadotrophines ont été évalués après différentes périodes d'irradiation. Au cours de ces études, les auteurs ont conclu que certains aspects de la fonction sertolienne sont contrôlés par les spermatides âgées, et en ont déduit une corrélation entre la sécrétion de certains facteurs sertoliens tels que l'ABP, et le nombre de spermatides âgées. Une étude *in vitro* a montré par exemple que des cellules germinales contrôlent leur apport en fer en stimulant le sécrétion de transferrine par les cellules de Sertoli (Le Magueresse et al., 1988). Les chercheurs de cette équipe ont montré que les produits de sécrétion des cellules de Sertoli répondent différemment aux facteurs protéiques originaires des cellules germinales (Onoda and Djakiew, 1990; Jégou, 1991); et en particulier, que ces facteurs peuvent affecter négativement la production de testines et de clusterine par les cellules de Sertoli. Leur action

peut donc avoir des effets opposés sur les différents produits sertoliens (Jégou et al., 1993) comme c'est la cas pour la transferrine et la clusterine (Pineau et al., 1993). Selon leur stade de différenciation, les cellules germinales à l'origine de ces facteurs ou « germines » influencent différemment la fonction Sertolienne (Gérard and Jégou, 1993). Les spermatides matures semblent avoir un rôle particulièrement important dans la régulation de la fonction de sécrétion des cellules de Sertoli (Onoda et al., 1991; Jégou et al., 1993; Onoda and Djakiew, 1993; Pineau et al., 1993). Les spermatocytes pachytène influencent également la sécrétion de plusieurs protéines sertoliennes : la céruloplasmine, les protéines de liaison aux lipides SGP 1 et 2 (Sulfated glycoprotein 1 et 2) et la transferrine (Onoda et al., 1991). Il existe donc des facteurs spécifiques de chaque stade de différenciation des cellules germinales qui ont des effets différents sur la fonction sertolienne. Les chercheurs ont tenté d'identifier et de purifier ces facteurs (Onoda and Djakiew, 1993; Pineau et al., 1993b), en particulier ceux spécifiques des spermatides, tel que la protéine PEBP-1 (Phosphatidylethanolamine binding protein), (Onoda and Djakiew, 1993). Un facteur issu des spermatides a été isolé, mais n'a pas pu être identifié, car sa séquence n'était pas dans les bases de données (Onoda and Djakiew, 1993).

Les chercheurs se sont donc heurtés aux difficultés inhérentes à la nature même des cellules germinales en différenciation qu'il est impossible de maintenir en culture afin de produire des milieux conditionnés pour tester leurs effets sur les cellules de Sertoli, ceci constituant une limite sérieuse à leur étude. Une deuxième limitation vient du fait que ces facteurs spécifiques peuvent être non caractérisés, et absents des bases de données.

De ces limitations est né le besoin d'utiliser des approches nouvelles pour l'étude de la spermatogenèse : les approches « Omiques ». La protéomique, étude de l'expression des protéines dans une cellule ou un organisme dans une condition donnée et à un instant donné est déjà beaucoup utilisée dans le domaine de la biologie reproductive. Ainsi, dans notre équipe, le protéome des cellules germinales a déjà été étudié. D'abord, dans les années 2000, un répertoire d'une cinquantaine de protéines par une approche d'électrophorèse bidimensionnelle (Guillaume et al., 2000), puis, une liste de plus de 153 protéines par la même approche a été obtenue (Com et al., 2003). Ainsi ont été découvertes les protéines MCM7 (Minichromosome maintenance protein 7), TCTP (Translationally-controlled tumor protein homolog) et la stathmine (Guillaume et al., 2001a, 2001b; Com et al., 2006). Les auteurs de ces travaux ont montré une expression différentielle de ces protéines dans la lignée germinale, et préférentielle dans les spermatogonies, ce qui laisse présupposer leur rôle dans

la spermatogenèse qui reste à définir. En revanche, ces études ne constituaient pas une étude différentielle à grande échelle.

Les études différentielles ont été initiées plus tard dans le but d'identifier des protéines qui sont spécifiques des cellules germinales et potentiellement impliquées dans les communications cellules de Sertoli/cellules germinales. La technique 2D-DIGE (Two-Dimensional Difference Gel Electrophoresis) apportée dans notre laboratoire dès 2005, a permis une telle étude différentielle (Rolland et al., 2007). Cette dernière étude a permis de faire un bond dans les identifications protéiques, avec la mise en évidence de 997 spots différentiels entre les différents types de cellules germinales, correspondant à 123 protéines non redondantes et significativement différentielles, c'est à dire avec un ratio d'expression ≥ 2,5 entre deux types cellulaires identifiées en spectrométrie de masse. L'expression de quelques-unes d'entre elles : PTBP2 (Neural polypyrimidine tract-binding protein), Smac/Diablo, Grp58 et GADPH (Glyceraldehyde-3-phosphate dehydrogenase I), a été validée par western blot et immunohistochimie. Parmi ces 123 protéines, la protéine inconnue nommée plus tard CLPH (Casein-like phosphoprotein), fortement exprimée dans les spermatides chez le rat et chez l'homme, a fait l'objet d'une étude ultérieure descriptive et fonctionnelle approfondie. Elle a été étudiée pour ses propriétés biochimiques, en particulier sa capacité à être phosphorylée par la caséine kinase 2, une enzyme indispensable à la morphogenèse des spermatozoïdes, et à lier le calcium (Calvel et al., 2009).

En parallèle, des chercheurs de l'IRSET se sont aussi intéressés à la régulation transcriptionnelle du programme d'expression des gènes dans la lignée germinale mâle chez les mammifères par des approches par puces à ADN (Wrobel and Primig, 2005; Chalmel et al., 2007a, 2007b, 2012). En effet, un grand nombre de gènes sont régulés temporellement au cours des différentes étapes de la spermatogenèse. Une étude transcriptomique importante de Frédéric Chalmel et collaborateurs (Chalmel et al., 2007a) a permis d'établir le transcriptome conservé de la gamétogenèse chez le rat, la souris et l'homme, avec une analyse des transcriptomes différentiels entre les différentes cellules de la lignée germinale mâle. Cette étude a suscité l'idée d'intégrer les données de la transcriptomique avec les données de la protéomique différentielle pour faciliter le choix de protéines, qui, de par leur profil d'expression dans les cellules germinales sont potentiellement impliquées dans le déroulement de la spermatogenèse. Depuis quelques années, les avancées des technologies de séquençage à haut débit comme le séquençage des ARNs ou RNA-seq (RNA sequencing) ont permis d'accéder à une vue globale du programme d'expression des gènes à un niveau de

détail sans précédent dans un processus biologique donné. En effet, dans leur étude récente, Frédéric Chalmel et ses collaborateurs ont rapporté le profil d'expression de nombreux transcrits (plus de 20.000) reconstruits et quantifiés dans les cellules testiculaires isolées du rat, en particulier aux différentes étapes de la différenciation des cellules germinales mâles (Chalmel et al., 2014). Un certain nombre d'entre eux, les TUTs (testicular unannotated transcripts), ne sont pas annotés, l'annotation du génome rn4 du rat n'étant pas complète. Ces derniers partagent les caractéristiques des longs ARNs non codants, les lncRNAs (long non coding RNAs) qui sont une sous classe d'ARNs non codants ncRNAs (non coding RNAs) récemment découverte (Mercer et al., 2009; Hung and Chang, 2010). Chalmel et collaborateurs rapportent aussi le fait que ces TUTs ainsi que des lncRNAs déjà connus s'accumulent pendant les phases méiotiques et post-méiotiques de la spermatogenèse. Les TUTs et les lncRNAs ont des caractéristiques génomiques proches semble-t-il. Or, seule une approche intégrative faisant appel à la protéomique permet de trancher entre des évènements transcrits inconnus mais qui codent pour des protéines, et des transcrits non codants exprimés de manière stade-spécifique au cours de la spermatogenèse. C'est dans ce but que nous avons décidé de nous focaliser sur la recherche de nouveaux évènements codants différentiellement exprimés au cours des stades méiotique et post-méiotique de la spermatogenèse, par une approche de protéomique qui intègre les données de la transcriptomique par RNA-seq. Ce travail sera présenté dans le premier chapitre de ce manuscrit. Un deuxième chapitre sera présenté en continuité de ce travail, et qui traitera de la découverte de nouvelles isoformes spécifiques des protéines germinales par la même approche.

Du côté de la protéomique, des approches différentielles de quantification relative à l'aide de marquage isotopiques, telles que l'ICPL (Isotope-coded protein label), plus puissantes que la DIGE, ont été adoptées par le laboratoire. Pour continuer le travail commencé par Rolland et collaborateurs qui commença d'établir le protéome différentiel de la spermatogenèse chez le rat, j'ai utilisé la technologie ICPL en me focalisant sur les protéines membranaires, pour accéder plus facilement aux protéines dont le profil d'expression laisserait supposer un rôle important dans le dialogue Sertoli / germinales. L'influence des spermatides matures sur les activités de sécrétion des cellules de Sertoli dont nous avons parlé plus haut pouvant être médiée par des changements de conformation de ces dernières induits par les spermatides au cours de la spermiogénèse et par la phagocytose des corps résiduels (Jégou, 1991) ; j'ai inclus les corps résiduels dans cette étude. Ce travail sera présenté dans le troisième chapitre.

Le devenir du corps résiduel et les mécanismes à l'origine de sa formation sont aussi un sujet d'étude historique dans l'unité (Pineau et al., 1991; Gérard et al., 1992; Syed et al., 1995). Il est connu que la phagocytose des corps résiduels déclenche une nouvelle vague spermatogénétique juste après la spermiation, mécanisme encore mal connu. Le devenir du corps résiduel au sein de la cellule de Sertoli est également mal connu. Deux hypothèses concomitantes sont actuellement formulées au sein de la communauté scientifique pour expliquer sa spécificité par rapport à la phagocytose des cellules apoptotiques par les cellules de Sertoli: celle d'une autophagie du corps résiduel couplée à sa phagocytose, et celle de la spécialisation des complexes protéiques membranaires apicaux (spécialisations ectoplasmiques) des cellules de Sertoli comparés aux autres complexes membranaires impliqués dans la phagocytose des cellules germinales apoptotiques. Afin d'identifier des protéines susceptibles d'avoir un rôle dans le devenir des corps résiduels, nous avons réalisé le protéome du corps résiduel et celui des cellules de Sertoli en utilisant une approche protéomique shogun itérative. Ce travail juste initié sera présenté dans le quatrième chapitre de ce manuscrit.

ABREVIATIONS

2D-DIGE: Two-Dimensional Difference Gel Electrophoresis

ABP: Androgen binding protein

AEBSF: 4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride

AMH: Anti mullerian hormone

AMPc: Adénosine monophospahte cyclique

AP-MS : Affinity purification mass spectrometry

AR: Androgen receptor

ARE-BPs: AU-rich elements binding proteins

ARE: Adenylate-uridylate-rich elements

ARE: Androgen response element

ARNm: ARN messager

BCA: Bicinchoninic acid assay

BHT: Barrière hémato testiculaire

BHT: Barrière hématotesticulaire

BioGRID: Biological General Repository for Interaction Datasets

C-ter: C- terminus

CBs: Chromatoid bodies

CDS: Coding sequence

CID: Collision-Induced Dissociation

CLPH: Casein-like phosphoprotein

CRABP: Cellular retinoic acid binding protein

CREM: Cyclic AMP response element modulator

CTB: Complexe tubullobulbaire

Da: Dalton

DHT: Dihydrotestostérone

DMR: Differentially methylated regions

DNAse: Désoxyribonucléase

DNMT: DNA methyltransferase

DTT: Dithiothréitol

E64: trans-Epoxysuccinyl-leucylamido(4-guanidino)butane

EDTA: Ethylenediaminetetraacetic acid

EGFR: Epidermal growth factor receptor

ENO4 : enolase family member 4

ERK: Extracellular signal regulated kinase

ERα: Estrogen receptor-alpha

ERβ: Estrogen receptor-beta

ES: Spécialisation ectoplasmique

ESI: Electrospray ionization (ionisation par électrospray)

EST: Expressed sequence tag

FDR: False discovery rate

FGF: Fibroblast growth factor

FPKM: FPKM = fragments per kilobase per million

FSH: Follicle Stimulating Hormone

FSTL3: Follistatin-like 3

GADPH: Glyceraldehyde-3-phosphate dehydrogenase I

GCNF: Germ cell nuclear factor

GnRH: Hormone gonadotropin releasing hormone

GO: Gene Ontology

GPR30: G protein-coupled receptor 30

HCD: Higher energy collisional dissociation

HGF: Hepatocyte growth factor

hnRNPs: heterogeneous nuclear ribonucleoprotein

ICPL: Isotope-coded protein label

IGF-1: Insulin like growth factor

IL-1: Interleukin 1

IL-6: Interleukin 6

KO: Knock-out

l'iCAT: Isotope coded affinity tag

LC-MS/MS: Chromatographie liquide et spectrométrie de masse en tandem

LC: Chromatographie liquide

LH: Luteinizing hormone

lncRNA: long non coding RNA

LTQ: Linear trap quadrupole

m/z: Rapport de la masse sur la charge

MALDI: Matrix-assisted laser desorption/ionization

MCM7: Minichromosome maintenance protein 7

miRNAs: micro-RNAs

MS: Mass spectrometry

MW: Poids moléculaire

MZT: Maternal-zygotic transition

N-ter: N-terminus

NCBI: National Center for Biotechnology Information

ncRNA: non coding RNA

NP40: Tergitol Type NP-40

ORF: Open reading frame

PABP: Poly (A) binding protein

PBS: Phosphate buffered saline

PEBP-1: Phosphatidylethanolamine binding protein

PGC: Primordial germ cells

pI: Point Isoélectrique

PIPES: 1,4-Piperazinediethanesulfonic acid, Piperazine-1,4-bis(2-ethanesulfonic acid)

piRNAs: Piwi-interacting RNAs

PIWL: Piwi-like protein

PTBP2: Neural polypyrimidine tract-binding protein

RARs: Retinoic acid receptors

RARγ: Retinoic acid receptor gamma

RBP: Retinol-binding protein

RBP: RNA binding protein

RGV: ReproGenomics Viewer

RISC: RNA-induced silencing complex

RNA-seq: RNA sequencing

RNAse: Ribonuclease

RNP: Ribonucleoprotein particle

RP-HPLC: Reversed phase liquid chromatography coupled with tandem mass spectrometry

RXRs: Retinoid X receptors

SDS: Sodium dodecyl sulfate

SGP 1: Sulfated glycoprotein 1

SGP 2: Sulfated glycoprotein 1 et 2

SGP: Sulfated glycoprotein

SILAC: Stable-isotope labelling by amino acids in cell culture

SNARE : Soluble N-éthylmaleimide-sensitive-factor Attachment protein

sncRNA: small non coding RNAs

SRC: Sarc-kinases

T-ENOL : Testicular enolase

TCTP: Translationally-controlled tumor protein homolog

TGF-β3: Transforming growth factor beta 3

TGFβ: Transforming growth factor beta

TMT:Tandem mass tag

TNF-α: Tumor necrosis factor alpha

TOF: Time Of Flight

TQ: Triple quadrupole

TUT: Testicular unannotated transcript

UTR: Untranslated region

VAMP7 : Vesicle-associated membrane protein 7

VAMP9 : Vesicle-associated membrane protein 9

INTRODUCTION **1**

I. La fonction testiculaire chez les mammifères 1

A. Le testicule *1*

B. La spermatogenèse chez les mammifères *7*

C. Régulation de la spermatogenèse *10*

II. L'analyse protéomique 39

A. De la cellule à la source du spectromètre *40*

B. La spectrométrie de masse en protéomique *42*

C. Approches de protéomique *52*

D. Les bases de données de séquence *63*

E. Ontologies et bases de données biologiques *68*

III. Les études protéogénomiques 73

A. La protéogénomique *74*

B. La protéogénomique au sens élargi *79*

IV. Approches « Omiques » et spermatogenèse 85

A. Transcriptomique et étude de la spermatogenèse *85*

B. La protéomique pour comprendre la spermatogenèse *88*

C. Génomique intégrative et spermatogenèse. *91*

OBJECTIFS **94**

MÉTHODOLOGIES **96**

II. Préparation des échantillons en vue des analyses protéomiques 98

A. Préparation de protéines *98*

B. Préparation d'extraits peptidiques *100*

III. Analyses protéomiques 101

A. Acquisition des données par analyse LC MS/MS *103*

B. Analyse des données *104*

C. Traitement des données de protéomique *105*

D. Utilisation de la Gene Ontology *108*

RÉSULTATS **110**

Chapitre 1 ..... **112**

Découverte de nouveaux loci codants par une approche protéomique informée par la transcriptomique (PIT) dans les cellules germinales de rat

I. Contexte et objectifs de l'étude 113
II. Résultats et discussion 114
A. Découverte de nouveaux gènes codants exprimés pendant la spermiogénèse 114
B. L'approche PIT et la réannotation du génome du rat 120
C. Une limite de l'approche PIT 127
III. Conclusion 129

ARTICLE 1 132

Forty-four novel protein-coding loci discovered using a PIT approach in rat male germ cells

Chapitre 2 134

Découverte de nouvelles isoformes spécifiques des cellules germinales chez le rat

I. Introduction 135
II. Résultats et discussion 137
A. Nouvelles isoformes potentielles spécifiques des cellules méiotiques et post méiotiques 137
B. Détection de nouveaux évènements d'épissage et d'UTR codants 141
C. Problèmes rencontrés lors de la détection de nouvelles isoformes potentielles 145
III. Conclusion 148

Chapitre 3 150

Analyse protéomique différentielle des protéines membranaires dans les cellules méiotiques et post-méiotiques et dans les corps résiduels chez le rat

IV. Contexte et objectifs de l'étude 151
V. Résultats et discussion 152

ARTICLE 2 154

Quantitative proteomic Isotope-Coded Protein Label (ICPL) analysis reveals 166 membrane proteins differentially expressed in rat meiotic and post-meiotic germ cells and in residual bodies

Chapitre 4 156

Identification de protéines impliquées dans le devenir des corps résiduels par protéomique Shotgun chez le rat

I. Contexte et objectifs 157
II. Méthodes 159

III. Résultats et discussion 160

DISCUSSION ET PERSPECTIVES 164

RÉFÉRENCES 174

ANNEXES..... 198

I. Communications scientifiques 199

A. Articles et revues à comité de lecture 199

B. Communications orales..... 199

C. Présentations affichées 201

II. Activités d'encadrement 204

III. Activité de vulgarisation scientifique 204

FIGURES

Figure 1. Anatomie du testicule de rat et structure de l'épithélium séminifère 3
Figure 2. Déroulement de la spermatogenèse et cycle de l'épithélium séminifère chez le rat 6
Figure 3. Représentation schématique des principales régulations humorales de la spermatogenèse 16
Figure 4. Transport des cellules germinales concomitants avec la spermiation dans l'épithélium séminifère de rat 22
Figure 5. Vue d'ensemble des régulations post-transcriptionnelles des ARNms 38
Figure 6. Les différents éléments composant un spectromètre de masse 43
Figure 7. Principe de la source MALDI 44
Figure 8. Principe de l'ionisation par éléctronébulisation 45
Figure 9. Schéma du LTQ-Orbitrap XL™ de ThermoFisher avec le trajet des ions 47
Figure 10. Expérience de LC-MS/MS générique en protéomique Shotgun 56
Figure 11. Comparaison schématique de deux approches protéomiques différentielles visant à quantifier les protéines dans différents échantillons 62
Figure 12. Annotation erronée dans UniProtKB 66
Figure 13. Parenté du terme de la Gene Ontology « acrosomal vesicle », vue de QuickGO 71
Figure 14. Structure d'un gène eucaryote comparée à celle d'un gène procaryote 73
Figure 15. Exemple d'alignements de séquences de peptides identifiés par MS sur une région génomique contenant un gène annoté 77
Figure 16. Cinq événements différents d'amélioration d'annotation d'un gène proposés par des peptides intragéniques 78
Figure 17. Principe du RNA sequencing 81
Figure 18. Principe de la protéogénomique 83
Figure 19. Schéma d'ensemble des différentes stratégies de protéomique utilisées 102
Figure 20. Stratégie d'analyse des protéines quantifiées par ICPL 107
Figure 21. Séquence traduite du transcrit TCONS_00003700 121
Figure 22. Profil d'expression du transcrit XLOC_003503 (TCONS_00003700) dans les cellules testiculaires isolées de rat, et dans le testicule comparé à d'autres tissus 123
Figure 23. Erreur sur le génome rn4 du rat au niveau du 1er exon du gène Wdr62 127
Figure 24. Proportion des différentes catégories de transcrits desquels dérivent les protéines identifiées par PIT dans les spermatocytes et les spermatides 138
Figure 25. Mise en évidence de nouvelles isoformes avec un saut d'exon, grâce aux peptides jonctionnels identifiés par l'approche PIT 142
Figure 26. Mise en évidence de nouveaux évènements d'épissage alternatif grâce aux peptides jonctionnels identifiés par l'approche PIT 144

Figure 27. Localisation d'un peptide identifié correspondant à un UTR : transcrit TCONS_00107908 sur le génome rn4 du rat _____ *144*

Figure 28. Nouvelle isoforme potentielle identifiée à cause d'une séquence écourtée _____ *146*

Figure 29. Nouvelle isoforme potentielle identifiée avec une séquence incorrecte _____ *147*

Figure 30. Localisation du gène Wdr62 et de nouveaux transcrits assemblés par RNA-seq à cette même localisation _____ *148*

Figure 31. Objectif de l'étude consistant à identifier des protéines potentiellement impliquées dans la formation et le devenir des corps résiduels chez le rat _____ *159*

TABLEAUX

Tableau 1. Liste des principaux facteurs sécrétés par les cellules de Sertoli _____ *20*

Tableau 2. Liste des gènes codant pour des facteurs de transcription pour lesquels l'invalidation chez la souris provoque des défauts de la fertilité chez le mâle _____ *28*

Tableau 3. Conditions d'élutriation des cellules germinales de rat adulte. _____ *98*

Tableau 4. Identification des séquences TCONS_00012662 et TCONS_00010279 (XLOC_001949) _____ *128*

Tableau 5. Séquences des protéines traduites de transcrits issus du même locus XLOC_001949 _____ *129*

Tableau 6. Termes de la Gene Ontology associés à 10 isoformes potentielles sélectionnées _____ *140*

Tableau 7. Nouvelles isoformes potentielles de WDR62 identifiées dans les cellules germinales _____ *147*

INTRODUCTION

I. La fonction testiculaire chez les mammifères

A. Le testicule

Le testicule est une glande amphicrine dont la fonction exocrine permet la production de gamètes mâles par le processus de spermatogenèse et dont la fonction endocrine concerne la production des hormones stéroïdes masculines (androgènes, essentiellement la testostérone). Chez les mammifères, le testicule est un organe des plus complexes, tant d'un point de vue structural que fonctionnel.

a) Anatomie et structure du testicule adulte

Le testicule, organe pair ovoïde est situé dans la plupart des cas chez les mammifères dans le scrotum, invagination péritonéale séparée de la cavité abdominale, ce qui lui permet d'être maintenu à une température de 34°C optimale pour la spermatogenèse normale chez les espèces exorchides (Stechell et al, 1994). Le testicule est maintenu à la base du scrotum par le ligament scrotal et est suspendu dans le sac scrotal par le cordon spermatique qui contient le canal déférent, des vaisseaux sanguins et lymphatiques, et des fibres nerveuses ortho et parasympathiques. La cavité scrotale communique avec la cavité abdominale par le canal inguinal. Le testicule est entouré de l'albuginée, une capsule conjonctive fibreuse, épaisse et résistante, riche en fibres de collagène. Les tubules séminifères qui composent le testicule sont séparés par du tissu interstitiel. Ils se rejoignent en débouchant par de courts segments rectilignes, les tubes droits, dans le rete testis qui se prolonge par les canaux efférents puis l'édididyme (Figure 1A).

Les tubules séminifères sont constitués de l'épithélium séminifère soutenu par une lame basale. Ces tubules sont séparés du tissu interstitiel par une gaine péri-tubulaire mince formée de la lame basale de l'épithélium séminifère, de fibres de collagène et des cellules myoïdes péritubulaires. Cette gaine est appelée la *lamina propria* ou *tunica* propria (Figure 1B). Entre les tubules, le tissu conjonctif lâche contient de nombreux vaisseaux sanguins et lymphatiques, des terminaisons nerveuses, des macrophages testiculaires, des mastocytes, des lymphocytes et des fibroblastes. Il contient aussi les cellules de Leydig isolées ou en petits îlots situés à proximité des capillaires. Ces cellules endocrines sécrètent essentiellement la testostérone et assurent ainsi la différenciation et le développement de l'appareil génital mâle, ainsi que l'apparition des caractères sexuels secondaires. La testostérone possède toutefois

une fonction paracrine envers les cellules de Sertoli. Celle-ci est cruciale, car elle contribue à l'initiation et au maintien de la spermatogenèse en collaboration avec l'hormone folliculo stimulante ou FSH (Follicle Stimulating Hormone).

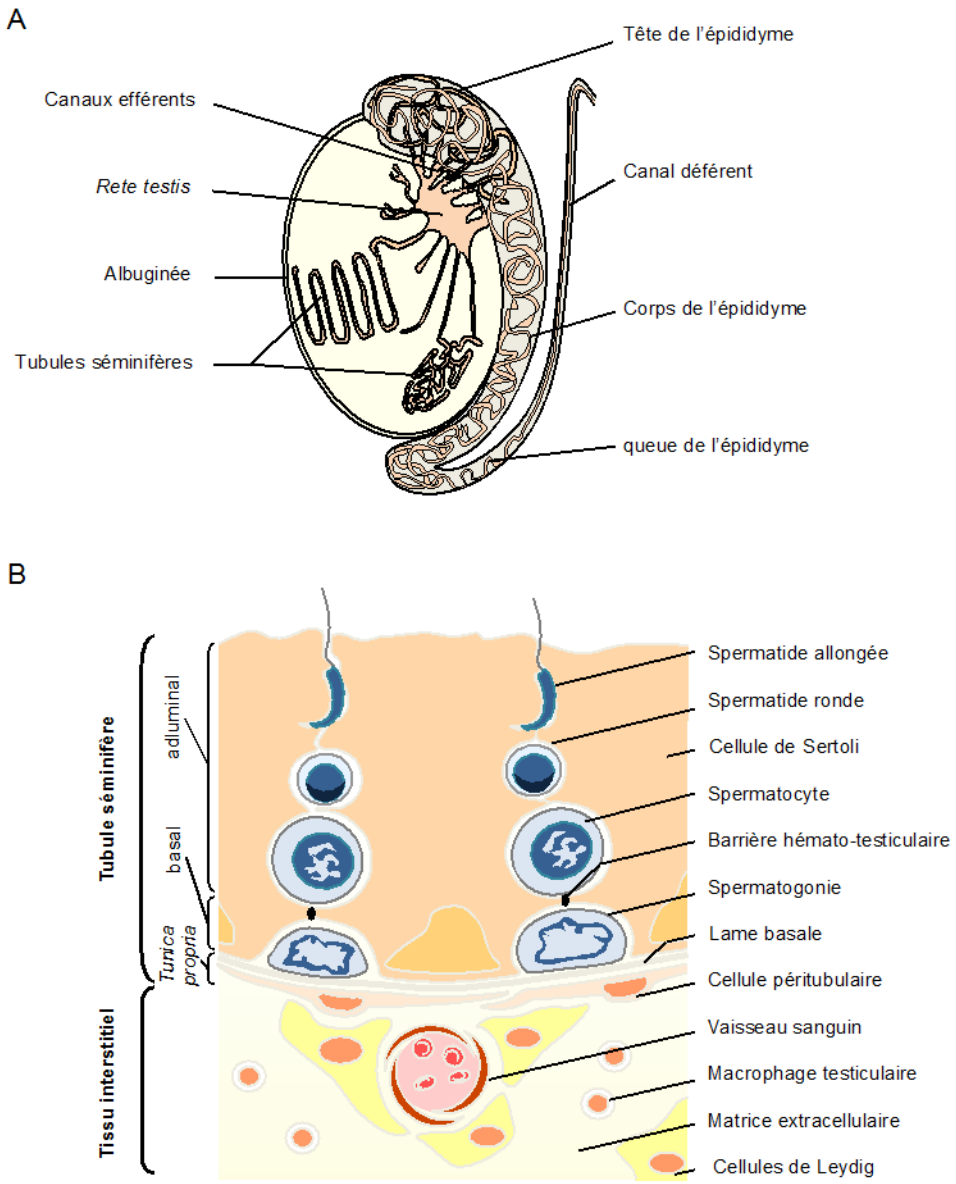

Figure 1. Anatomie du testicule de rat et structure de l'épithélium séminifère

A, Anatomie du testicule de rat. Une capsule fibreuse appelée albuginée entoure les testicules. Comme chez la plupart des mammifères, les tubules forment des boucles convolutées, leurs deux extrémités se connectant au *rete testis*. Les canaux efférents partent du rete testis et rejoignent l'épididyme. B, L'épithélium séminifère compose la paroi des tubules séminifères. Cet épithélium comporte deux types cellulaires : les cellules germinales (bleu) d'une part, et les cellules de Sertoli (beige) d'autre part. Les cellules de Sertoli s'étendent de la lame basale à la lumière du tube séminifère. Elles sont cylindriques et possèdent une structure tridimensionnelle complexe avec leurs nombreux prolongements cytoplasmiques qui leur permettent d'englober les cellules germinales en développement. Au tiers basal de l'épithélium, des jonctions serrées entre cellules de Sertoli sont observables. Ces complexes jonctionnels entrent dans la formation de la barrière hémato-testiculaire scindant en deux compartiments l'épithélium (basal et adluminal). Les spermatogonies, les spermatocytes pré-leptotène et leptotène se retrouvent donc dans le compartiment basal tandis que les autres cellules germinales se trouvent dans le compartiment adluminal, de l'autre côté de la barrière hémato-testiculaire.

b) L'épithélium séminifère

(1) Les cellules de Sertoli

Les cellules de Sertoli décrites pour la première fois par Enrico Sertoli en 1865 sont les cellules somatiques dans lesquelles s'enchâssent les cellules germinales mâles en développement : les spermatogonies, les spermatocytes, les spermatides rondes et les spermatides allongées. L'ensemble de ces cellules compose l'épithélium séminifère. Les celules de Sertoli reposant sur la membrane basale du tubule séminifère sont de forme pyramidale et leurs faces latérales sont en contact étroit avec les autres cellules de Sertoli et les cellules germinales, fournissant ainsi un soutien physique à ces dernières. Chaque cellule de Sertoli établit des jonctions serrées avec les cellules de Sertoli adjacentes. Elles créent ainsi une barrière immunologique imperméable appelée BHT (barrière hémato testiculaire) qui sépare les spermatogonies et les spermatocytes en début de méiose au côté basal, des spermatocytes terminant leur méiose et des spermatides au côté adluminal de l'épithélium séminifère (Figure 1B). Cette barrière est indispensable au développement et à la maturation de gamètes mâles normaux au sein du testicule adulte chez les mammifères. Les cellules de Sertoli sont en étroite interaction fonctionnelle avec les cellules germinales pour lesquelles elles exercent la fonction de cellules nourricières (Griswold et al., 1988). Elles entretiennent avec ces dernières le dialogue nécessaire au bon déroulement de la spermatogenèse et de la spermiogénèse. De ce dialogue dépend la production de 100 millions de spermatozoïdes par jour chez le rat, et jusqu'à 200 millions chez l'homme adultes (Dadoune and Démoulin, 1991). Certains de leurs produits de sécrétion, les facteurs Sertoliens agissent directement sur la membrane des cellules germinales (Jégou, 1995). Nous reviendrons plus en détail sur sujet dans la section « Dialogue entre les cellules de Sertoli et les cellules germinales » (page 17).

(2) La barrière hémato testiculaire

La BHT divise l'épithélium séminifère en un compartiment basal et un compartiment apical. Elle sépare ainsi les évènements du développement post méiotique des cellules germinales, c'est à dire la spermiogenèse et la spermiation ayant lieu dans le compartiment apical, de la circulation systémique. La BHT est constituée de complexes protéiques d'adhésion comme les jonctions serrées (formées par des occludines) coexistant avec des desmosomes (formés par exemple par la desmogleine 2, la desmocollin-2), des jonctions « gap » (formées par la connexin 43), des jonctions adhérentes nommées interface cellulaire basale Sertoli-Sertoli ou

spécialisations ectoplasmiques, et des faisceaux de filaments d'actine pris en sandwich entre les citernes de réticulum endoplasmique et la membrane apposée de la cellule de Sertoli. Au moment de la libération des spermatides âgées par l'épithélium séminifère, des cytokines telles que le facteur de croissance transformant TGF-β3 (Transforming growth factor beta 3) et le facteur de nécrose tumorale TNF-α (Tumor necrosis factor alpha) ainsi que la testostérone induisent une endocytose des protéines membranaires formant la BHT. Ces protéines sont ainsi internalisées par les cellules de Sertoli, ce qui déstabilise le site de la BHT et ouvre les jonctions serrées pour le transit des spermatocytes préleptotène entrant en méiose. Certaines protéines des jonctions serrées ou les N-cadhérines internalisées par endocytose sont destinées à la dégradation, tandis que d'autres sont recyclées pour la formation du nouveau site de la BHT sitôt le passage du spermatocyte préleptotène terminé. La synthèse *de novo* de protéines de la BHT est médiée par la testostérone. Des protéines de la BHT sont aussi présentes à la membrane des cellules germinales et facilitent leur transit en établissant des interactions homotypiques entre la cellule germinale et les cellules de Sertoli, empêchant ainsi à la BHT de s'ouvrir pendant la migration. Ces évènements ne compromettent donc pas l'intégrité de la barrière immunologique procurée par la BHT. Le passage de la BHT par les jeunes spermatocytes (préleptotène) est critique pour leur différenciation. Il initie leur progression dans le cycle cellulaire qui aboutit à la méiose. La dynamique de la BHT est sans doute soumise à d'autres régulations et est particulièrement sensible à un certain nombre de toxiques (Cheng and Mruk, 2012).

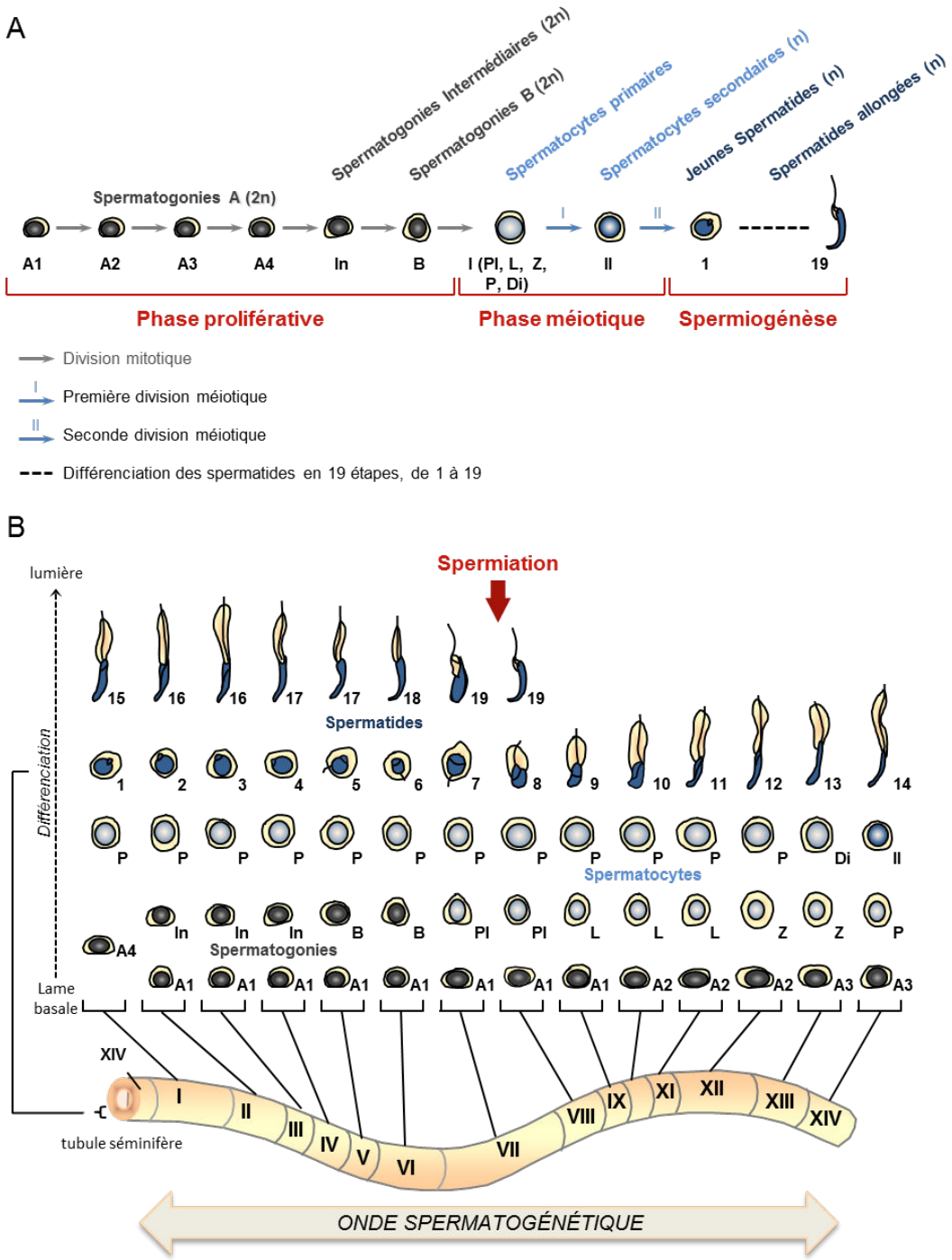

Figure 2. Déroulement de la spermatogenèse et cycle de l'épithélium séminifère chez le rat

A, Représentation schématique de la spermatogenèse chez le rat illustrant les phases proliférative et méiotique et la spermiogenèse à partir des spermatogonies différenciées (A1). Les stades d'autorenouvellement et de réplication des spermatogonies souches ne sont pas représentés. B, Représentation schématique des 14 stades du cycle de l'épithélium séminifère chez le rat sur la base du développement de l'acrosome des spermatides (selon Leblond et Clermont, 1952; Jégou et al., 1995). A1-A4: spermatogonies de type A; In: spermatogonies intermédiaires; B: spermatogonies de type B, Pl: spermatocytes pré-leptotène; L, Z, P, Di: spermatocytes leptotène, zygotène, pachytène et diplotène; II: spermatocytes secondaires, 1-19: les différents stades de la spermiogénèse; 1-8: jeunes spermatides ou spermatides rondes; 9-19: spermatides âgées en cours d'allongement ou allongées. Le début de la méiose c'est à dire l'entrée en prophase des spermatocytes Pl, a lieu au stade VII et VIII; les divisions méiotiques ont lieu au début et à la fin du stade XIV; la spermiation (flèche rouge) se produit au stade VIII. La différenciation des cellules germinales s'effectue dans le sens transversal des tubules, c'est à dire de la partie basale vers la lumière des tubules. Les différentes associations germinales, ou stades, sont coordonnées le long du tubule séminifère selon un ordre chronologique: c'est l'onde spermatogénétique. La taille de chaque segment représenté le long du tubule est proportionnelle à la durée de chaque stade du cycle.

B. La spermatogenèse chez les mammifères

a) Déroulement de la spermatogenèse

Chez les mammifères, la spermatogenèse est le processus de différenciation cellulaire ayant pour but la production de spermatozoïdes à partir des cellules germinales souches. Elle se déroule à partir de la puberté et tout au long de la vie adulte. La spermatogenèse a lieu au sein des tubules séminifères et est classiquement divisée en trois phases. Au cours de la première phase proliférative ou mitotique, les cellules germinales primitives, les spermatogonies, subissent une série de divisions mitotiques. Pendant la seconde phase méiotique, les spermatocytes primaires subissent deux divisions consécutives pour produire des spermatides haploïdes. Au cours de la troisième phase, la spermiogénèse, les spermatides se différencient en spermatozoïdes. La durée de la spermatogenèse varie selon les espèces, avec une durée de 35 jours chez la souris, 52 jours chez le rat et 74 jours chez l'homme (Clermont, 1972).

(1) Première phase, la phase mitotique:

Cette phase consiste en la production de spermatogonies matures, puis de spermatocytes primaires suite à une série de mitoses (Figure 2A). Notons avant tout qu'Il y a deux populations de cellules souches spermatogoniales indifférenciées qui dérivent des cellules germinales primordiales. La première est une population de cellules souches stables qui a la capacité de s'auto renouveler. La seconde est une population de cellules souches potentielles qui ne s'auto-renouvelle pas dans une situation normale. Les cellules de cette dernière population sont rapidement remplacées dans le testicule normal, ce qui suggère leur appartenance à une population d'amplification transitoire plutôt qu'à une population de cellules souches dormantes (Nakagawa et al., 2007), pour fournir le testicule en spermatogonies différenciées de manière continue. Ces cellules souches potentielles sont donc des progéniteurs et non des cellules souches "vraies". Ce sont ces progéniteurs qui se divisent pour donner les spermatogonies différenciées A1, qui vont subir un certain nombre de mitoses caractérisant cette phase, pour donner les spermatocytes primaires. Les cellules issues d'un même progéniteur restent liées par des ponts intercellulaires, marques d'une séparation incomplète des cytoplasmes après chaque division. La dernière étape de la phase mitotique est marquée par la transition des spermatogonies de type B matures, aux spermatocytes préleptotène par une dernière mitose (Figure 2A). Chaque spermatogonie A1

issue d'un progéniteur donne théoriquement 32 spermatogonies de type B et 64 spermatocytes primaires, (Dym and Fawcett, 1971).

(2) Seconde phase, la phase méiotique:

Cette phase concerne les spermatocytes qui subissent la méiose consistant en deux divisions successives permettant la formation de cellules haploïdes (à n chromosomes) à partir de cellules diploïdes (à 2n chromosomes). Les spermatocytes primaires préleptotène entrent dans la première division méiotique avec une prophase très longue divisée en cinq étapes: les stades leptotène, zygotène, pachytène, diplotène (cf. Figure 2A,B), et la diacinèse. La synthèse d'ADN a lieu au stade préleptotène. Pendant la prophase de cette première division et plus exactement pendant le stade zygotène se forme le complexe synaptonémal (Heyting et al., 1988) entre les chromosomes homologues, avec la formation de chiasmata (Moens, 1978), permettant des remaniements chromosomiques au sein de la chromatine. Ces remaniements permettent le brassage de l'information génétique. Les spermatocytes primaires qui terminent leur première division de méiose donnent naissance à deux spermatocytes secondaires. Ces spermatocytes secondaires subissent la seconde division de méiose, donnant naissance chacun à deux spermatides haploïdes.

(3) La troisième phase, la spermiogénèse:

Au cours de cette dernière phase, les spermatides issues de la méiose vont subir une importante différenciation morphologique et fonctionnelle qui va leur permettre de se différencier en spermatozoïdes aptes à la motilité lorsqu'ils se détachent des cellules de Sertoli et sont libérés dans le tubule séminifère. Cette phase ne comporte pas de division, mais une série de processus très spécialisés de différenciation cellulaire. Ces processus de différenciation incluent la formation de l'acrosome qui joue un rôle primordial lors de la fécondation. Ils incluent l'apparition des éléments clés pour l'acquisition de la motilité cellulaire, à savoir la formation et l'élongation du flagelle, le remodelage et la condensation de la chromatine grâce au remplacement des histones par les protéines nucléaires de transition puis par les protamines (Boskovic and Torres-Padilla, 2013; Lewis et al., 2003b; Yu et al., 2000) ; et enfin le réarrangement des mitochondries en une gaine mitochondriale dans la pièce intermédiaire des spermatozoïdes. Afin d'être motiles une fois libérées dans la lumière du tubule séminifère, les spermatides allongées réduisent de 80% leur volume cytoplasmique, puis se débarrassent du corps résiduel (Figure 2B) qui sera phagocyté par la

cellule de Sertoli. Il est probable que les processus de dégradation du corps résiduel au sein de la cellule de Sertoli jouent un rôle dans la suite du déroulement de la spermatogenèse, à savoir le déclenchement de nouvelles divisions des spermatogonies et donc d'une nouvelle vague spermatogénétique.

b) Le cycle et la vague spermatogénétiques

Les cellules germinales s'associent en compositions fixes, ou stades, constituant le cycle de l'épithélium séminifère. Chez le rat, le cycle de l'épithélium séminifère est divisé en 14 stades (I à XIV) selon Leblond et Clermont (Leblond and Clermont, 1952). La spermiogénèse, définie comme la transformation morphologique des spermatides en spermatozoïdes est de nouveau divisée en 19 étapes de différenciation (de 1 à 19, Figure 2A, B). Cette dernière phase de développement des cellules germinales est un exemple frappant et unique de différenciation cellulaire. Il implique les transformations morphologiques évoquées à la section précédente. Ces transformations permettent aux spermatozoïdes immatures d'acquérir la mobilité une fois l'épididyme atteint. La spermatogenèse s'achève au moment de la spermiation, au stade VIII de l'épithélium séminifère et au stade 19 de différenciation des spermatides. Les cycles de l'épithélium séminifère essentiels à une production continue de sperme sont dépendants de nombreux facteurs et sont espèce-spécifiques (Hess and Renato de Franca, 2008).

Il existe un ordre distinct des associations cellulaires le long des segments des tubules séminifères. Ces segments correspondent à différentes associations cellulaires. Un segment est défini par une portion longitudinale du tubule séminifère qui présente un seul stade. Les cellules germinales au sein de chaque couche de l'épithélium séminifère évoluent de manière synchrone avec les cellules d'autres couches, produisant la séquence des étapes décrites dans la Figure 2A. Les cellules ne migrent pas latéralement le long du tubule séminifère; en revanche, une succession des étapes y est observée. Ces étapes séquentielles se produisent à répétition le long des tubules et constituent la «vague» de l'épithélium séminifère (Perey et al., 1961). Autrement dit, au moins chez les rongeurs, la phase I est suivie de la phase II, puis III, etc jusqu'au stade XIV, puis de nouveau apparaît la phase I. Une vague englobe 14 segments chez le rat (Leblond and Clermont, 1952). 12 segments chez la souris (Ahmed and de Rooij, 2009) et 6 segments chez l'homme (Amann, 2008).

C. Régulation de la spermatogenèse

Nous avons vu que la spermatogenèse, processus de différenciation cellulaire unique, permet la production quotidienne de millions de spermatozoïdes. Cette fonction implique l'expression coordonnée de gènes spécifiques et la génération de produits de gènes spécifiques à chacune des étapes du processus, conjointement avec une communication continue entre les cellules germinales en développement et les cellules somatiques testiculaires. Les mécanismes de régulation de l'expression de ces gènes et de leurs produits sont nombreux, et ne sont pas complètement élucidés de nos jours. Le but de cette partie est d'illustrer le fait que leur complexité et leur complémentarité font de la spermatogenèse l'un des processus de différenciation les plus complexes de l'organisme, mais cette description de différents mécanismes de régulation ne se veut pas exhaustive.

a) Contrôle humoral

La spermatogenèse chez les mammifères nécessite l'action d'un assortiment complexe de peptides et d'hormones stéroïdes, dont chacun joue un rôle important dans le fonctionnement normal de l'épithélium séminifère. Ces messagers hormonaux sont essentiels à la régulation du développement des cellules germinales et à la prolifération et la fonction des cellules somatiques nécessaires pour le bon développement du testicule (McLachlan et al., 2002), dans l'interstitiel, des cellules de Leydig ont pour fonction principale la production de testostérone (Mendis-Handagama, 1997). Les cellules myoïdes péritubulaires contractiles (Maekawa et al., 1996) sont impliquées dans le transport des spermatozoïdes et du fluide testiculaire le long des tubules. Enfin, les cellules de Sertoli, constituent un soutien physique et nutritionnel pour la spermatogenèse (Griswold et al., 1988). Chacun de ces types cellulaires est la cible directe d'une ou de plusieurs hormones dont les actions sont essentielles pour la fertilité mâle.

(1) Les gonadotrophines

L'hormone FSH (Follicle-stimulating hormone*)* ou hormone folliculostimulante et l'hormone LH (Luteinizing hormone) ou hormone lutéinisante sont des hormones glycoprotéiques sécrétées par l'hypophyse antérieure sous le contrôle d'un "interrupteur principal" générateur d'impulsions, la GnRH (Gonadotropin Releasing Hormone) ou gonadolibérine, neurohormone synthétisée dans l'hypothalamus. Elle est elle même sous le contrôle de deux

systèmes de rétroaction séparés permettant un contrôle indépendant des androgènes (LH-testostérone) d'un côté et de la production de spermatozoïdes (FSH-inhibine) de l'autre. La FSH et la LH agissent directement sur les testicules pour stimuler la fonction des cellules somatiques en appui de la spermatogenèse. Chez le mâle, l'expression du récepteur de la FSH est limitée aux cellules de Sertoli testiculaires (Rannikki et al., 1995), tandis que les récepteurs à la LH se trouvent principalement dans les cellules de Leydig (Lei et al., 2001). La régulation de la synthèse de la testostérone semble être la seule fonction indispensable de la LH dans le testicule adulte (Lei et al., 2001). Chez les rongeurs, il semble que le rôle principal de la FSH dans la spermatogenèse soit la stimulation de la prolifération des cellules de Sertoli au cours du développement pré-pubère (Heckert and Griswold, 2002). Or, le nombre de cellules de Sertoli détermine en grande partie du nombre de cellules germinales à l'âge adulte (Orth et al., 1988). La FSH semble être impliquée dans la prolifération et la différenciation des spermatogonies chez le macaque (Simorangkir et al., 2009).

La FSH pourrait également jouer un rôle dans le processus de spermiation (Saito et al., 2000), bien que les androgènes soient des régulateurs plus importants de ce processus (O'Donnell et al., 2011). En effet la suppression de la FSH et des androgènes agit notamment sur la spermiation par le biais des cellules de Sertoli particulièrement sensibles à la suppression de ces hormones. La FSH et les androgènes agissent sur les cellules de Sertoli au stade VIII pour contrôler l'expression de micro-ARNs dont les cibles sont des gènes importants pour l'adhésion focale et de régulation du cytosquelette d'actine, processus connus comme associés à l'adhésion des spermatides aux cellules de Sertoli. Notons que deux de ces gènes, Pten, une phosphatase intracellulaire, et Eps1, un médiateur de l'endocytose, sont sous-exprimés par le retrait de ces hormones *in vivo* et possèdent des sites cibles du micro ARN miR-23b dans leur 3' UTR (Untranslated region), (Nicholls et al., 2011).

(2) Les stéroïdes

La testostérone et un de ses métabolites, la DHT (dihydrotestostérone) ainsi que l'estradiol (17β-oestradiol) sont désignées collectivement comme les hormones sexuelles en raison de leur rôle primordial dans la régulation des gonades et le développement des cellules germinales chez le mâle et la femelle ainsi que dans la différenciation sexuelle chez le mâle. La testostérone est produite sous l'influence de la LH par les cellules de Leydig localisées dans l'interstitium. Elle est l'un des régulateurs les plus importants de la spermatogenèse dans l'axe hypothalamo-hypophyso-testiculaire; en parallèle du 17β-oestradiol produit par les

cellules de Leydig et de Sertoli, ainsi que par les cellules germinales (Carreau and Hess, 2010; O'Donnell et al., 2001; Walker, 2010).

(a) Actions des androgènes sur la spermatogenèse

Les androgènes et leur récepteur nucléaire AR (androgen receptor) sont bien connus pour avoir un rôle dans la spermatogenèse normale et la fertilité. Il existe un seul AR chez le rat codé par le gène Ar sur le chromosome X. Les androgènes agissent en stimulant la spermatogenèse *via* une signalisation AR dans les cellules de Sertoli, et dans les cellules péritubulaires myéloïdes (Welsh et al., 2009). Bien que la stimulation de la spermatogenèse par les androgènes requière une action directe sur les cellules de Sertoli (O'Shaughnessy et al., 2010), la signalisation AR dans les cellules péritubulaires est aussi essentielle à la fonction testiculaire notamment la production de fluide séminifère et l'expression de gènes androgéno-dépendants dans les cellules de Sertoli (Welsh et al., 2009). L'AR dans les cellules de Leydig est nécessaire à la production de testostérone et à une spermatogenèse normale (Xu et al., 2007). L'AR s'exprime dans les artérioles, les cellules myéloïdes péritubulaires et les cellules de Leydig indépendemment du stade de l'épithélium séminifère adjacent, mais pas dans les cellules germinales (Bremner et al., 1994). Ces dernières ne semblent pas avoir besoin d'un AR intrinsèque, alors que son expression et ses fonctions dans les autres cellules testiculaires sont cruciales pour la spermatogenèse.

L'AR est exprimé dans les cellules de Sertoli entre les stades II et VIII du cycle spermatogénique et particulièrement aux stades VII et VIII, (Bremner et al., 1994), c'est à dire au moment où les spermatocytes primaires passent la BHT. Ce stade de l'épithélium séminifère correspond aussi au moment de la spermiation. Les androgènes sont d'ailleurs essentiels à la formation de la BHT en régulant les protéines de jonction (Meng et al., 2005; Wang et al., 2006). L'absence d'AR dans les cellules de Sertoli résulte en un blocage complet de la méiose indiquant que ces cellules sont le médiateur majeur des effets des androgènes sur la spermatogenèse (Verhoeven et al., 2010). L'action principale des androgènes sur la méiose vise à assurer la survie des spermatocytes pachytène et l'entrée en division des spermatocytes diplotène (Abel et al., 2008; De Gendt et al., 2004; Haywood et al., 2003). Le développement des spermatozoïdes matures requiert également la présence des androgènes et les stades de différenciation entre spermatides rondes et spermatides allongées sont particulièrement sensibles à la fonction AR dans les cellules de Sertoli (Holdcraft and Braun, 2004). Le rôle joué par les androgènes dans la spermiogénèse et la spermiation n'est pas clair.

En revanche, on sait que l'adhésion des spermatocytes aux cellules de Sertoli est dépendante des androgènes, puisque leur action est nécessaire pour éviter la rétention des spermatides allongées, matures, et la libération prématurée de spermatides rondes (O'Donnell et al., 1996, 2011). Notons que très récemment, une liste d'environ 1000 transcrits androgéno-dépendants et principalement exprimés dans les cellules de Sertoli ont été identifiés par RNA-seq et RiboTag chez la souris (De Gendt et al., 2014). Fait intéressant, il a également été montré qu'en plus de la voie classique médiée par l'action génomique du récepteur nucléaire AR, la testostérone peut agir par l'intermédiaire d'une voie non génomique qui implique le recrutement par l'AR de kinases SRC (sarc-kinases) et l'activation du récepteur au facteur de croissance épidermique EGFR (Epidermal growth factor receptor), (Cheng et al., 2007; Fix et al., 2004). Les androgènes agiraient par cette voie pour augmenter l'adhérence des spermatocytes et des spermatides aux cellules de Sertoli suivie par la libération des spermatides allongées et la maturation des spermatozoïdes (Shupe et al., 2011).

Les androgènes sont liés, *via* une action sur l'environnement tubulaire et un contrôle des concentrations en acide rétinoïque, au métabolisme de ce dernier et son action dans les testicules (O'Shaughnessy et al., 2007) qui sont associés à la régulation de la méiose. Le rétinol délivré aux cellules germinales lié à la RBP (retinol-binding protein), ou directement délivré aux spermatogonies par le sérum ou les cellules de Sertoli, est internalisé par le récepteur membranaire Stra6. Il est converti dans les spermatogonies en acide rétinoïque, qui agit, lié aux protéines CRABP (cellular retinoic acid binding protein) sur des récepteurs nucléaires tels que le RARγ (retinoic acid receptor gamma). Ce récepteur activé agit sur la transcription de nombreux gènes comme *Stra8* qui est indispensable pour la méiose (Hogarth and Griswold, 2010). Enfin, les androgènes sont également essentiels à la prolifération des cellules de Sertoli au cours du développement pré-pubertaire (O'Shaughnessy et al., 2012), et comme le nombre de cellules de Sertoli régule le nombre de cellules germinales (Orth et al., 1988) la stimulation androgénique au cours du développement permet de déterminer le nombre de ces cellules chez l'adulte (Orth et al., 1988).

(b) Action des oestrogènes sur la spermatogenèse

La complexité de l'action des oestrogènes sur la spermatogenèse vient des effets indirects multiples que ces hormones ont sur le testicule *via* une régulation endocrine. Par exemple les oestrogènes peuvent inhiber la spermatogenèse par l'inhibition de la sécrétion de LH, et donc l'inhibition des niveaux intra testiculaires de testostérone (Hunt et al., 2009). Par ailleurs, le

testicule possède une activité aromatase qui convertit les androgènes en oestrogènes (Dorrington et al., 1978). Les cellules germinales expriment, en plus des récepteurs nucléaires aux oestrogènes ERα (estrogen receptor-alpha) et ERβ (estrogen receptor-beta), le récepteur membranaire GPR30 (G protein-coupled receptor 30) qui médie une action membranaire rapide non génomique des oestrogènes *via* la voie EGFR/ERK (Extracellular signal regulated kinase), (Carreau et al., 2012; Chimento et al., 2010). L'expression de l'aromatase et du récepteur aux oestrogènes ERα sont nécessaires à la fertilité, et plus particulièrement à l'initiation de la spermiogenèse (Robertson et al., 1999; Sinkevicius et al., 2009). L'action des oestrogènes *via* la signalisation ERα est nécessaire pendant la période néonatale pour assurer la spermatogenèse chez l'adulte (Sinkevicius et al., 2009), probablement par une action sur la maturation des cellules de Sertoli, ou, chez les rats nouveau-nés, par une augmentation du nombre des gonocytes (Li et al., 1997). Les effets les plus clairs des oestrogènes sur la spermatogenèse sont visibles sur la spermiogenèse. De plus, il a été suggéré que la biogenèse de l'acrosome pourrait être un processus également dépendant des oestrogènes (Cacciola et al., 2013).

(3) L'activine et l'inhibine

L'activine et l'inhibine sont des hormones peptidiques produites dans les gonades, membres structurellement apparentés de la superfamille des facteurs de croissance transformant TGF-β (Transforming growth factor beta) pléiotrophes. Isolées pour la première fois à partir d'extraits d'ovaires, sur la base de la capacité à inhiber la production de FSH. L'inhibine est composée de deux sous-unités, d'une seule sous-unité α et de l'une des deux -sous-unités β homologues, A ou B, codées par des gènes distincts. Ces hétérodimères ont été nommés inhibine A et l'inhibine B, respectivement. Cependant, des homodimères et des hétérodimères des sous-unités β ont été découverts dans des extraits d'ovaires et de testicules ayant des actions opposées à celles de l'inhibine, c'est à dire, stimulant la production de FSH, et ont été nommés activines (Ling et al., 1986). L'activine et l'inhibine peuvent agir comme des régulateurs autocrine et/ou paracrine de la fonction testiculaire, et peuvent donc être considérés comme des facteurs locaux plutôt que des hormones. Les cellules de Sertoli produisent l'inhibine et sont la principale source de l'activine A dans le testicule de rat normal, avec des contributions supplémentaires potentiellement significatives des cellules péri-tubulaires. En effet, l'activine et les protéines de la même famille sont produites par la plupart des types cellulaires testiculaires et doivent plutôt être considérés comme des facteurs

de croissance locaux plutôt que des hormones dans le périmètre de la spermatogenèse. Les récepteurs à l'activine sont exprimés dans les cellules germinales et les cellules de Sertoli, avec pour certains une expression stade-spécifique. La modulation de la production et de l'activité de l'activine A agit dans le contrôle du cycle de l'épithélium séminifère (Hedger and Winnall, 2012). Au cours du cycle de l'épithélium séminifère chez le rat, il existe une relation inverse significative entre l'expression de l'activine A et l'inhibine B. Le cycle de l'épithélium séminifère est sous le contrôle des cellules spermatiques en développement (França et al., 1998) ce qui implique par conséquent, que les changements cycliques de l'inhibine B et l'activine A sont régis par les cellules spermatiques Les cellules germinales et en particulier les jeunes spermatides sont impliquées dans le contrôle de la production d'inhibine par les cellules de Sertoli (Pineau et al., 1990). L'activine A agit sur la spermatogenèse adulte, en stimulant et modulant la prolifération des spermatogonies et le développement de spermatocytes. Elle module également la régulation de la réponse hormonale de la cellule de Sertoli, en particulier à la spermiation ou immédiatement après la spermiation où elle connaît un pic d'expression entre la fin du stade VIII et le stade XI. L'activine A ou B, mais pas l'inhibine, a des effets régulateurs sur la prolifération des spermatogonies (Mather et al., 1990). Les protéines apparentées, la follistatine et la FSTL3 (follistatin-like 3) qui lient l'activine, agissent comme des antagonistes sur son activité, et leur surexpression provoque l'infertilité sans effets clairs sur les taux de FSH (Guo et al., 1998). Dans l'ensemble, les données disponibles montrent que les activines peuvent jouer un rôle de régulation dans le maintien de la spermatogenèse et assurer le développement normal et l'activité des cellules de Sertoli.

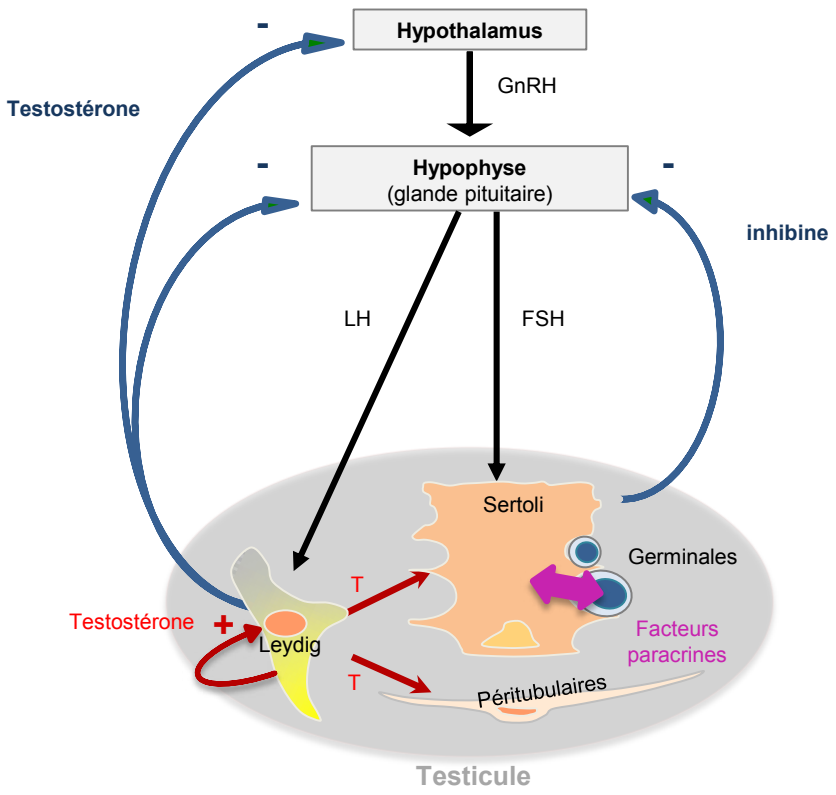

Figure 3. Représentation schématique des principales régulations humorales de la spermatogenèse

T, testostérone ; GnRH, gonadolibérine; FSH, hormone folliculo-stimulante ; LH, hormone lutheinisante.

b) Régulations autocrines et paracrines

Comme nous l'avons vu, la cellule de Sertoli de par sa position stratégique au sein de l'épithélium séminifère, est au centre des régulations de la spermatogenèse. Les cellules germinales et les cellules de Sertoli communiquent d'un point de vue anatomique et biochimique à l'aide d'un ensemble de structures (Jégou, 1993) qui vont permettre l'adhésion, le façonnage et le mouvement de cellules germinales du pôle basal au pôle apical de l'épithélium séminifère (jonctions adhérentes, desmosome-like, spécialisations ectoplasmiques), qui préviennent aussi la fuite des cellules germinales immatures de l'épithélium (Kopera et al., 2010). Il existe également des structures mixtes assurant le transfert de molécules (jonctions lacunaires) et des structures permettant un transfert de molécules par différents procédés d'endocytose. De plus, les cellules de Sertoli produisent le fluide des tubules séminifères (Rato et al., 2010) dans lequel sont sécrétés des facteurs solubles du compartiment basal au compartiment adluminal. Ce fluide joue également un rôle dans la spermiation et le transport des spermatozoïdes immatures.

(1) Dialogue entre les cellules de Sertoli et les cellules germinales

Séparées du sérum ou de la lymphe dans le compartiment adluminal par la BHT, les cellules germinales méiotiques et post-méiotiques sont au cœur d'un environnement local régi par les produits de sécrétion des cellules de Sertoli, et qui influence la méiose ainsi que le développement des spermatocytes et spermatides (Jégou, 1993). Il existe aussi bien une régulation de la spermatogenèse par les cellules support et nourricières qu'une régulation de la structure et de la fonction des cellules de Sertoli par les cellules germinales tout au long du processus spermatogénétique (Jégou et al., 1992). Les preuves se sont accumulées dans les années 1990 pour dire que la sécrétion de facteurs paracrine et autocrine par les cellules de Sertoli et les cellules germinales sont importantes pour le fonctionnement des deux types de cellules ; pour revues (Griswold, 1995; Griswold et al., 1988; Jégou, 1995). Ainsi, la sécrétion de facteurs Sertoliens relargués dans le fluide séminifère, aurait un rôle dans la différenciation des cellules germinales mâles (Jégou, 1993). On estime qu'une centaine de protéines différentes pourraient être sécrétées par la cellule de Sertoli. Une liste des principaux facteurs sécrétés par les cellules de Sertoli est présentée dans le Tableau 1. On peut les classer de la manière suivante :

- Hormones, facteurs de croissance et de différenciation
- Protéines de liaison et de transport
- Protéases et inhibiteurs de protéases tels que l'activateur du plasminogène
- Composants de la lame basale et de la matrice extracellulaire
- Agents antioxydants
- Métabolites énergétiques
- Constituants des complexes jonctionnels
- Protéines ayant une fonction inconnue

Il est connu à l'inverse, que les cellules germinales modulent la fonction des cellules de Sertoli *via* la sécrétion de facteurs solubles dans le fluide séminifère (Jégou et al., 1993; Onoda and Djakiew, 1993; Onoda et al., 1991; Pineau et al., 1993). Par exemple, des milieux conditionnés de cellules germinales stimulent la sécrétion des testines, de la clusterine et de la transferrine par les cellules de Sertoli (Le Magueresse et al., 1988; Pineau et al., 1993). La sécrétion de SGP1-SGP2 et CP-2/cathepsin L par les cellules de Sertoli est aussi modulée par des spermatides (McKinnell and Sharpe, 1997). Le rôle joué par les spermatides en cours d'allongement ou allongées dans le contrôle de la structure et la fonction des cellules de

Sertoli peut être positif (sécrétion de fluide tubulaire, d'ABP, inhibine (Jégou, 1991; Pineau et al., 1990)) ou négatif (testines, production de AMPc dépendante de la FSH (Jégou, 1993; Pineau et al., 1993)).

Les autres cellules germinales peuvent aussi contrôler la fonction des cellules de Sertoli à chaque stade de leur développement, mais les cellules de Sertoli semblent le plus régulées par les cellules germinales de la génération la plus avancée dans l'épithélium séminifère (Jégou, 1993).

Facteurs	Rôles supposés
Hormones ; facteurs de croissance et de différenciation	
Activine	Stimule la prolifération des spermatogonies
Inhibine	Inhibe la prolifération des spermatogonies, inhibe la FSH (Pineau et al., 1990)
TGFβ (Transforming growth factor beta) et IGF (Insulin like growth factor) 1, 2 et 10	Division et différenciation des cellules germinales. Le TGF β inhibe les fonctions leydiguiennes, régule l'immunité. L'IGF-1 : réplication de l'ADN des spermatogonies, stimule la stéroïdogenèse leydiguienne.
IL1α et IL6 (Interleukines 1α et 6)	L' IL1a stimule les mitoses et la méiose tandis que l'IL6 les inhibe. L'IL1a régule les fonctions sertoliennes et leydiguiennes, l'IL6, les fonctions sertoliennes uniquement.
Facteur Steel	Au cours du développement, ce facteur guiderait les cellules germinales primordiales vers les crêtes génitales et permettrait la prolifération des cellules germinales.
AMH (Anti mullerian hormone)	Elle induit la régression des canaux de Müller.
3HP (3 α -hydroxy-4-pregne-20-one)	Stéroïde qui inhibe la FSH et stimule le développement des spermatocytes primaires.
SCF (Stem cell factor)	Action directe sur le réceptur c-kit des spermatogonies
GDNF (glial cell-derived neurotrophic factor)	Maintien des spermatogonies dans un état indéfférencié (Sato et al., 2011b)
FGF (Fibroblast growth factor) -2	Induction de l'expression de la 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase PF3FB4 PFKFB3 dans les cellules germinales (Gómez et al., 2012)
FGF-9	Inhibition de la méiose des spermatogonies par l'induction de NANOS2 (Rossi and Dolci, 2013)
SCSGF (Sertoli cell secreted growth factor)	Prolifération des cellules de Sertoli (Lamb et al., 1991)
bFGF (basic fibroblast growth factor)	Suurvie (Li et al., 2012)
EGF (epidermal growth factor)	
JAGGED1	Induction des lymphocytes T régulateurs (Campese et al., 2014)
BMP4 (bone morphogenetic protein 4)	Différenciation des spermatogonies (Pellegrini et al., 2003)
Facteurs de mort cellulaire	
FasL	Immunoprotection (Cupp, 2014)
Inhibiteurs du complément	
CD59	Probablement présents à la surface des cellules germinales en différentiation (Cupp, 2014) puis, présents à la surface des spermatozoïdes les protègeant contre la lyse médiée par le complément dans le tractus génital femelle (Bozas et al., 1993)
DAF (decay activating factor) (MAC-inhibitory protein)	
Clusterine (SP-40)	
Protéines de liaison et de transport	
Transferrine et céruloplasmine	Transporteurs du fer et du cuivre, respectivement.
ABP (Androgen-Binding Protein)	Transport et stockage des androgènes.
RBP (Retinol-Binding Protein)	Transport du rétinol vers les cellules méiotiques et post-méiotiques qui le transforment en acide rétinoïque.
SGP1 (Sulfated GlycoProtein 1)	Transport de précurseurs de lipides et d'acides gras spécifiques.
SGP2 (Sulfated GlycoProtein 2)	Transport des lipides.
alpha-2-macroglobuline	Transport de facteurs impliqués dans les divisions germinales (cytokines, facteurs de croissance,...).
Gamma-GTP (g-Glutamyl TransPeptidase)	Transport du glutathion.
Protéases et inhibiteurs de protéases	
Activateur du Plasminogène	A certains stades, les AP dégradent les jonctions entre cellules de Sertoli ou entre les cellules de Sertoli et les cellules germinales.
CP2 (Cyclic Protein 2) / procathépsine L	probablement dans la libération des spermatozoïdes.
Cystatine C	Inhibiteur de la cathepsine L

Collagénase de type IV et autres métalloprotéinases	Impliqués dans le remodelage permanent de l'épithélium séminifère et de la membrane basale.
Composants de la matrice extracellulaire (MEC) et de la lame basale	
Collagène de type I et IV, laminine et protéoglycanes, fibulines	La MEC est indispensable à la polarisation des cellules de Sertoli, au stockage et divers facteurs, dont les facteurs de croissance.
Métabolites énergétiques	
Lactate et pyruvate	Indispensables aux cellules germinales qui ne peuvent pas métaboliser le glucose.
Agents anti-oxidants	
Glutathion	Pourrait être transféré aux cellules germinales.
Constituants des complexes jonctionnels	
Testines, ZO-1, ZO-2 (Zonula occludens 1 et 2), ...	Protéines des complexes jonctionnels.
Autres composants membranaires	
LRP (Liver-Regulated Protein)	Présente sur les cellules de Sertoli et les spermatocytes, la LRP permettrait leur interaction.

Tableau 1. Liste des principaux facteurs sécrétés par les cellules de Sertoli

Cette liste est adaptée de Jégou et al., 1995, (les références bibliographiques sont en partie tirées de (Jégou, 1993)).

(2) Les régulations liées au trafic cellulaire

Un trafic très actif au sein des cellules de Sertoli est nécessaire au bon support trophique des cellules germinales car celles-ci sont séparées de la circulation par la BHT. De nombreuses voies de trafic dans ces cellules ont donc pour fonction de médier la signalisation hormonale comme c'est le cas avec la FSH (Parvinen, 1982) ou de recycler des protéines carrier telles que la transferrine qui assure l'apport de fer aux cellules germinales au delà de la barrière hémato-testiculaire; ou bien encore, de se débarrasser des cellules germinales apoptotiques et des corps résiduels (Morales et al., 1985). Il existe deux compartiments d'endocytose dans les cellules de Sertoli. Les endosomes impliqués dans le captage de fluide séminifère ou la phagocytose de cellules apoptotiques et de corps résiduels à l'apex des cellules de Sertoli; et les endosomes impliqués dans la phagocytose récepteur-dépendante de différents ligands comme la transferrine au pôle basal des cellules de Sertoli (Morales and Clermont, 1986; Morales et al., 1985). La transferrine sérique et son récepteur sont internalisés, et contrairement aux hormones, aux LDL (low density lipoprotéines), aux virus et toxines qui sont aussi internalisés par leur récepteur dans les endosomes et adressés aux lysosomes pour être catabolisés; le complexe transferrine-récepteur reste intact, et la transferrine est restituée dans l'espace péricellulaire.

La phagocytose des corps résiduels par les cellules de Sertoli est un élément important dans la synchronisation de l'épithélium séminifère. Elle a lieu au moment de la spermiation au stade VIII (Figure 2 et 4) et provoque la sécrétion d'IL-1α par les cellules de Sertoli (Gérard et al., 1992). Or l'IL-1α est un régulateur majeur de la spermatogenèse par la régulation de la dynamique de la BHT (Lie et al., 2011), et par l'activation de la production de l'activateur du plasminogène par les cellules de Sertoli (Sigillo et al., 1998). L'activateur du plasminogène a lui même une action sur l'ouverture de la BHT, sur la synthèse d'ADN dans les spermatocytes, ainsi que sur la rupture des jonctions d'adhésion des spermatides matures avec les cellules de Sertoli, au moment de la spermiation. Il aurait un rôle dans la phagocytose des corps résiduels par les cellules de Sertoli (Guo et al., 2007; Liu, 2007),(Figure 4). Malgré les quelques connaissances sur le contenu du corps résiduel et des principaux phénomènes suivant sa phagocytose par les cellules de Sertoli, le rôle de ce dernier dans la synchronisation du processus spermatogénétique est loin d'être compris. Les corps résiduels, ainsi que les flagelles des spermatides expriment fortement de l'ER β (Sasso-Cerri, 2009), ce qui renforce le rôle de ces récepteurs et de leur signalisation associée, dans la spermiation. Les corps résiduels contiennent des corps multivésiculaires, des gouttelettes lipidiques, des aggrégats de ribosomes et des mitochondries condensées (Morales et al., 1985; Pineau et al., 1991). Ils contiennent aussi des peroxysomes impliqués dans le métabolisme des lipides et des espèces actives de l'oxygène (Dastig et al., 2011), dans la régulation de l'homéostasie des molécules de signalisation qui régulent la spermatogenèse tels que les rétinoïdes, et dans la protection des cellules germinales contre le stress oxydant. Les corps résiduels renferment aussi le corps chromatoïde (pour revue, voir (Parvinen, 2005)), impliqué dans la régulation de la traduction au cours de la spermatogenèse.

Dans ces régulations liées au trafic, il convient de mentionner les régulations de la BHT par des cytokines telles que le TGF-β3, le TNF-α, et par la testostérone, qui induisent l'endocytose des protéines d'adhésion de la BHT intégrées à la membrane des cellules de Sertoli. Ces protéines sont internalisées provoquant l'ouverture des jonctions serrées et la translocation des spermatocytes préleptotène vers le compartiment adluminal. Ces protéines sont ensuite, soit adressées vers la dégradation par la cellule de Sertoli, soit recyclées à un endroit de la membrane cellulaire derrière la cellule en transit, refermant ainsi la BHT sur son passage (Cheng and Mruk, 2012; Qian et al., 2014), (Figure 4). La BHT est aussi régulée par le facteur de croissance des hépatocytes ou HGF (Hepatocyte growth factor) *via* l'activation

de l'expression du TGF-β et la réduction de la quantité d'actine dans la région de la BHT qui modifie la morphologie du cytosquelette (Catizone et al., 2012).

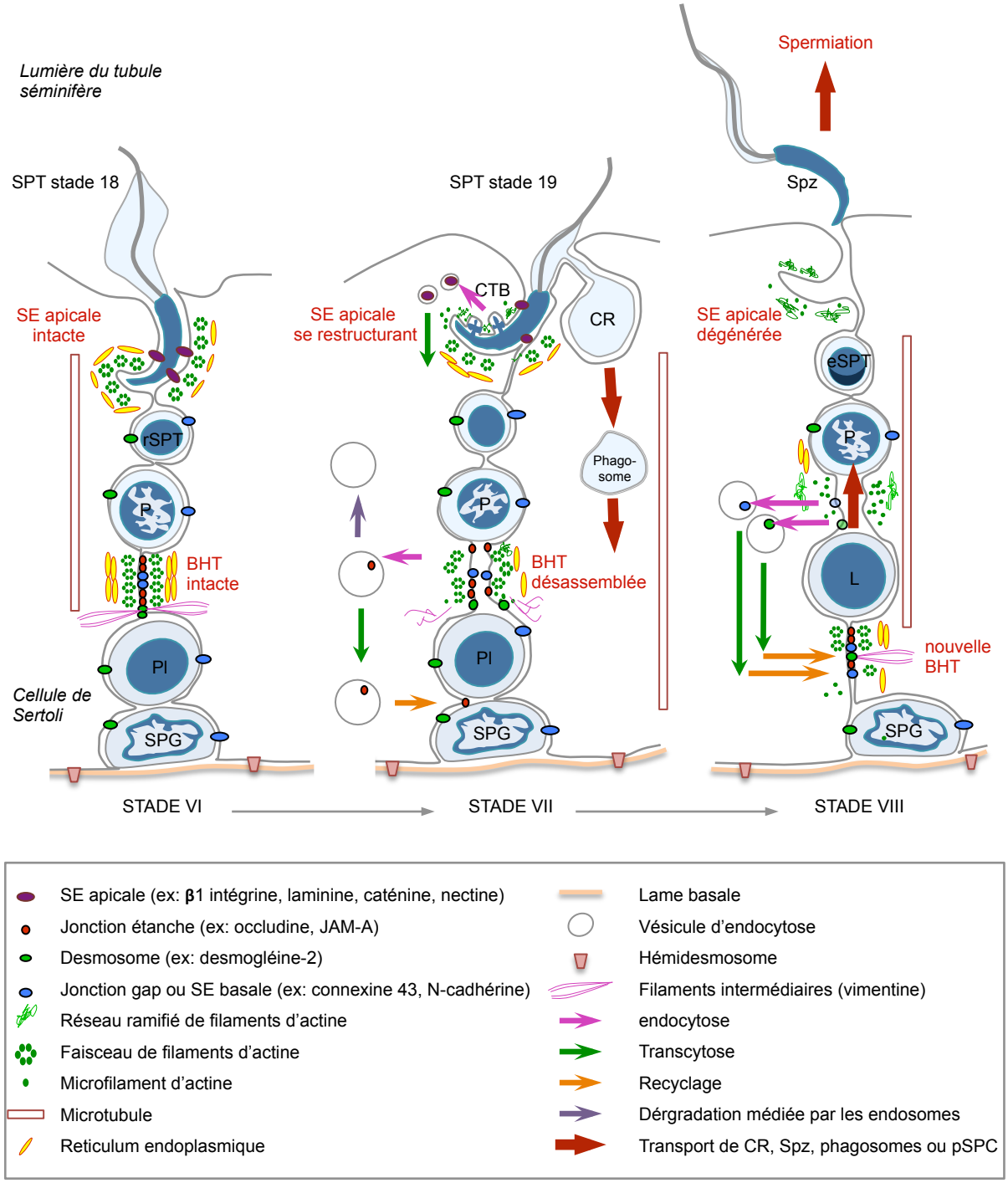

Figure 4. Transport des cellules germinales concomitants avec la spermiation dans l'épithélium séminifère de rat

Quelques complexes de protéines d'adhésion sont montrés: desmosomes, jonctions serrées, jonctions gap ou spécialisations ectoplasmiques (SE) entre les cellules germinales et les cellules de Sertoli, ou formant la BHT. Pendant le cycle de l'épithélium séminif!ère comme au stade VI, la SE apicale et la SE basale/BHT sont intactes, avec des microifilaments d'actine organisés en faisceaux maintenant l'intégrité de la BHT (gauche). Au stade VII à début du stade VIII, le complexe tubulobulbaire (CTB) apical apparait sur la face concave (ventrale) des têtes de spermatides, ce qui représente la restructuration de la SE apicale. En même temps, les protéines de la SE apicale telles que l'intégrine bêta-1 et les nectines 2 ou 3 peuvent subir la transcytose et être recyclées pour former une nouvelle SE apicale (milieu). Des évènements similaires ont lieu à la BHT, grâce à l'action de la testostérone ou de cytokines (ex: TGF-β3, TNF-α), provoquant l'endocytose de certaines protéines de la BHT et de leurs adapteurs, ceci étant aussi probablement du à un changement de leur état de phosphorylation par la FAK ou la Kinase SRC. L'ancienne BHT au dessus du spermatocyte preleptotène (Pl) en transit au travers de la BHT est alors restructurée, certaines molécules endocytosées sont vouées à la dégradation et certaines molécules comme l'occludine et JAM-A sont transcytosées et recyclées pour assembler la nouvelle BHT sous le spermatocyte Pl au stade VII-VIII. La formation de cette nouvelle BHT est aussi facilitée par une synthèse *de novo* de protéines de la BHT sous l'action de la testostérone. En même temps, la SE apicale qui entoure les têtes des spermatides au stade 19 continue de dégénérer jusqu'à ce que tous les microfilaments d'actine soient désorganisés, facilitant le relargage des spermatozoïdes immatures dans la lumière du tubule séminifère (spermiation). BHT, barrière hématotesticulaire; CTB, complexe tubullobulbaire; SE, spécialisation ectoplasmique, SPG, spermatogonie; Pl, spermatocyte préleptotène; L, spermatocyte leptotène; P, spermatocyte pachytène; SPT, spermatide; Spz, spermatozoïde; rSPT, spermatide ronde; eSPT, spermatide en allongement; CR, corps résiduel. Adapté de (Cheng and Mruk, 2012; Qian et al., 2014).

c) Régulation de l'expression des gènes et de leurs produits au cours de la spermatogenèse

Comme nous venons de l'entrevoir, la spermatogenèse est un processus complexe impliquant une succession d'étapes cruciales. Leurs régulations conditionnent la production des gamètes mâles et la survie de l'espèce. Le contrôle de la spermatogenèse implique de nombreux produits de gènes spécifiques du testicule dont beaucoup sont spécifiques des cellules germinales mâles et qui subissent des régulations strictes au cours du processus. Le testicule est donc considéré comme un des organes, si ce n'est l'organe le plus compliqué du corps. Il exprime le plus grand nombre de gènes tissus-spécifiques, comme le montrent un certain nombre d'études transcriptomiques à grande échelle récentes (Chalmel et al., 2007a, 2012, 2014a; Djureinovic et al., 2014; Laiho et al., 2013; Meikar et al., 2014; Soumillon et al., 2013). Des études de l'expression transcriptionnelle à l'échelle du génome ont permis d'identifier des milliers de gènes régulés dans le temps et l'espace au cours de l'ontogenèse testiculaire et de la différenciation des cellules germinales mâles (Chalmel et al., 2007a, 2012; Eddy, 2002; Schlecht et al., 2004; Schultz et al., 2003; Shima et al., 2004; Son et al., 2005; Wrobel and Primig, 2005). L'expression des protéines qui en découlent est tout aussi régulée dans l'espace et le temps comme le montrent un certain nombre d'études protéomiques différentielles (pour revues, voir (Calvel et al., 2010; Chocu et al., 2012)). Malgré l'importance biologique du développement des cellules germinales mâles, les

mécanismes sous-jacents de la régulation des gènes et de leurs produits spécifiques à chaque étape et transition cellulaire pendant la spermatogenèse restent encore mal documentés. Jusqu'à tout récemment, les études génomiques ont été largement limitées aux gènes codant pour des protéines. Mais il est aujourd'hui évident que les gènes codant pour des protéines ne représentent qu'un faible pourcentage du transcriptome (ensemble des transcrits, ou ARNs, produits par le génome d'un organisme donné); la majorité correspond à des transcrits non codants qui ne sont pas traduits en protéines. Bien que les petits ARNs non codants, les sncRNA (small non coding RNAs) tels que les micro-ARNs ou miRNAs (micro-RNAs), les ARNs interférants ou siRNA (small interacting RNAs), et les ARNs partenaires de protéines PIWI ou piRNAs (Piwi-interacting RNAs) soient largement étudiés dans le contexte de développement des cellules germinales mâles, une catégorie a nouvellement été mise en évidence, les longs ARNs non codants ou lncRNAs (long non coding RNAs) (Bánfai et al., 2012; Derrien et al., 2012; Mercer et al., 2009). Cette catégorie de transcrits semble d'après de récentes découvertes avoir une importante fonction de régulation de l'expression des gènes au cours de la spermatogenèse ((Bao et al., 2013; Chalmel et al., 2014; Laiho et al., 2013, 2013; Liang et al., 2014; Son et al., 2005), pour ne citer que quelques études). En effet, un grand nombre de lncRNAs sont exprimés dans le testicule chez les mammifères. Chez la souris, près de deux fois plus de lncRNAs sont transcrits dans le testicule que dans le cerveau et cinq fois plus que dans le foie (Soumillon et al., 2013) en accord avec ce qui a été montré chez l'homme (Cabili et al., 2011). Ils sont transcrits en grand nombre dans la lignée germinale et davantage dans les spermatocytes et les spermatides. Ces lncRNAs montrent un niveau de méthylation des CpG plus faible sur leurs régions promotrices prédites dans les spermatides comparé aux tissus somatiques et l'ouverture de la chromatine aux régions promotrices prédites de ces gènes transcrits en lncRNAs facilite également leur transcription dans les cellules germinales (Soumillon et al., 2013). Ces lncRNAs semblent spécifiques de chaque stade de la différenciation (Laiho et al., 2013) et s'expriment séquentiellement comme les ARNms (ARNs messagers) au cours de la spermatogenèse. Les lncRNAs et les ARNms présentent des changements coordonnés au cours de la spermatogenèse chez la souris (Liang et al., 2014). Cependant, leur rôle dans la régulation de la différentiation des cellules germinales mâles reste mal connu. Dans cette partie seront présentés les mécanismes de régulation connus de l'expression des gènes et de leurs produits dans la lignée germinale mâle.

(1) Régulations de la transcription

La régulation stringente et stade-spécifique de l'expression génique qui a lieu dans les cellules germinales, ainsi que la vague massive de transcription qui se produit juste après la méiose sont gouvernées par un mécanisme de transcription hautement spécialisé. Les profils d'expression des gènes restreints dans le temps et l'espace requièrent l'action de facteurs de transcription spécifiques, ou de facteurs généraux de transcription exprimés de manière différentielle dans les cellules germinales tout au long de spermatogenèse (pour revue (Lui and Cheng, 2008). Le rôle d'un certain nombre d'entre eux a été mis en évidence par des études d'invalidation de gènes (K.O.) chez la souris (Tableau 2).

(a) Facteurs de transcription généraux

L'expression différentielle de facteurs généraux de transcription et de leurs isoformes spécifiques du testicule dans les cellules germinales joue un rôle crucial pour assurer la transcription dans la lignée germinale. Ainsi le facteur de transcription TFIIB, la TBP (Tata binding protein) ou protéine de liaison aux boîtes TATA, et l'ARN polymérase s'accumulent dans les cellules germinales précoces haploïdes. Leurs niveaux d'expression y sont beaucoup plus élevés que dans les cellules somatiques testiculaires permettant aux jeunes spermatides d'accumuler assez d'ARNms pour leur développement jusqu'à la phase finale de la spermiogénèse. Des facteurs de transcription obtenus par clivage des facteurs de transcription généraux par des protéases spécifiques du testicule telles que la Taspase1 qui clive le TFIIA assurent une transcription « tissu-spécifique » dans le testicule (Oyama et al., 2013). Dans le cas mentionné, le clivage de TFIIA est nécessaire à la transcription des gènes codants pour les protéines de transition et les protamines.

(b) Les récepteurs nucléaires

Les androgènes essentiels à la spermatogenèse exercent leurs effets par le biais de l'AR, récepteur nucléaire qui régule l'expression des gènes cibles par la liaison sur leur région promotrice à un élément de réponse aux androgènes ou ARE (androgen response element) portant le consensus GGTACAnnnTGTTCT. L'expression du gène à homéo box X-linked Rhox5/PEM est un exemple typique de régulation par les androgènes dans le testicule. Des expériences d'invalidation de gènes avec des souris SC AR, souris K.O. pour l'AR dans les cellules de Sertoli, montrent bien le rôle que ce récepteur a sur la spermatogenèse et la stéroïdogenèse (Tableau 2). L'un des phénotypes observé de ces souris K.O. est une

augmentation de l'expression de l'AMH ou hormone anti Mullérienne conduisant à une baisse de la production de testostérone par les cellules de Leydig. Une réduction de l'apoptose des cellules germinales est observée chez ces souris, ainsi, notamment, qu'une réduction de l'expression de la cyclin A1 et de certains facteurs comme Sperm 1 importants pour le développement aux stades tardifs de la différenciation des cellules germinales.

D'autres récepteurs nucléaires jouent un rôle essentiel dans la spermatogenèse comme les récepteurs RARs (Retinoic acid receptors) de l'acide rétinoïque et les récepteurs X aux rétinoïdes, les RXRs (Retinoid X receptors), qui sont exprimés dans le testicule. Ils ont une fonction régulatrice des gènes cibles par la liaison à des éléments de réponse à l'acide rétinoïque appelées RARE (Retinoic-acid response element) ou RXRE (Retinoid X response element. L'activation des RARs et RXRs est essentielle pour la spermatogenèse. En effet, les rats déficients en vitamine A et les souris transgéniques pour RAR et RXR sont stériles.

(c) Les récepteurs orphelins

Les récepteurs orphelins sont des récepteurs nucléaires qui n'ont pas de ligand connu. Parmi eux le GCNF (Germ cell nuclear factor) est un nouveau membre de la superfamille des récepteurs nucléaires. Il régulerait l'expression temporelle des gènes cibles en coopération avec le modulateur de l'élément de réponse à l'AMPc ou CREM (cyclic AMP response element modulator) (Rajkovic et al., 2010). Le GCNF serait indispensable à l'expression spatio-temporelle de certains gènes pendant les phases méiotique et haploïde précoce de la spermatogenèse tels que les protamines pm-1 et-2 dans les spermatides rondes (Hummelke and Cooney, 2004).

(d) Autres facteurs de transcription impliqués dans une fonction testiculaire

De nombreux autres facteurs de transcription ont un rôle dans la fonction testiculaire. Les membres de la famille Basic-Domain-Leucine-Zipper (b-zip) sont connus pour être exprimés dans le testicule. Ils comprennent le modulateur CREM, la protéine de liaison de l'élément de réponse à l'AMPc ou CREB (cyclic AMP response element-binding protein). La protéine CREB active le facteur de transcription ATF1 en réponse à la voie de signalisation de l'AMPc, et le complexe CREB/ATF1 formé, se lie à l'élément de réponse à l'AMPc , le CRE (Cyclic AMP response element). De nombreuses isoformes de CREB sont générées par épissage alternatif dans le testicule.

Les facteurs de transcription de la famille des homéobox contiennent le motif de boîte homéotique qui est un domaine de liaison à l'ADN hautement conservé. Deux sous-familles de ces facteurs sont impliquées dans la spermatogenèse : la sous famille des Rhox (X-linked Reproductive homeobox) et une autre sous famille dont fait partie sperm 1 (Pearse et al., 1997).

Le facteur de transcription Oct-4 est exprimé dans les pro spermatogonies jusqu'à la naissance. Son expression continue dans les spermatogonies de type A (Pesce et al., 1998a). Sa baisse d'expression semble être l'un des déclencheurs de l'engagement en méiose des cellules germinales mâles (Pesce et al., 1998b).

Les facteurs de transcription de la famille de récepteurs à doigts de zinc tels que Gata-4 qui régule la différenciation testiculaire et d'expression de Sry (transcription factor sex-determining region Y) (Bagheri-Fam et al., 2010); Plzf (Costoya et al., 2004), et WT1(Gao et al., 2006), ont un rôle important dans la spermatogenèse.

La famille des NFk-B (Nuclear Factor Kappa B) est une famille de facteurs de transcription qui régulent une grande variété de gènes impliqués dans la spermatogenèse en activant et en réprimant la transcription de gène spécifiquement exprimés dans le testicule. Par exemple, le TNF-α induit la liaison de NF-kB à la protéine CREB dans les promoteurs AR dépendants, et augmente l'activité de ces promoteurs dans les cellules de Sertoli (Delfino and Walker, 1999).

Autre facteur, le facteur d'assemblage de la chromatine CAF-1 (Chromatin assembly factor-1) est un co-régulateur de RXR bêta dans la régulation de la transcription dans les cellules testiculaires somatiques, RXR étant exprimé dans les cellules de Sertoli et dans les cellules de Leydig (Nakamura et al., 2004).

Gène invalidé	Phenotype mâle
AR	Arrêt complet au stade pachytène, apparence féminine, baisse de la concentration sérique de testostérone
RARalpha	Arrêt complet de la spermatogenèse, dégénerescence de l'épithélium séminifère
RXR bêta	Stérilité, arrêt partiel au stade spermatocyte, anomalies des spermatozoïdes
GCNF	Létal
TR2	Testicule fonctionnel, nombre de spermatozoïdes normal et motilité normale

TR4	Retard dans le première phase de spermatogenèse, stades XI et XII prolongés, fertilité réduite
CREM	Arrêt complet au stade spermatocyte pachytène
CREB alpha et delta	Fertile
CREB Alpha, bêta et delta	Mort après la naissance
Plzf	Perte progressive de spermatogonies et augmentation de l'apoptose avec l'âge
Rhox5	Subfertile, fréquence supérieure de spermatocytes apoptotiques
Sperm-1	Subfertile
WT1	Souris knock-out conditionnel montrent une altération de la spermatogenèse
GATA-1, 4, 6	Létal
MSY2	Infertile
CAF1	Infertile

Tableau 2. Liste des gènes codant pour des facteurs de transcription pour lesquels l'invalidation chez la souris provoque des défauts de la fertilité chez le mâle

selon (Lui and Cheng, 2008).

(2) Régulations épigénétiques dans les cellules germinales

Le phénomène d'empreinte parentale mis en évidence dans les années 1990 est à l'origine du fait que le développement embryonnaire chez les mammifères requiert la présence d'un génome maternel et d'un génome paternel. En effet, la plupart des gènes des mammifères sont exprimés de façon bi- allélique, cependant certaines régions chromosomiques portent des gènes dont l'expression est mono-allélique et dépend de l'origine parentale. Ces gènes à empreinte sont regroupés dans le génome en domaines chromosomiques. Les premiers gènes soumis à empreinte parentale mis en évidence ont été les gènes de l'Igf2 (insulin-like growth factor2) et de son récepteur Igf2R (Barlow et al., 1991; DeChiara et al., 1990). Pour ces gènes, c'est l'empreinte parentale qui détermine leur expression différentielle aussi bien au cours du développement embryonnaire que pendant la vie adulte. Cette empreinte est caractérisée par une méthylation différentielle des di-nucléotides CpG ou CG (cytosine suivie d'une guanine) de l'ADN. Des éléments de contrôle, les DMR (differentially methylated regions), ont été identifiés au niveau de la plupart des gènes soumis à empreinte, et la méthylation de ces séquences contrôle la transcription des gènes concernés. Cette méthylation est cruciale pour le développement embryonnaire, comme en témoigne la létalité du K.O de la DMT1 (ADN méthyl transférase 1) chez les embryons de souris (Li et al., 1992).

Les cellules de la lignée germinale sont soumises à des modifications de l'empreinte parentale (méthylations et déméthylations programmées) sur leur génome car l'établissement des différences épigénétiques entre les deux chromosomes parentaux a lieu au cours de la gamétogenèse mâle et femelle. Ces modifications consistent en l'effacement de l'empreinte sur les deux chromosomes parentaux dans la nouvelle lignée germinale puis l'établissement d'une nouvelle empreinte au cours de la gamétogenèse selon le sexe de l'embryon. Il existe d'abord un processus d'effacement de l'empreinte par la perte de méthylation de l'ADN dans les PGC (primordial germ cells) ou cellules germinales primordiales, suivi d'une méthylation *de novo* pendant la phase d'établissement de l'empreinte dans les cellules germinales mâles et femelles (Reik et al., 2001). Alors que chez la femelle la mise en place de cette méthylation *de novo* n'est complète qu'après la naissance dans les ovocytes matures, elle intervient chez le mâle pendant la période prénatale dans les gonocytes quiescents, puis elle s'achève dans les spermatocytes avant la fin du stade pachytène (Reik et al., 2001). Après la fécondation, les premières phases du développement de l'embryon sont caractérisées par une vague de déméthylation complète du génome, suivie d'une reméthylation au stade blastocyste. Cette déméthylation massive permet d'effacer toute régulation provenant des cellules germinales parentales pour permettre la pluripotence des cellules embryonnaires et l'établissement d'un nouveau profil d'expression génique caractéristique du zygote. Cependant, les méthylations spécifiques des gènes soumis à empreinte établies pendant la gamétogenèse ne subissent pas cette déméthylation (Seisenberger et al., 2013).

Etant donné l'importance des méthylations de l'ADN au cours de la gamétogénèse, les ADN méthyltransférases ou DNMTs (DNA methyltransferases) et leurs régulateurs ont un rôle crucial notamment DNMT3A (DNA (cytosine-5)-methyltransferase 3A) ou DNMT3B (DNA (cytosine-5)-methyltransferase 3B) pour la spermatogenèse (Kaneda et al., 2004). L'invalidation du gène Dnmt3A génère un phénotype d'azoospermie et une altération dans la première vague de méiose, se traduisant par une perte de cellules germinales au stade leptotène, zygotène et pachytène (Yaman and Grandjean, 2006). Dnmt3a et Dnmt3L (DNA (cytosine-5)-methyltransferase 3-like) sont nécessaires pour la méthylation des régions imprimées dans les cellules germinales, mais suggèrent aussi l'implication d'autres facteurs (Kaneda et al., 2004). La DNMT3L en particulier contribue à la méthylation de l'ADN des loci à empreinte paternelle, et son K.O. entraîne un phénotype d'hypogonadisme sévère caractérisé par une perte progressive des cellules germinales par apoptose avec à l'âge adulte un phénotype « Sertoli Cell Only », un défaut de mitose des spermatogonies et un délai

d'entrée en méiose. Comme l'expression de Dnmt3L est restreinte aux gonocytes, on peut dire que la reprogrammation précoce du génome pendant la spermatogenèse est responsable de changements au niveau de la chromatine nécessaires à la suite normale la spermatogenèse (Webster et al., 2005). L'ADN méthyltransférase DNMT1 (DNA (cytosine-5)-methyltransferase 1) semble aussi nécessaire à la spermatogenèse car son K.O. entraîne l'apoptose des cellules germinales (Takashima et al., 2009).

Bien qu'une proportion importante du profil de méthylation du génome spécifique des cellules germinales soit acquis avant le stade spermatogonie A, il a été montré que les deux phénomènes de méthylation *de novo* et de déméthylation peuvent se produire au cours de la spermatogenèse d'une manière séquence-spécifique, mais pas uniquement pour les régions imprimées. Ces altérations ont lieu au cours des étapes précoces de la différenciation des cellules germinales et sont achevées au stade spermatocyte pachytène. Alors que certaines séquences subissent une méthylation *de novo*, à l'inverse un certain nombre de séquences se retrouvent progressivement déméthylées au cours de la progression vers le stade spermatocyte pachytène (Oakes et al., 2007). L'étude de méthylation de l'ADN au niveau des gènes Pgk-2 (Phosphoglycerate Kinase 2), Apo A1 (Apolipoprotein A1) et Pou5f1 (POU Class 5 Homeobox 1) a montré que ces gènes sont déméthylés au cours de la spermatogenèse tandis qu'ils se retrouvent méthylés au cours du transit des spermatozoïdes dans l'épididyme (McCarrey et al., 2005).

L'établissement de la méthylation sur une séquence d'ADN s'accompagne d'un changement de conformation de la chromatine *via* les modifications post-traductionnelles des histones telles que la méthylation et l'acétylation des résidus lysine. Des enzymes capables de modifier ces histones, comme par exemple les histone-désacétylases ou HDAC ; les histone-méthyltransférases ou HMT, participent au remodelage de la chromatine qui peut être permissive ou réprimée du point de vue de la transcription. L'acétylation des histones est associée à l'activation transcriptionnelle et la désacétylation à la répression. Ces modifications des histones sont également impliquées dans la régulation des gènes soumis à empreinte et sont étroitement associées à la méthylation de l'ADN (Lewis et al., 2004; Umlauf et al., 2004). Il est également reconnu que les ARNs non codants et les lncRNAs jouent un rôle dans la méthylation de l'ADN au cours de la spermatogenèse (Bao et al., 2013;

Peschansky and Wahlestedt, 2014). De même, les scnRNAs, comme les piRNAs qui peuvent recruter des modificateurs de la chromatine, induisant des changements de marques des histones (perte ou gain de méthylation) qui induisent l'activité *de novo* du complexe tétramérique DNMT3A/DNMT3L, essentiel pour la spermatogenèse normale (Aravin and Bourc'his, 2008).

d) Régulations post-transcriptionnelles

Les mécanismes de régulation post-transcriptionnelle jouent un rôle central dans la régulation de l'expression des gènes au cours de la spermatogenèse, ainsi qu'au cours de l'ovogenèse et du développement embryonnaire précoce. Chez la femelle, les ovocytes bloqués au stade diplotène de prophase de méiose subissent à la puberté une phase de croissance importante pendant laquelle ils acquièrent la capacité à reprendre leur méiose (qui reprend dans les follicules pré ovulatoires et se termine après la fécondation), puis acquièrent la compétence à assumer la fécondation et le développement de l'embryon: capacité au développement. Comme la spermatogenèse, la croissance ovocytaire est accompagnée de changements radicaux dans l'expression des gènes. Au milieu de la phase de croissance, la transcription diminue et les ovocytes arrivés au terme de leur croissance sont inactifs d'un point de vue transcriptionnel (Moore and Lintern-Moore, 1978). La quiescence transcriptionnelle accompagnée de la redistribution de la chromatine condensée autour du nucléole, et de modifications post-traductionnelles des histones (Kageyama et al., 2007), est nécessaire à l'acquisition de la compétence de développement (De La Fuente and Eppig, 2001). L'ovocyte en croissance accumule avant cette quiescence transcriptionnelle une grande quantité d'ARNs qui seront nécessaires au développement précoce de l'embryon. Après la fécondation, le génome de l'embryon reste en effet inactif d'un point de vue de la transcription, et bien que certains gènes zygotiques puissent être transcrits précocement, l'activation majeure du génome zygotique n'a lieu qu'après un certain stade de clivage qui dépend alors de l'espèce (Andéol, 1996), lors de la transition materno-zygotique ou MZT (maternal-zygotic transition). La stabilité des ARNs maternels joue donc un rôle majeur pendant les stades précoces de développement de l'embryon. Elle est assurée dans l'ovocyte par des mécanismes de régulation négative qui visent à réprimer la traduction des ARNs cibles jusqu'à ce que leur traduction ne soit activée. Ces mécanismes sont étroitement connectés à la localisation des ARNs messagers. La stabilité des ARNms dans l'ovocyte ainsi que la régulation de leur traduction et leur localisation impliquent des variations dans la longueur de

leur queue poly (A) ; la liaison des ARNs à des protéines de liaison aux ARNs, les RBPs (RNA binding protéines) ; l'intervention de micro ARNs ou de petits ARNs interférants, les siRNA dérivés de pseudogènes (Tam et al., 2008) ; ainsi que l'interaction entre miRNA et RBPs (Bettegowda and Smith, 2007). Pendant la croissance ovocytaire, beaucoup d'ARNms sont donc synthétisés, puis réprimés du point de vue de la traduction, et transportés dans des particules ribonucléoprotéiques, les RNPs (ribonucleoprotein particles), les granules de cellules germinales, les GCGs (Germ cell granules), qui correspondent aux granules polaires chez la drosophile et aux granules germinales chez le xénope ; à différentes localisations de l'ovocyte. Ainsi, les GCGs contiennent des protéines impliquées dans l'initiation de la traduction, le contrôle de la traduction et la dégradation des ARNms selon leur fonction de régulation de l'expression des ARNs maternels (Anderson and Kedersha, 2006).

La biogénèse de miRNAs maternels, nécessaire à la maturation des ovocytes et transmis au zygote, les plus abondants étant ceux de la famille Let-7, est cruciale au développement précoce de l'embryon chez les mammifères (Tang et al., 2007). En revanche, les miRNAs d'origine spermatique ne semblent pas contribuer aux miRNAs présents dans l'embryon précoce (Amanai et al., 2006). Des miRNAs zygotiques d'autres familles, les plus abondants sont ceux du miR-290 cluster, prennent le relais après le MZT (Tang et al., 2007). Au cours de la MZT, le génome de l'embryon prend le contrôle du développement et les produits de gènes maternels doivent alors être éliminés. La déstabilisation des ARNms maternels qui peut s'étaler sur plusieurs stades de l'embryogenèse incluant blastula et gastrula, est médiée par la liaison : 1) d'ARNs régulateurs (miRNAs) sur des séquences spécifiques de leur 3'UTR, dont une fonction clé est de réprimer l'expression des ARNms cibles par une complémentation de séquence qui aboutit à la réduction de leur abondance et / ou l'inhibition de leur traduction (Bagga et al., 2005) ou 2), de protéines régulatrices telles que Smaug dans la région 3'UTR des messagers maternels conduisant à leur déadénylation puis leur dégradation. (Schier, 2007). Des petits ARNs interférants, les siRNAs qui ont pour fonction de réguler le niveau de production de protéines, entrent aussi en jeu dans la dégradation des ARNms maternels aux stades précoces de l'embryogénèse, permettant le développement au delà de la MZT (Lykke-Andersen et al., 2008). Rappelons que les siRNAs sont produits à partir de long ARN double brin par clivage médié par la RNase III Dicer. En revanche, les miRNAs proviennent de longs transcrits primaires présentant des structures en épingle à cheveux qui sont clivées par un complexe microprocesseur composé de la protéine DGCR8 et de la RNase III Drosha, puis les miRNAs sont produits à partir des précurseurs en épingle à cheveux *via* leur clivage

par Dicer. Les siRNAs et les miRNAs sont chargés sur des protéines Argonaute (AGO) qui sont acteurs du silencing d'ARNms (Rana, 2007).

Chez le mâle, la transcription *de novo* s'arrête à la moitié de la spermiogénèse, après la méiose, contrairement à ce qui se produit dans l'ovocyte, de par la condensation de la chromatine juste après le stade spermatides rondes, bien avant la différenciation complète (Sassone-Corsi, 2002). Pour compenser le manque de nouveaux ARNs transcrits, les ARNs synthétisés doivent être stockés, pour les mêmes raisons que dans l'ovocyte, de sorte que leur traduction puisse coïncider avec la demande en temps voulu, aux étapes plus tardives de la différenciation, car la production de protéines stade-spécifiques ne peut pas être anticipée. Les ARNms de protéines des spermatides allongées et des spermatozoïdes, par exemple ceux des protéines de transition et des protamines remplaçant séquentiellement les histones, (Kleene et al., 1984) doivent donc être préparés à l'avance. Il y a alors nécessairement une régulation de la traduction de ces ARNms aux stades plus précoces de la spermatogenèse pendant lesquels la transcription est encore possible (Figure 5).

Chez les mammifères, les gamètes mâles ou femelles en formation gèrent donc le stockage, l'activation de la traduction régulée dans le temps et le silencing d'une grande quantité et diversité d'ARNms selon des mécanismes communs tels que la modification de la queue poly (A) des transcrits, le transport dans les RNPs et la concurrence avec les machineries de traduction. Certains composants des RNPs tels que la protéine VASA peuvent être communs aux cellules germinales mâles et femelles (Eulalio et al., 2007). Ceux développés ci-après sont ceux impliqués dans la régulation post-transcriptionnelle pendant la spermiogénèse.

L'allongement de la queue poly (A) d'un transcrit peut augmenter les taux de la traduction jusqu'à 100 fois. Ce processus est médié par la protéine de liaison poly (A) PABP (poly (A) binding protein), qui recrute des facteurs d'initiation de la traduction, EIF4G puis EIF4E (Sheets et al., 1994). Cependant, un mécanisme inverse se produit dans les spermatides où la traduction est associée à un raccourcissement rapide de la queue poly(A) (Yanagiya et al., 2010).

Un mécanisme plus important consiste à diriger le transcrit soit aux RNPs, compartiments cytoplasmiques particuliers qui sont généralement corrélés avec l'inhibition traductionnelle,

soit aux polyribosomes, ou polysomes, qui sont les sites de traduction (Iguchi et al., 2006). Les protéines liant les ARNs, ont été décrites pour leur rôle dans le décalage temporel entre la transcription et la traduction (Sassone-Corsi, 2002). Ces RBPs ont donc un rôle crucial dans le bon déroulement de la spermatogenèse. Typiquement, une RBP est exprimée dans le noyau et se lie à l'ARNm précurseur (pré-ARNm). Après maturation du transcrit, la RBP peut y rester pendant un certain temps avant d'être modifiée pour passer dans le cytoplasme, portant avec elle le transcrit. Là, le transcrit sera adressé à certains compartiments cytoplasmiques: les RNPs. En cas de besoin pour la cellule, la RBP s'associera aux polysomes et le transcrit associé sera traduit. Les motifs présents sur les extrémités non traduites d'un transcrit, les UTRs, sont décisifs pour leur transport du noyau au cytoplasme, leur stabilisation et leur efficacité traductionnelle. Ce sont ces motifs que reconnaissent les RBPs pour réguler le devenir des transcrits dans les RNPs.

(1) Les RNPs : compartiments cytoplasmiques régulant le devenir des ARNms

Les différents RNPs contiennent des enzymes différentes. Par exemple, les exosomes et les P-bodies (Processing bodies) possèdent des exonucléases qui dégradent les ARNms, alors que les autres n'en possèdent pas (Anderson and Kedersha, 2008). Les principaux RNPs sont décrits ci-dessous et régulent différemment les ARNms du point de vue de la traduction :

- les polysomes (polyribosomes) ou ergosomes, sites principaux de la traduction (Iguchi et al., 2006), sont parfois considérés comme des RNPs.

- les granules de stress, centres de tri en condition de stress provoquant en général l'arrêt de la traduction (Ivanov and Nadezhdina, 2006).

- les P-bodies associés à la dégradation des transcrits ou à leur stockage, qui possèdent des enzymes hélicases telles que DDX6 à domaine DEAD (Asp-Glu-Ala-Asp), qui sont des enzymes impliquées dans l'élimination de la coiffe et la dégradation 5'-3' des transcrits (Bettegowda and Wilkinson, 2010),

- les exosomes, sites de dégradation des transcrits qui contiennent des exoribonucléases 3'-5' (à ne pas confondre avec les exosomes de sécrétion (van Dijk et al., 2007)).

- les granules germinales mâles : IMC (Inter-mitochondrial cement) ou ciment intermitochondrial et CBs (Chromatoid bodies) ou corps chromatoïdes, qui sont deux

catégories majeures de RNPs détectées dans les cellules germinales mâles en développement. Les CBs sont des structures denses périnucléaires apparaissant au stade spermatocyte pachytène, et sont éliminés des spermatides allongées dans le corps résiduel. Les corps chromatoïdes sont composés d'ARNs et de RBPs ainsi que d'autres protéines impliquées dans la régulation des ARNs. Ils contiennent des piRNA pachytène, des ARNms, et des lncRNAs, et sont un RNP spécifique des cellules germinales (Meikar et al., 2014). Ils accumulent des ARNs naissant pendant les étapes de différenciation des spermatides rondes. Ces structures sont similaires aux P-bodies, mais n'existent que dans les cellules germinales mâles. La formation des corps chromatoïdes semble essentielle à la spermatogenèse (Chuma et al., 2009; Nguyen Chi et al., 2009). Basé sur le fait qu'ils contiennent des ARN-hélicases, des enzymes d'élimination de la coiffe, un homologue de VASA (DDX4), des protéines AGO, des PABPs, des miRNAs, et la protéine HUR ; on leur attribue des fonctions de stockage des ARNs de longue durée de vie, de dégradation des transcrits déadénylés, et de régulation de l'activité des miRNAs (Bettegowda and Wilkinson, 2010; Nguyen Chi et al., 2009). La protéine HUR comme la protéine MHV sont très importantes, car leur mouvement est associé à la régulation des transcrits pendant les étapes 1 à 5. Les transcrits associés sont accumulés dans les corps chromatoïdes, puis sont dirigés vers les polysomes après l'étape 6 (Kotaja et al., 2006; Nguyen Chi et al., 2009). Notons que DICER1, une endonucléase impliquée dans la biogenèse des miRNAs est indispensable pour la méiose et la spermiogénèse (Romero et al., 2011).

- les splicéosomes, RNPs spécialisées dans l'épissage, agissant dans le noyau. Elles enlèvent les introns d'un pre-mRNA, et lient les exons pour former un transcrit mature. Ses composants clés sont des petits ARNs nucléaires ou snRNAs (small nuclear RNAs) et leurs protéines associées. Ensemble, ils sont aussi appelés snRNPs (Kaida et al., 2010).

- les complexes hnRNPs (heterogeneous nuclear ribonucleoprotein) ou protéines RNP nucléaires hétérogènes, lient les pre-mRNAs (ou hnRNAs) et sont responsables de l'ajout de la queue poly (A), et de la coiffe 5' 7 méthylguanosine aux pre-mRNAs (DeGracia et al., 2008). Ils sont souvent responsables de l'export nucléaire (Xie et al., 2003).

- les granules Elav/HU, centres de relais qui facilitent le transfert des transcrits à d'autres RNPs, aussi appelés « RNA opérons » (Keene, 2007).

(2) Eléments « codes secrets » dans les UTRs des transcrits

Les motifs liés par les RBPs et les miRNAs sont d'une importance vitale. Ceux-ci sont les éléments qui déterminent quels complexes protéiques vont se lier aux transcrits. Le nombre et la localisation relative des autres éléments ou motifs ainsi que les nuances de séquences vont réguler finement la liaison à des RBPs et des miRNAs variés, déterminant ainsi le devenir post-transcriptionnel du transcrit. Ces motifs reconnus par les RBPs et les miRNAs et qui sont généralement présents sur les 3'UTRs des ARNms, sont des éléments *cis* régulateurs. Ils existent sur une grande variété de transcrits dans différents types cellulaires. Ils sont d'une grande variabilité et il est impossible de décrire lesquels vont être reconnus par telle ou telle RBP. On sait toutefois qu'une altération du « code » altérera la régulation du transcrit. Il en existe de nombreux dont les plus connus sont les éléments riches en adénylate/uridylate ou ARE (Adenylate-uridylate-rich elements) impliqués dans la dégradation ARE; les éléments EDEN (Embryo deadenylation elements), impliqués dans la déstabilisation et la répression des transcrits ; ainsi que les éléments de polyadénylation cytoplasmiques CPE (Cytoplasmic polyadenylation elements), encourageant l'adénylation et la traduction, (pour revue, voir (Idler and Yan, 2012).

(3) Les RBPs, les effecteurs ultimes dans le contrôle des mRNAs

On trouve trois catégories de RBPs. Celles qui ne sont pas séquence-spécifiques, celles qui s'associent aux petits ARNs non codants ou sncRNAs dans des complexes effecteurs, et celles qui lient des éléments (motifs) spécifiques des ARNs décrits plus haut. Les RBPs sont nombreuses. Elles peuvent exister aussi bien dans le noyau que dans le cytoplasme et sont en général capables de transiter entre les deux. Elles peuvent aussi être classées dans les catégories suivantes : 1) celles qui s'associent aux sncRNAs : les sncRNA-associated RBPs ; 2) les RBPs qui lient les ARE : les ARE-BPs (ARE elements binding proteins) ; 3) les RBPs qui ne lient pas les ARE : les non-ARE-BPs.

(a) Les RBPs associées aux scnRNA

Parmi les RBPs associés aux sncRNAs (small non coding RNAs), les RBPs dépendantes des miRNAs lient ces derniers comme composants du complexe RISC (RNA-induced silencing complex). Ces miRNAs se lient à leur séquence cible spécifique, habituellement dans le 3' UTR des longs transcrits, conférant une activité motif-spécifique au complexe RISC associé

(Agami, 2010). Jusqu'à récemment, on pensait que ceci résultait en la dégradation de l'ARNm ou l'inhibition de sa traduction par la perturbation d' EIF4 (facteur eucaryote d'initiation de la traduction 4) et EIF6 par les protéines AGO (Argonaute), (Chendrimada et al., 2007). Toutefois, de nouveaux éléments prouvent que sous certaines conditions, les miRNAs de mammifères peuvent encourager la traduction en recrutant un complexe Ago2-FXR1 (Argonaute-fragile X mental retardation) (Vasudevan and Steitz, 2007).

Un important sous ensemble de protéines AGO a un domaine appelé le domaine PIWI (dont les protéines Miwi et Mili chez la souris) qui est vital pour la spermatogenèse. Au lieu de s'associer avec le complexe RISC et les miRNAs, la famille de protéines PIWI s'associe aux piRNAs (impliqués dans la répression des éléments transposables par ARN interférence). Tandis que les miRNAs (environ 22 nucléotides) sont associés au complexe RISC, les piRNAs sont des ARNs plus longs (24-31 nucléotides) qui s'associent avec les protéines PIWIL (Piwi-like protein) 1, 2 et 4. Les miRNAs sont créés par une voie DICER-dépendante, alors que les piRNA sont produits indépendamment de DICER. Certaines protéines PIWI sont cruciales pour la méiose et la spermiogénèse (Carmell et al., 2007; Deng and Lin, 2002; Kuramochi-Miyagawa et al., 2004). Les protéines PIWL ont été montrées comme associées aux CBs et aux polysomes, ce qui indique leur rôle dans la régulation temporelle de la traduction (Grivna et al., 2006; Parvinen, 2005). L'exemple de Mili (PIWL2) et Miwi (PIWL1) est représentatif. Ces protéines sont exprimées séquentiellement avant (Mili) et après (Miwi) le stade spermatocyte pachytène et lient les précurseurs de piRNAs pendant le processus nucléolytique vers les piRNAs matures. Miwi en revanche, stabilise les ARNms spermiogéniques dans les RNPs réprimées du point de vue de la traduction, en se liant directement à eux, sans utiliser les piRNAs comme guides. La liaison directe de Miwi sur les ARNms spermiogéniques les protège des RNases jusqu'à ce que les ARNms soient prêts pour la traduction au cours des étapes les plus tardives de la spermiogénèse (Vourekas et al., 2012).

(b) Les ARE-BPs

Les protéines liant les éléments riches en A-U sont nombreuses, elles comprennent entre autres les protéines AUF1 et HUR (mentionnée plus haut) ayant un profil d'expression similaire, et CELF1 (pour revue, voir (Idler and Yan, 2012)).

(c) Les non-ARE-BPS

Les plus connues sont les protéines de la famille DAZ dont la fonction exacte n'est pas élucidée, la GRTH (Gonadotropin-regulated testicular helicase) ou DDX25, les protéines de la famille Nanos impliquées dans la répression de la transcription, et les protéines dites « mouse Y-box » généralement impliquées dans la stimulation de la transcription (pour revue, voir (Idler and Yan, 2012)). Les RBPs peuvent subir des modifications (phosphorylations, méthylation, acétylation, etc.) qui altèrent leur activité et leur localisation cellulaire, contrôlant ainsi leurs effets sur le devenir post transcriptionnel des ARNms.

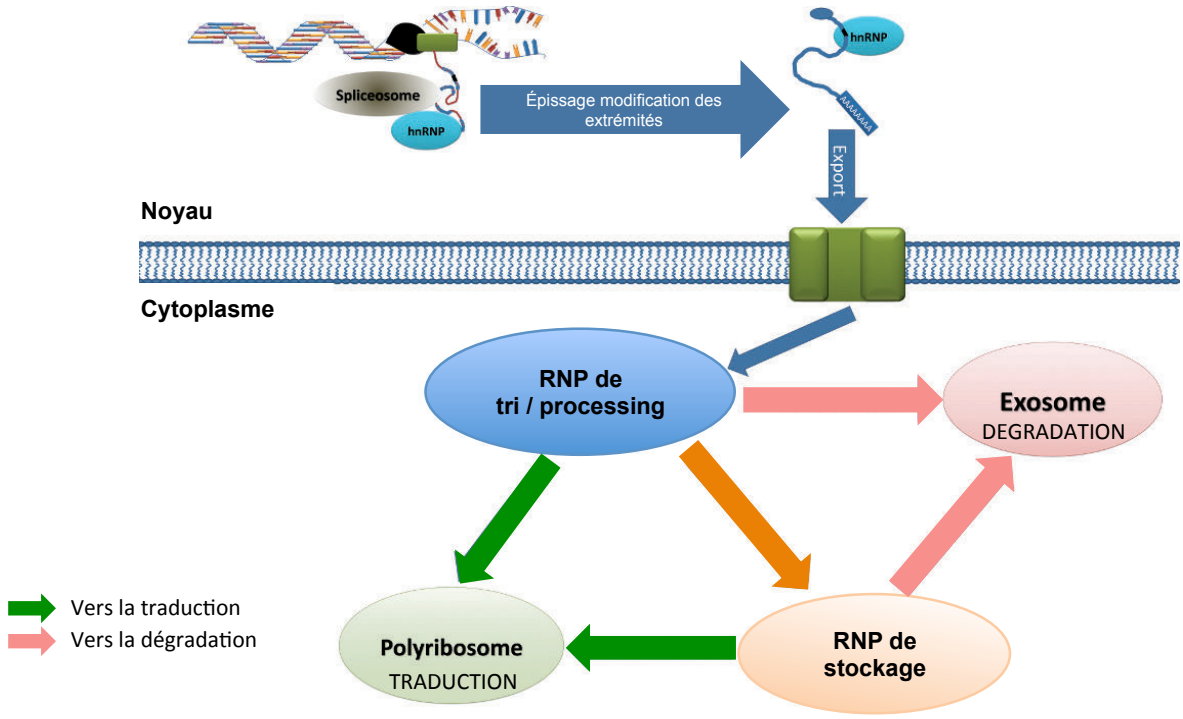

Figure 5. Vue d'ensemble des régulations post-transcriptionnelles des ARNms

(Adapté de Idler and Yan, 2012).

II. L'analyse protéomique

Le mot « protéome » a été inventé par Marc Wilkins lors du 4ème congrès de Sienne en 1994 et est une contraction des mots «protéine» et «génome». Le terme englobe la nature complexe et dynamique de l'expression des protéines de la cellule à l'organisme Considérant que les génomes sont essentiellement invariants dans différentes cellules d'un organisme, le protéome, ainsi que le transcriptome, varient de cellule en cellule, en fonction du temps et des stimuli et / ou de stress environnementaux. La protéomique est le domaine de recherche révélant la dynamique temporelle des protéines dans un compartiment biologique donné et à un moment donné. La définition a jusque récemment couvert les protéines en tant que produits des gènes. Récemment, en effet, la définition de la protéomique a été modifiée pour inclure non seulement des produits de l'expression des gènes, mais aussi la modification de structure et les modifications chimiques de ces produits de gènes, à savoir, les modifications post-traductionnelles (Aebersold et Mann, 2003). Cette définition pourrait bien encore changer, puisqu'il est désormais possible d'accéder à la dynamique de complexes protéiques et à leur mécanisme d'action en utilisant la spectrométrie de masse native (Sharon et al., 2007; Taverner et al., 2008). La protéomique comprend aujourd'hui des champs très divers et repose sur des approches et des stratégies variées, tant en amont qu'en aval. Parmi ces grandes « thématiques » il est possible de citer : 1) l'identification des protéines présentes dans un échantillon donné ; 2) la quantification des protéines en fonction du temps, de leur cellule d'origine ou de leur état ; 3) la localisation de protéines seules ou liées à d'autres molécules dans un organe ; 4) la caractérisation de modifications post-traductionnelles de protéines et leur changement sous certaines conditions ; 5) la caractérisation d'interactions entre certaines protéines; 6) la caractérisation de la structure spatiale des protéines ou de complexes de protéines et leur dynamique.

La protéomique est donc une approche de choix, car les protéines sont modifiées et maturées de façon non apparente sur la séquence du gène, et parce qu'elles reflètent l'état des systèmes biologiques étudiés. De plus, leur niveau d'expression n'est pas forcément accessible en transcriptomique, comme par exemple dans les spermatozoïdes, inactifs du point de vue de la transcription (Nynca et al., 2014).

A. De la cellule à la source du spectromètre

La préparation des échantillons et la séparation des protéines sont cruciaux en protéomique, car elles conditionnent l'obtention de résultats de spectrométrie de masse interprétables et cohérents avec la nature de l'échantillon. La préparation des protéines pour une séparation sur gel 1D ou 2D puis une analyse en spectrométrie de masse nécessite généralement

- l'extraction la plus large possible des protéines présentes dans l'échantillon de départ ;
- leur solubilisation et la prévention des phénomènes de précipitation ;
- l'élimination des substances pouvant interférer avec les protéines pendant leur séparation et leur analyse en MS : sels, lipides acides nucléiques, polysaccharides, détergents.

Selon la nature de l'échantillon, différents tampons d'extraction des protéines peuvent être utilisés avant la séparation des protéines par électrophorèse qui nécessite leur dénaturation. Les agents dénaturants peuvent être utilisés comme par exemple le SDS ou le déoxycholate pour défaire les liaisons polaires des protéines, en présence de sels pour rompre les liaisons électrostatiques. Des détergents sont utilisés pour éliminer les lipides qui peuvent se lier à des protéines et en altérer la migration que ce soit en fonction de leur point isoélectrique (IEF, Isoelectric focusing), ou de leur poids moléculaire. Les ponts disulfures doivent être réduits grâce à des agents réducteurs comme le DTT, et les liaisons hydrogènes rompues à l'aide des agents chaotropes (urée, thiourée) qui vont aussi solubiliser les segments hydrophobes des protéines (Rabilloud, 1996). En revanche, dans le cas d'analyses sans séparation des protéines par électrophorèse, il est indispensable d'éviter les détergents qui vont interférer avec la spectrométrie de masse et d'appliquer des procédés de dessalage par chromatographie (Cañas et al., 2007), ou d'adapter la préparation avec certains procédés de précipitation (Zhou et al., 2012). Il existe de nombreuses techniques de séparation des protéines, mais selon la nature des protéines d'intérêt, certaines techniques s'avèrent inadaptées. Par exemple, si les protéines sont plutôt hydrophobes, une séparation par fractionnement IEF n'est pas adaptée et il faut utiliser une séparation sans IEF. Cela même si une électrophorèse bidimensionnelle est envisagée. Le problème ne se pose pas pour les protéines hydrophiles. On peut optimiser la préparation en fonction de l'échantillon en utilisant une variété de détergents et chaotropes, par exemple pour l'isolement de protéines membranaires. Il n'existe pas de solutions universelles pour la préparation d'échantillons en protéomique. La préparation sera adaptée à

la technologie utilisée, s'il y a besoin d'un marquage des peptides ou des protéines, par exemple ou d'une purification de certaines protéines en particulier.

La réduction de la complexité des échantillons est un prérequis pour toute analyse protéomique qui consiste à identifier le plus de protéines possibles, ou de rester sensible dans la détection de protéines faiblement exprimées dans un échantillon biologique. En effet, une cellule peut exprimer plusieurs milliers de protéines différentes, à un nombre de copies très variable, dont certaines parfois exprimées en très faibles quantités par rapport aux protéines majoritaires et dont il faut pouvoir atténuer la supériorité en termes de nombre de copies. Pour remédier à ce problème, les approches sont déférentes selon les questions biologiques auxquelles les analyses protéomiques vont tenter de répondre. On peut par exemple procéder à un fractionnement subcellulaire et ne s'intéresser qu'à certains organites pour étudier un sous protéome, par exemple par centrifugations différentielles (Abdolzade-Bavil et al., 2004) ou sur gradient (Fialka et al., 1997). Un sous protéome ainsi constitué contiendra en proportion plus de protéines discrètes que l'échantillon complexe. L'élimination de protéines majoritaires peut être réalisée par immunodéplétion (Chromy et al., 2004) ou égalisation protéique (González-Iglesias et al., 2014). La séparation des peptides sans séparation des protéines est aussi utilisée en protéomique. Dans ce cas, la digestion des protéines se fait en solution sans séparation préalable, avant la séparation des peptides par plusieurs techniques, par exemple sur colonnes échangeuses d'ions ou par chromatographie liquide haute performance en phase inverse, avant l'analyse en MS ou MS/MS (spectrométrie de masse en tandem).

Les techniques de séparation des protéines et des peptides sont nombreuses et évoluent de manière constante. On peut donner l'exemple de l'utilisation de colonnes à phase monolithique pour chromatographie liquide (Guryca et al., 2008). Ces techniques de séparation peuvent être couplées pour améliorer la séparation des peptides. C'est le cas en protéomique Shotgun basée sur la séparation multidimensionnelle (Fournier et al., 2007), où des colonnes qui vont permettre de séparer les peptides en fonction de propriétés différentes (résine échangeuse d'ions, phase inverse) sont couplées avant la source du spectromètre de masse.

B. La spectrométrie de masse en protéomique

La spectrométrie de masse ou MS (Mass spectrometry) initialement utilisée pour identifier les protéines séparées par électrophorèse est devenue un outil d'analyse incontournable pour les approches protéomiques. L'essor de cette technologie est du à des développements fondamentaux et instrumentaux permettant des performances analytiques améliorées, en particulier dans le domaine de l'analyse les peptides (peptidomique) et des protéines (protéomique). L'apparition des sources d'Ionisation par Electronébulisation ou ESI (ElectroSPray Ionization) et des sources de Désorption/Ionisation Laser Assistée par Matrice (Matrix Assisted Laser Desorption Ionization, MALDI) vers la fin des années 80, ont révolutionné la MS et ses applications en sciences du vivant. La rapidité des technologies, l'amélioration des couplages avec la chromatographie liquide, ainsi que l'établissement des banques de données ont aussi contribué à l'essor de la spectrométrie de masse en protéomique. Elle est une méthode de choix dans l'analyse des protéines dans la mesure où elle peut nous fournir de l'information sur leur séquence. D'autre part, les instruments utilisés sont sensibles et permettent d'accéder de plus en plus facilement aux protéines discrètes d'un échantillon. De plus, les spectromètres de masse sont de plus en plus rapides.

a) Principe de la spectrométrie de masse

La spectrométrie de masse se fonde sur la détermination du rapport masse sur charge (m/z) des composés en phase gazeuse. Le principe de la mesure du rapport m/z repose sur la possibilité pour un flux d'ions: 1) de traverser un champ électrodynamique (Simple Quadripôle); 2) de décrire un mouvement périodique dans un champ électrostatique (Orbitrap); 3) de se séparer dans le temps par leur vitesse (TOF); 4) de se disperser en fonction de leur moment ou énergie cinétique (champ magnétique, FTICR). Cette détermination impose aux molécules d'un échantillon d'être préalablement ionisées et désolvatées, avant d'être dirigées vers l'analyseur. Un spectromètre de masse comporte les éléments suivants:

- Un moyen d'introduction de l'échantillon qui peut être un appareil chromatographique (injection) ou un pousse seringue (infusion) ;
- une source d'ionisation/désorption : ionise et transfère les composés en phase gazeuse ;

- un analyseur qui trie, isole, active et fragmente les ions en fonction de leur rapport m/z ;
- un détecteur qui permet une détection des ions préalablement triés et fournit un signal électrique proportionnel au nombre d'ions détectés ;
- Un système informatique d'acquisition des données.

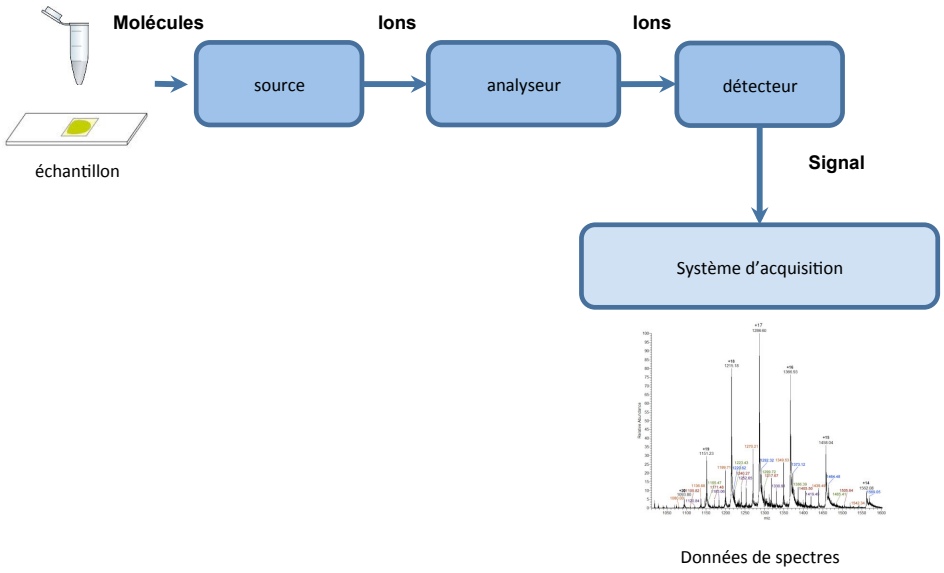

Figure 6. Les différents éléments composant un spectromètre de masse

b) Les différentes sources d'ions

Plusieurs sources d'ions sont utilisables pour l'analyse des protéines. Ici seront détaillées seulement la source MALDI et la source ESI.

(1) La source MALDI

Le principe du MALDI (Matrix Assisted Laser Desorption / Ionisation) fut mis au point par Michael Karas, Franz Hillenkamp et leurs collaborateurs (Karas et al., 1985). Il repose sur l'excitation d'un échantillon solide, dispersé dans une matrice cristalline (par exemple, le Alpha-Cyano-4-hydroxycinnamic acid (CHCA) et le 2,5-Dihydroxybenzoic acid (DHB) valables pour l'analyse des peptides et des protéines, mais aussi des phospho-peptides), par des photons issus d'un laser dont la longueur d'onde est située dans la bande d'absorption de la matrice. L'irradiation de la matrice et de l'échantillon provoque l'apparition

d'une grande quantité d'énergie dans la phase condensée, provenant de l'excitation des molécules de la matrice. Des ions, formés par transfert de protons ou d'électrons entre la matière photoexcitée et l'analyte, désorbent de la matrice (Figure 7). Cette source est utilisée avec des analyseurs de type TOF (Time of Flight) détaillés plus loin, dans lesquels les ions seront séparés selon leur rapport masse sur charge le long du tube de vol. Cette source est utilisable pour les composés de haut poids moléculaire. Cette source est utilisée en imagerie pour la recherche de biomarqueurs sur des coupes de tissus. En effet, une corrélation peut être obtenue entre l'histologie d'un tissu et son image MALDI, et cette dernière permet d'accéder à la localisation précise de protéines d'intérêt avec une résolution de 20µm (Lagarrigue et al., 2011). La résolution spatiale de l'imagerie MALDI peut désormais être de l'ordre du sub-cellulaire (1 à 5 microns) et même inférieure au micron (Zavalin et al., 2012).

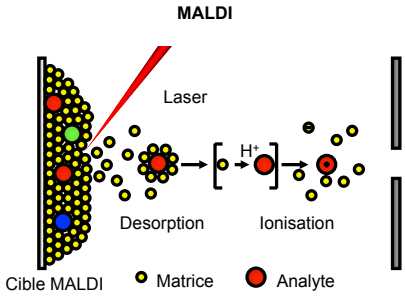

Figure 7. Principe de la source MALDI

(2) La source ESI

La source d'ionisation par électronébulisation, ou ESI (ElectroSpray Ionisation, (Whitehouse et al., 1985)), a été développée par J.B. Fenn (prix Nobel 2002) et a révolutionné la spectrométrie de masse, notamment dans le domaine de l'analyse de protéines (molécules non volatiles) et la protéomique. En effet, elle permet de transformer directement des molécules d'un échantillon en solution en ions en phase gazeuse. Le principe repose sur la formation, à pression atmosphérique, d'un spray de gouttelettes chargées qui s'évaporent et provoquent l'expulsion d'ions solvatés, qui seront ensuite libérés de leurs molécules de solvant en phase gazeuse (Figure 8). L'échantillon est infusé à débit constant à travers un capillaire, auquel est appliquée une différence de potentiel à l'aide d'une contre-électrode (skimmer). Ce capillaire est enchâssé dans un tube au travers duquel circule un débit constant

d'un gaz inerte, appelé gaz de nébulisation. Il se forme un spray par le cône de Taylor, de microgouttelettes chargées en suspension dans l'air à la sortie du capillaire. Durant le trajet de ces microgouttelettes vers le skimmer, le solvant qui les compose s'évapore. La densité de charge au sein de chaque gouttelette (1) augmente jusqu'à ce qu'elle devienne trop importante et qu'il se produise une explosion coulombienne, provoquant une libération des ions solvatés en phase gazeuse. Les agrégats chargés ainsi générés sont ensuite transférés pour être évaporés (2) avant d'atteindre la limite de Rayleigh provoquant une explosion coulombienne (3) et dirigés vers l'analyseur.

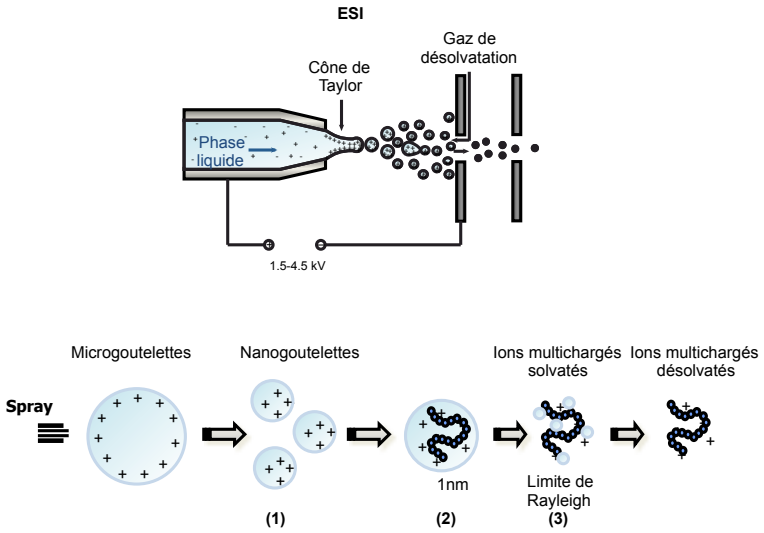

Figure 8. Principe de l'ionisation par éléctronébulisation

Pour les solvants en ESI, le meilleur compromis consiste en un mélange entre un solvant organique (température d'ébullition, viscosité et tension superficielle faibles) et de l'eau (constante diélectrique élevée). Ces ions sont générés par protonation ou déprotonation des molécules de l'échantillon. Le spectre de masse permet donc l'accès aux rapports m/z des ions moléculaires protonés ou déprotonés ([M+H]+ ou [M-H]-). La particularité de cette technique est qu'elle permet également de former des ions multichargés. Cette source est utilisée dans les approches de protéomiques dites « Shotgun » (Lee et al., 2004), et les approches shotgun itératives, c'est à dire qui consistent en la répétition de l'injection d'un même échantillon pour augmenter le nombre d'identifications de peptides, et donc de protéines, dans cet échantillon, (Lavigne et al., 2012).

c) Les analyseurs

Il existe plusieurs sortes d'analyseurs, ceux utilisés en routine au sein du laboratoire sont les analyseurs de type TOF (Time Of Flight), et l'analyseur de type Orbitrap. Les analyseurs TOF se basent sur le temps de vol des ions *via* l'utilisation d'un faisceau mono-énergétique créé par application d'une différence de potentiel à la sortie de la source d'ionisation. Cette tension confère une énergie cinétique au faisceau ionique homogénéisée à la valeur Vo. Le rapport m/z est déduit directement de la vitesse v, donc du temps (temps de vol) que met un ion à franchir une distance. Les ions volent d'autant plus vite qu'ils sont légers, et donc frappent le détecteur en premier, et les ions les plus lourds frappent le détecteur en dernier.

Les « Orbitrap » sont parties intégrantes de spectromètres de masse hybrides comme le LTQ-Orbitrap. Le système LTQ-Orbitrap est un spectromètre de masse à FT particulièrement adapté à la MS couplée à la chromatographie liquide en phase inverse RP-HPLC-MS/MS (Reversed phase liquid chromatography coupled with tandem mass spectrometry). Il permet d'analyser les peptides en MS en tandem (MS/MS). C'est un spectromètre de masse hybride qui consiste en un piège à ions linéaire couplé, *via* une C-trap, à une Orbitrap. Cette technologie permet d'obtenir des résolutions très élevées ainsi qu'une excellente précision de masse, sans sacrifier la sensibilité, et sur une grande gamme dynamique (Hu et al., 2005; Makarov, 2000; Makarov et al., 2006). Des résolutions de Rs = 60.000 à la masse m/z 400 en 1s, juqu'à Rs = 100.000 à la masse m/z 400 en 1,5s peuvent être obtenues avec le LTQ Orbitrap. Aujourd'hui des spectromètres de ce type offrent de plus en plus hautes résolutions jusqu'à 240.000 à m/z 400 en 768 ms pour l'Orbitrap Elite, soit 4 fois plus qu'une LTQ Orbitrap "classique", ainsi que des modes de fragmentation des ions CID (Collision induced dissociation) et HCD (Higher energy collisional dissociation) séquentielles et complémentaires, ameliorant les analyses MS/MS (Michalski et al., 2012). Cette technologie offre une grande sensibilité et un faible seuil de détection. Il est possible d'identifier de l'ordre de 100 attomoles (10^{-18} mole) de protéines digérées à la trypsine. La gamme de masse offerte par l'Orbitrap XL™ couvre de 15 à 4000 m/z.

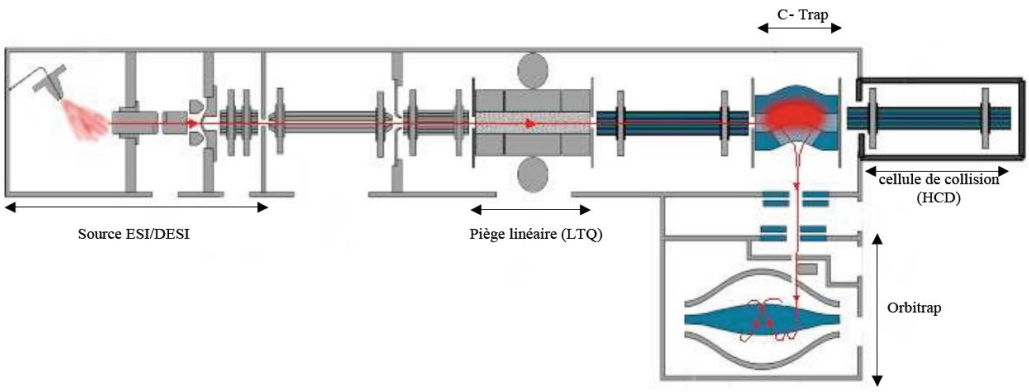

Figure 9. Schéma du LTQ-Orbitrap XL^{TM} de ThermoFisher avec le trajet des ions

Les ions générés par ESI sont collectés dans le LTQ, forme linéaire de piège à ions pouvant être employé comme un filtre de masse sélectif, ou comme un piège réel en créant un potentiel le long de l'axe des électrodes. Ces ions sont éjectés *via* un champ électrique quadripolaire au sein du LTQ, vers la C-trappe qui stocke les ions avant l'injection dans le piège orbital. La fragmentation des ions se fait dans le piège à ions linéaire collision, CID, mais peut aussi se faire dans une chambre de collision par dissociation par collision de haute énergie ou HCD. Les ions transférés depuis l'extrémité de la C-Trappe sont capturés dans le piège orbital par l'augmentation rapide du champ électrique. Chaque ion entrant dans l'analyseur Orbitrap aura à la fois un mouvement de rotation autour de l'électrode, et un mouvement axial le long de celle-ci. La fréquence des oscillations harmoniques selon l'axe de l'électrode central dépend du ratio masse / charge des ions (m/z). Ce mouvement axial génère un courant induit qui est enregistré par les deux moitiés externes de l'Orbitrap. Les signaux provenant de chaque extrémité du piège orbital sont amplifiés, et la fréquence axiale et donc le rapport m/z des ions déterminée par détection du courant induit et des calculs par algorithmes rapides de transformée de Fournier (FT), génèrent finalement un spectre de masse.

Le LTQ joue le rôle d'accumulateur d'ions dans cette MS en tandem, grâce à l'utilisation d'un vide très poussé et d'une injection rapide des ions depuis la C-Trap qui permet de réduire la dispersion des ions lors de l'injection = focalisation. Les ions sont donc stables pendant plusieurs secondes dans l'Orbitrap ce qui permet d'atteindre des résolutions très élevées (Rsmax > 100.000 à m/z 400), et une excellente précision de masse.

d) L'identification des protéines

Il existe trois façons d'identifier les protéines à partir de données spectrales, la recherche d'ions MS/MS, faisant partie de l'analyse LC-MS/MS sera davantage détaillée.

(1) Identification par empreinte peptidique (PMF, peptide mass fingerprint)

L'identification par PMF se fait à partir d'un ensemble de masses obtenues de la digestion enzymatique d'une protéine, formant « l'empreinte peptidique » de cette protéine, reconnue comme telle dans les bases de données. Ce type d'identifications a émergé d'un besoin d'une méthode rapide, efficace, pour identifier les protéines fréquemment observées dans les gels d'électrophorèse. L'idée est née en 1989, mais fut plutôt utilisée après la venue de l'instrumentation commerciale beaucoup plus sensible des spectromètres basés sur la technologie MALDI-TOF-MS (Henzel et al., 2003). Elle est simple, rapide et très sensible, mais nécessite la présence de la séquence de la protéine dans les bases de données UniProt et NCBInr, base de données de séquences protéiques de NCBI régulièrement mise à jour et téléchargeable pour Mascot, l'algorithme de recherche utilisé pour faire des identifications PMF (http://www.matrixscience.com). La séquence protéique ou son homologue proche doit être présente dans la base de données. Ce type d'identification ne convient pas aux mélanges complexes, particulièrement si l'on recherche des protéines peu abondantes.

(2) La recherche d'étiquettes de séquences (Sequence tag)

La première approche de recherche dans les bases de données à partir des données de spectres de fragmentation de peptides (spectre MS/MS) a été « l'étiquette de séquence » (Mann et Wilm, 1994). Quand la qualité d'un spectre MS/MS typique n'est pas assez bonne pour une interprétation de séquence *de novo*, il est souvent possible de lire séquentiellement trois ou quatre résidus de séquence facilement identifiables (tag). Par contre, même dans une petite base de données, cette courte séquence de quelques acides aminés peut se répéter de nombreuses fois. Toutefois un court tronçon de la séquence d'acides aminés pourrait fournir suffisamment de spécificité pour une identification sans ambiguïté si elle est combinée avec les valeurs de masse des ions fragments qui l'entourent, la masse du peptide, et la spécificité de l'enzyme. Cette méthode consiste donc à comparer les tags de séquence peptidiques aux

séquences protéiques dans les bases de données. Rapide, cette méthode permet d'identifier un peptide même avec une modification post-traductionnelles inconnue. Elle est donc tolérante aux erreurs, et requiert parfois une interprétation manuelle des spectres, elle n'est donc pas appropriée pour la protéomique à haut débit.

(3) Identification de protéines avec les données MS/MS

L'analyse des spectres MS/MS pour l'identification des protéines est typiquement utilisée en protéomique Shotgun. Comme nous le verrons, dans une analyse LC-MS/MS classique, un ion précurseur est sélectionné lors d'un scan MS puis fragmenté par CID ou HCD. Le spectre MS/MS de l'ion précurseur est interprété grâce aux intervalles entre les pics du spectre. Il est donc possible de retrouver la séquence du précurseur à l'aide de ce spectre et du rapport m/z du précurseur. La limite de l'interprétation est ainsi liée à la qualité des spectres. Pour chaque spectre MS/MS, un logiciel est utilisé pour déterminer quelle séquence de peptide dans une base de données de protéines ou de séquences d'acides nucléiques (traduites en acides aminés dans les six cadres de lecture possibles en fonction de la taxonomie selon laquelle le code génétique peut changer), donne la meilleure correspondance. Chaque « entrée » dans la base de données choisie est digérée in silico en utilisant la spécificité connue de l'enzyme choisie pour l'analyse, et les masses des peptides intacts sont calculées. Si la masse théorique calculée d'un peptide correspond à celle d'un peptide observé dans l'analyse, les masses des ions fragments attendus sont aussi calculées et comparées avec les valeurs expérimentales (Cottrell, 2011). De nombreux algorithmes de « scoring » comme ceux utilisés par Mascot (Perkins et al., 1999) ont été conçus pour décider quelle séquence peptidique correspond le mieux à un spectre donné. Compte tenu que les données de chaque spectre MS/MS correspondent à un peptide isolé, cela n'a aucune importance si l'échantillon original était une seule protéine ou alors un mélange. Des séquences peptidiques sont identifiées dans ce mélange, alors l'ensemble des séquences de peptides est utilisé pour déduire des protéines qui sont probablement présentes.

(a) Taxonomie

Si une base de données contient des informations concernant la taxonomie, la plupart des moteurs de recherche peuvent l'utiliser pour limiter la recherche aux entrées d'un organisme particulier ou au rang taxonomique. En diminuant la taille de la base de données, cela accélère la recherche. La limitation de la taxonomie simplifie également le résultat, car elle

élimine les protéines homologues d'autres espèces. Dépendamment de l'étude, il n'est pourtant pas toujours approprié de spécifier une taxonomie trop étroite dans la recherche. En effet, si la protéine correcte de la bonne espèce n'est pas présente dans la base de données, il peut s'avérer plus qu'utile de voir une correspondance avec une protéine homologue dans une autre espèce. Ceci est particulièrement vrai pour les espèces peu représentées.

(b) Enzyme et modifications

Les différentes modifications: fixes (carbamidométhylation des cystéines due à la préparation de l'échantillon) et variables (oxydations de méthionines, acétylation des lysines, phospohorylations des sérines) sur les résidus d'acides aminés doivent être précisément rentrées dans les paramètres de recherche pour l'assignement correct des spectres MS/MS aux masses théoriques correspondantes. Une autre classe de modifications utilisée pour le marquage de masse isotopique stable de protéines ou de peptides dans une expérience de quantification, doit être rentrée dans les paramètres de recherche. Certains peptides porteront une étiquette légère voire pas d'étiquette, et d'autres peptides porteront une étiquette lourde, mais aucun ne pourront à la fois être marqués « léger » et « lourd ». Ces étiquettes de masse peuvent être considérées comme deux modifications fixes distinctes, une pour chacune des étiquettes légère et lourde. La spécificité de l'enzyme utilisée doit être précisément rentrée également dans les paramètres de recherche. Des coupures manquantes par l'enzyme peuvent être autorisées.

(c) Le « scoring » des peptides

Beaucoup d'algorithmes ont été développés pour évaluer le score de la correspondance des peptides. Ils sont basés sur un système de probabilité, la recherche dans une base de données étant un processus statistique. La plupart des spectres MS/MS ne codent pas la séquence peptidique complète, il y a des lacunes et des ambiguïtés. Heureusement, la plupart du temps, la correspondance correcte peut être signalée c'est à dire un « vrai positif », mais pas toujours. Si la séquence du peptide n'est pas dans la base de données, et que l'on obtient une correspondance en dessous du score ou le seuil de significativité, nous avons un « vrai négatif». Un «faux positif» correspond à une correspondance correcte sur une mauvaise séquence. Un «faux négatif» arrive lorsque une correspondance n'est pas signalée, alors que la séquence correcte existe dans la base de données. La manière habituelle de mesurer la qualité d'un ensemble de résultats de recherche s'appuie sur l'utilisation du taux de faux

positifs (Nesvizhskii et al., 2007). Beaucoup de moteurs de recherche utilisent des E-values (expected values) à la place ou en plus des scores. Une E-value correspond au nombre de fois que l'on s'attend à obtenir un score au moins aussi élevé, par chance. Les petites E-values sont bonnes, et une correspondance avec une E-value de 1 ou plus indique une correspondance aléatoire.

(d) La recherche « cible-leurre »

Pour les études à grande échelle, il est nécessaire d'estimer le taux de faux positifs des correspondances des spectres MS/MS. L'un des moyens les plus fiables pour le faire repose sur une recherche appelée « cible-leurre » ou « target-decoy search ». C'est une façon très simple mais puissante de valider les résultats de recherche. La recherche est répétée, en utilisant des paramètres de recherche identiques, contre une base de données dans laquelle les séquences ont été inversées ou randomisées. Le nombre de correspondances sur des séquences de la base leurre est une excellente estimation du nombre de faux positifs dans les résultats de la base de données cible (Elias et Gygi, 2007). Les bases cibles et leurre peuvent être, ou pas, concaténées c'est à dire réunies dans une même base. L'important est d'estimer le taux de faux positifs par une recherche sur la base leurre.

(e) L'inférence des protéines

La correspondance des spectres MS/MS dans la base de données identifie des peptides, et non des protéines. Déduire les séquences peptidiques des protéines qui étaient présentes dans l'échantillon d'origine est particulièrement difficile car un grand nombre des séquences de peptides lors d'une recherche typique peut être assigné à plusieurs protéines. Le principe de l'inférence de protéines est de créer une liste minimale de protéines à partir des peptides identifiés. Autrement dit, le nombre minimal de protéines qui peuvent expliquer les peptides observés. Certaines personnes appellent cette approche le principe de parcimonie ou rasoir d'Occam (Nesvizhskii et Aebersold, 2005). Le taux de faux positifs des protéines n'est pas le même que le taux de faux positifs des peptides. Il peut être supérieur ou inférieur, selon les règles d'acceptation d'une protéine ou d'une famille de protéines. Il est généralement conseillé d'exiger qu'une protéine soit inférée de plus d'une séquence peptidique distincte. Une protéine avec seulement une séquence peptidique unique est souvent considérée comme suspecte. C'est en fait une légère sur-simplification. Dans une recherche avec un grand nombre de spectres et une petite base de données, même si le FDR de peptide est faible, une

protéine peut rassembler plusieurs fausses correspondances par hasard. Il faut alors regarder si les règles d'acceptation d'une protéine dans la base de données cible ne donnent pas de protéines fausses positives de la base leurre.

C. Approches de protéomique

a) La protéomique Shotgun

La protéomique dite Shotgun est appelée ainsi en comparaison à un « fusil de chasse « qui tire presque au hasard pour analyser par spectrométrie de masse en tandem des peptides générés par la digestion enzymatique, le plus souvent par la trypsine, de mélanges de protéines complexes. La protéomique Shotgun s'inscrit dans la protéomique de découverte tout comme la 2-DE-MS, et s'oppose à la protéomique ciblée, qui cherche à analyser des protéines connues dans un ou plusieurs échantillons, comme par exemple par SRM (selected reaction Monitoring) qui est apparue récemment en complément des approches de type shotgun. La protéomique ciblée par SRM, hautement reproductible et permettant une grande précision dans la quantification, est utilisée pour quantifier un ensemble préétabli de protéines parmi plusieurs échantillons. Elle sert notamment au monitoring de biomarqueurs et à leur validation (Hüttenhain et al., 2012).

La protéomique Shotgun repose sur le préfractionnement des protéines d'un échantillon complexe, puis l'analyse par spectrométrie de masse des protéines digérées présentes dans les fractions. Ce sont en effet les peptides enzymatiques présents dans chaque fraction qui sont séparés par chromatographie puis analysés par spectrométrie de masse en tandem, car la spectrométrie de masse des peptides est plus sensible que la spectrométrie de masse de protéines entières pour l'identification des protéines. On pourrait dire que la masse détectée par MS de gros composés est moins précise que la masse de petits composés à cause de la dispersion des signaux de masse des plus gros composés. Si la résolution de l'analyseur st faible, la masse mesurée n'est en effet pas la masse monoisotopique (celle du premier pic du profil isotopique c'est-à-dire celle qui ne prend en compte que les masses des isotopes les plus stables : ^{12}C, ^{1}H, ^{16}O, ^{32}S, ^{14}N, ..., mais la masse moyenne prenant en compte les différents isotopes stables naturels, c'est à dire le centroïde des masses des pics constituant le profil ou cluster isotopique. Or, la distribution isotopique augmente avec la masse du

composé, induisant une dispersion des signaux de masse détectés. Ceci diminue la précision de la masse moyenne détectée pour des gros composés de 3000-4000Da (Yergey et al., 1983). Cependant, la haute résolution des analyseurs (l'Orbitrap peut permettre une précision de masse de l'ordre du ppm pour 100.000 de résolution) conduit à la distinction des pics du profil isotopique et permet une précision de masse plus importante. D'autre part, la technique d'ionisation par électronébulisation (ESI) offre la possibilité d'analyser des composés de plusieurs centaines a plusieurs millions de Da (Fenn et al., 1989; Sanglier et al., 2003). En effet, l'ESI permet de générer des ions multichargés et donc de calculer la masse à partir des m/z de la série d'ions à différents états de charges pour un même composé, sans la nécessité de disposer d'un analyseur à gamme de balayage m/z élevée.

Le fait que la digestion des protéines facilite l'identification des protéines n'est pas tant dû à la sensibilité de détection des peptides par rapport à celle des protéines, mais parce que dans un mélange complexe l'hétérogénéité des protéines fait que de nombreuses protéines ayant des séquences différentes peuvent avoir la même masse, et l'identification protéique peut ne pas être univoque. La masse de la protéine intacte est donc par elle-même insuffisante pour son identification. La fragmentation des protéines en peptides trypsiques, ions parents eux même refragmentés en ions fils lors de la spectrométrie de masse en tandem permet d'obtenir un maximum d'informations sur la séquence de ces derniers. En effet, la spécificité de la trypsine qui coupe en C-ter des acides aminés basiques Lysine ou arginine sauf quand ils ont suivis d'une proline, et la longueur moyenne des peptides générés de 10 résidus d'acide aminés, donnent accès grâce à une succession d'ions fragmentés à l'information de toute leur séquence en plus de l'information de leur masse. Ceci donne une information structurale de la protéine et permet par corrélation des séquences peptidiques identifiées aux séquences théoriques d'accéder à une certaine couverture de séquence de la protéine et de l'identifier. La spectrométrie de masse en tandem est donc une méthode destructive permettant de retrouver la séquence des peptides à partir de leurs fragments, puis de la protéine par inférence à partir de ces séquences peptidiques. Cette méthode d'identification du « bas vers le haut » est appelée protéomique « Bottom up ».

La spectrométrie de masse de protéines entières peut cependant permettre d'identifier précisément des protéines grâce à la protéomique « Top down », méthode « du haut vers le bas » selon laquelle les protéines intactes sont directement ionisées puis fragmentées. Dans ce cas, les ions MS/MS générés à partir de la protéine permettent d'obtenir une grande couverture de séquence (jusqu'à 100%), et d'accéder aux modifications post-traductionnelles

ainsi qu'aux différents protéoformes (Siuti et Kelleher, 2007). La caractérisation de protéines complètes par Top down convient pour l'analyse des protéines simples ou des mélanges simples d'intérêt biologique, mais pose des difficultés techniques en termes de couverture du protéome, de sensibilité et de débit comparé à l'approche Bottum up. Cependant, les progrès récents en termes de séparation, d'instrumentation et les outils bioinformatiques ont propulsé l'approche Top Down comme un complément puissant et peut-être une alternative viable aux approches fondées sur la digestion (Catherman et al., 2014). Les stratégies Top Down et Bottum up n'ont pas les mêmes applications, et pour des mélanges complexes de protéines, une stratégie de type Bottum up est classiquement utilisée.

L'analyse de type Shotgun s'appuie sur le couplage de la nano chromatographie liquide directement couplée à la spectrométrie de masse. Au cours d'une analyse MS/MS, les ions sont analysés dans un premier analyseur, puis fragmentés dans une chambre de collision et les fragments sont envoyés vers un second analyseur de masse. Les analyseurs ne sont pas nécessairement du même type. Dans le LTQ-Orbitrap, ces deux analyseurs sont la trappe linéaire, puis l'Orbitrap. En protéomique Shotgun, les analyseurs de type LTQ-Orbitrap sont classiquement utilisés. L'analyse MS/MS se déroule en cinq étapes pour l'analyse des protéines (Figure 10). Au cours de la première étape, les protéines à analyser sont isolées à partir de lysats de cellules ou obtenues à partir de tissus par fractionnement biochimique ou sélection par affinité. Cela comprend souvent une étape finale d'électrophorèse sur gel unidimensionnelle, qui définit le sous-protéome à analyser **(étape 1).** Les protéines sont dégradées par une enzyme **(étape 2),** le plus souvent par la trypsine, ce qui génère des peptides avec des acides aminés C- terminaux protonés, fournissant un avantage dans le séquençage peptidique ultérieur. Cette stratégie est appelée « bottum up » par opposition à une stratégie « top down » évoquée plus haut, car elle consiste à identifier les peptides trypsiques afin de caractériser la protéine complète par corrélation de séquences. Dans une troisième étape **(étape 3)**, pour avoir le maximum d'identifications de cet échantillon complexe, les peptides sont séparés par une ou plusieurs étapes de chromatographie en phase liquide (phase inverse ou échange d'ions) à haute pression dans les capillaires très fins. Ils sont ensuite élués dans une source d'ions électrospray (Figure 8) où ils sont nébulisés dans de petites gouttelettes hautement chargées. Après évaporation, les peptides multi-protonés entrent dans le spectromètre de masse et, dans une quatrième étape **(étape 4)**, un spectre de masse des peptides élués à un temps de rétention donné est acquis (spectre MS, ou spectre de masse normal). L'ordinateur génère une liste prioritaire de ces peptides pour la fragmentation,

et une série d'analyses MS/MS s'ensuit **(étape 5)**. Cela consiste à isoler un peptide donné, à le fragmenter par choc énergique avec du gaz, et à enregistrer le spectre de masse en tandem ou spectre MS/MS. Les spectres MS et MS/MS sont stockés pour interroger les bases de données de séquences de protéines afin de chercher les correspondances avec les séquences peptidiques théoriques présentes dans les bases de séquence. Le résultat de l'expérience est la séquence des peptides identifiés aux différents temps de l'analyse, puis l'identification des protéines présentes dans la population de protéines « purifiées », par regroupement de l'information de séquence des différents peptides identifiés, et qui sont assignées à ces protéines.

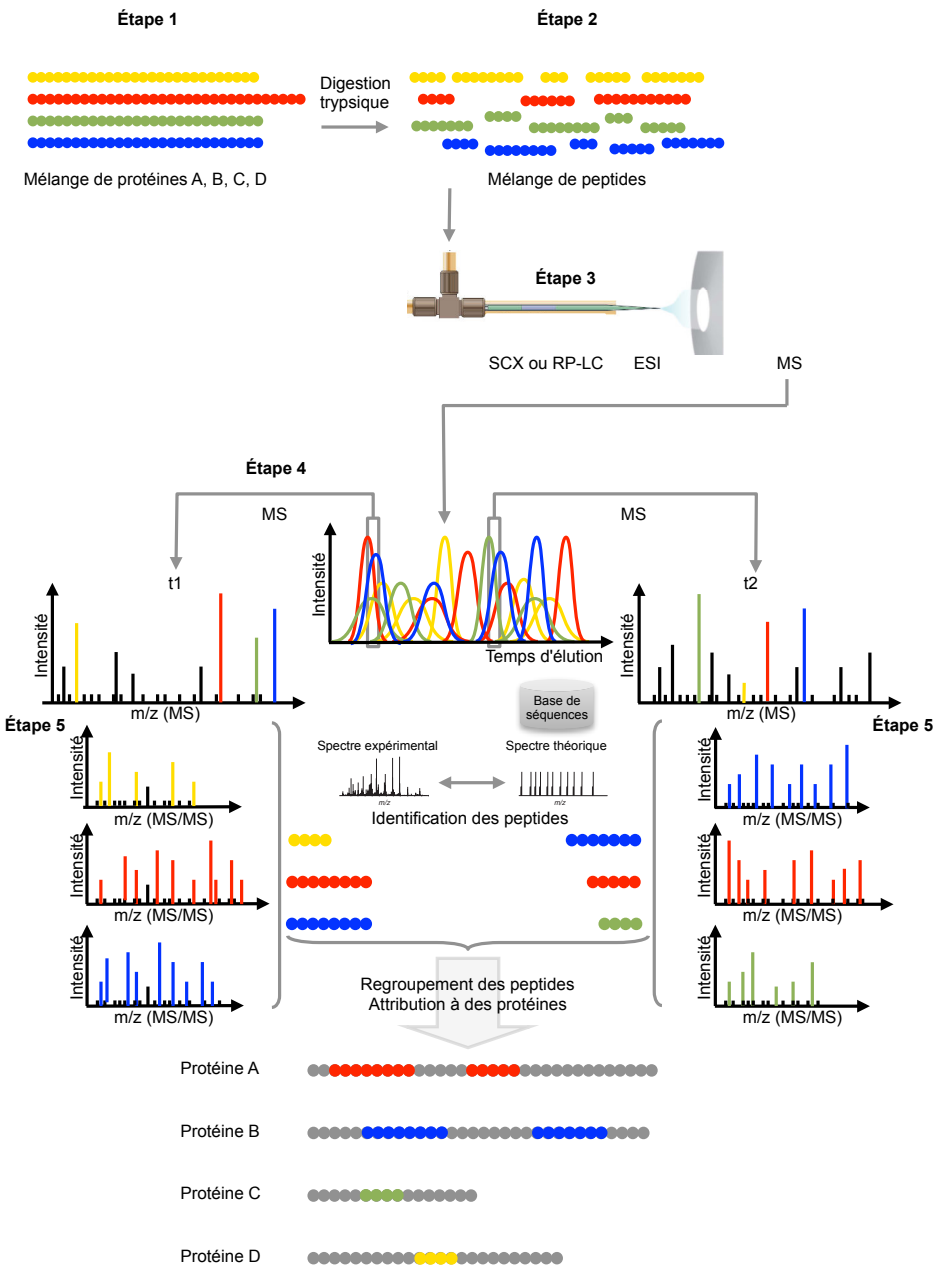

Figure 10. Expérience de LC-MS/MS générique en protéomique Shotgun

Cette expérience est utilisée pour des mélanges très complexes afin d'identifier un maximum de protéines. Les protéines sont digérées par la trypsine généralement, les peptides sont séparés par chromatographie liquide sur colonne échangeuse d'ions (SCX, strong cation exchange), ou phase inverse (RP, reverse phase) couplée à un spectromètre de masse. Dès leur élution de la colonne, ils sont nébulisés par une source ESI et passent dans le spectromètre de masse. A chaque temps d'acquisition, un spectre ou scan MS est réalisé et la masse des peptides élués à ce temps de rétention est mesurée. Les peptides les plus intenses sont fragmentés pour obtenir un spectre MS/MS par peptide sélectionné. Les masses des peptides et de leurs fragments correspondants obtenus tout au long de l'analyse sont confrontées aux masses théoriques des peptides trypsiques des protéines contenues dans les bases de données. Ceci permet d'une part d'identifier par information de séquence les peptides à différents temps de l'analyse et d'identifier les protéines dont ils sont issus, en regroupant ces informations de séquence.

b) Approches de protéomique différentielle

La protéomique différentielle consiste à comparer l'expression différentielle des protéines en mélanges complexes dans différents échantillons. Il existe plusieurs approches d'analyse différentielle s'appuyant sur l'utilisation de gels 2D, ou directement sur la spectrométrie de masse. Dans ce second cas, on peut séparer deux types d'approches : les approches de quantification relative des protéines dans différents échantillons qui utilisent des techniques de marquage isotopique de masse, et les approches de quantification des protéines sans marquage dites Label-free. Il existe également des techniques de quantification absolue qui sont davantage destinées aux analyses en protéomique ciblée plutôt qu'en protéomique exploratoire, et qui ne seront pas développées ici. On peut tout de même mentionner parmi elles le QconCat (Quantification conCATemer), (Austin et al., 2012); les peptides AQUA (Absolute QUAntification), (Warnken et al., 2013); les PSAQ (Protein Standard for Absolute Quantification), (Dupuis et al., 2008). Ces approches de quantification absolue permettent de quantifier précisément quelques protéines ciblées sur un grand nombre d'échantillons ; alors que la quantification relative permet de comparer des protéomes « entiers ». Les approches dont il sera question dans les deux prochaines sections font largement appel à la protéomique Shotgun expliquée plus haut, et s'appuient sur la spectrométrie de masse en tandem pour la quantification et l'identification des peptides.

L'électrophorèse bidimensionnelle, ou 2DE pour two dimensional electrophoresis dont le principe est étendu à la protéomique différentielle a longtemps été considérée comme l'outil de référence pour les analyses protéomiques, et est encore beaucoup utilisée aujourd'hui. Cette technique a pour principe la séparation des protéines selon leur point isoélectrique sur un gradient de pH dans la première dimension, puis selon leur taille par électrophorèse sur gel de polyacrylamide en présence de SDS dans la seconde dimension (Emes et al., 1975; Klose, 1975; O'Farrell, 1975). Par cette double séparation, une cartographie des protéines d'un échantillon peut être obtenue, ce qui a permis de référencer les données de protéomique 2DE en associant les identifications de protéines à leurs coordonnées sur les gels bidimensionnels, comme dans la banque SWISS-2DPAGE (Hoogland et al., 2004). Sur les gels bidimensionnels, les spots protéiques sont détectés par différentes méthodes de coloration, notamment le bleu de Coomassie ou le nitrate d'argent (coloration plus sensible), ou bien par des fluorochromes de type Cyanines. Les spots d'intérêt mis en évidence par analyse d'image sont excisés des gels bidimensionnels, et les protéines sont identifiées par spectrométrie de masse comme l'empreinte peptidique (PMF) après digestion à la trypsine.

On peut distinguer la comparaison de l'intensité des spots protéiques entre différents gels par 2DE, et l'approche 2D-DIGE ou DIGE (Two-Dimensional Difference Gel Electrophoresis), (Unlü et al., 1997)), qui permet de quantifier les protéines présentes dans différents échantillons. La comparaison d'échantillons par 2DE a longtemps été réalisée par analyse d'images de gels contre-colorés au bleu de Coomassie ou au nitrate d'argent pour mettre en évidence des différences apparentes d'abondance entre spots protéiques. Cette approche ne permet en revanche pas de quantification, étant donné le manque de linéarité entre la quantité des protéines et l'intensité de la coloration. La DIGE, grâce à l'utilisation de 3 fluorochromes aux spectres d'émission distincts permet, elle, la co-migration sur un même gel de 2 échantillons et d'un standard interne, autorisant alors la comparaison de plusieurs gels entre eux, et donc de quantifier les protéines par les différences d'abondance relatives des spots protéiques entre ces échantillons (Alban et al., 2003). La DIGE est utilisée dans de nombreux domaines de la biologie. Cette technique a été utilisée dans notre laboratoire pour établir le protéome différentiel des cellules germinales mâles chez le rat (Rolland et al., 2007).

Depuis plusieurs années, de nombreuses autres techniques d'analyse différentielle basées sur la spectrométrie de masse ont émergé. Beaucoup utilisent la protéomique Shotgun afin de quantifier et identifier un grand nombre de protéines dans les échantillons à comparer. La spectrométrie de masse n'est pas une technique quantitative en tant que telle, car différents peptides d'une même protéine ont des intensités (signal en MS) différentes. Cependant une quantification relative peut être mise en place, basée sur la comparaison de mêmes signaux peptidiques initiaux entre différents échantillons, traités et acquis dans des conditions similaires. L'avantage de ces techniques par rapport aux approches basées sur les gels 2D est que l'identification et la quantification se font parallèlement.

(1) Les approches de quantification par marquage isotopique

Le principe général de cette méthode de quantification relative est de marquer des échantillons à comparer avec des composés différentiables en spectrométrie de masse et de mêmes propriétés physico-chimiques. On utilise majoritairement les isotopes stables (^{2}H, ^{13}C, ^{15}N, ^{18}O) pour conserver strictement les mêmes propriétés en chromatographie liquide et spectrométrie de masse. Après marquage par différentes techniques, les échantillons sont mélangés et analysés par MS où l'on va rechercher des pics présentant un écart de masse

précis correspondant à la différence de masse entre les isotopes utilisés pour le marquage. L'intensité des pics est mesurée et le ratio entre les aires des deux pics d'intérêt reflète la différence d'expression de la protéine entre les deux échantillons. Quelques techniques de quantification basées sur la MS sont commentées ci dessous, mais cette liste n'est pas exhaustive.

Parmi les différentes méthodes utilisées, on peut en distinguer deux grands types. Celles qui consistent à marquer les protéines entières comme le SILAC (Stable-isotope labelling by amino acids in cell culture) qui est un marquage métabolique, l'iCAT (Isotope coded affinity tag) ou l'ICPL (Isotope-coded protein label) qui sont des marquages chimiques ; et celles avec lesquelles les peptides issus de la digestion enzymatique des protéines sont marqués comme l'iTRAQ, marquage chimique, et la TMT (Tandem Mass Tag). La méthode SILAC (Ong et al., 2002) permet de marquer les protéines dès la culture des cellules. Ces dernières sont cultivées en présence d'acides aminés marqués ou non par des isotopes stables (lysine et/ou arginine) jusqu'à incorporation des acides aminés alourdis. Après extraction protéique, les protéines des différents échantillons marqués sont mélangées et peuvent être fractionnées avant la digestion trypsique et l'analyse par LC-MS/MS. Les peptides alourdis ou non sont distingués par une différence de masse précise en mode MS et identifiés en mode MS/MS. Le problème de cette méthode est qu'elle ne s'applique qu'aux cellules qui peuvent être cultivées, elle est pourtant très utilisée dans des systèmes cellulaires variés. L'approche SILAC est beaucoup utilisée en couplage avec d'autres techniques de purification, comme par exemple avec la purification d'affinité: (Quantitative affinity purification MS). Cette méthode est utilisée par exemple pour caractériser des complexes protéiques chez la levure (Piechura et al., 2012). Les auteurs qualifient le SILAC de méthode la plus adaptée pour quantifier les protéines de cellules en culture. Elle est utilisée pour l'identification de modifications post-traductionnelles dynamiques importantes du point de vue fonctionnel, comme les phosphorylations, méthylations, acétylations, ubiquitinations (Størvold et al., 2013). L'ICAT (Gygi et al., 1999) est un marquage chimique des cystéines des protéines par un composé léger ou lourd ayant une différence de masse de 8Da, grâce à un groupement qui réagit avec les thiols. Les échantillons à comparer sont mélangés et peuvent être fractionnés avant digestion. Les peptides marqués sont enrichis sur colonne d'affinité par interaction avidine/biotine grâce à la biotine des composés ICAT, puis séparés par LC-MS/MS. Les peptides lourds et légers qui peuvent être distingués par une différence de masse précise de 8Da seront quantifiés en mode MS et identifiés en mode MS/MS. L'ICAT permet

uniquement la comparaison de deux échantillons en parallèle. L'ICPL est un marquage des protéines sur leurs groupements amine-libre (lysines ou en N-terminal). Le composé ICPL est un dérivé de l'acide nicotinique où 6 carbones peuvent être remplacés par des ^{13}C et/ou liés à des deutérium (Schmidt et al., 2005). Les protéines marquées sont digérées, et séparées sur gel avant analyse LC-MS. Les peptides séparés marqués par des étiquettes ICPL sont distingués par une différence de masse précise qui dépend des étiquettes utilisées. Ce marquage est en effet réalisable en duplex, triplex ou quadruplex. Ils sont quantifiés en mode MS et identifiés en mode MS/MS. L'ICPL permet de comparer jusqu'à 4 échantillons complexes en parallèle (Lottspeich et Kellermann, 2011). L'iTRAQ (Serada et Naka, 2014) est un marquage chimique des amines libres des peptides par un tag isobarique et se fait donc après la digestion des protéines des différents échantillons. Le composé iTRAQ est isobarique, c'est-à-dire qu'il possède la même masse (145 Da) quels que soient les isotopes stables incorporés. En effet, les tags possèdent un groupe de quantification « reporter group » et un groupe « balance » entre lesquels sont répartis différemment les isotopes selon les tags, sans que la masse totale de ces deux groupes ne varie. Un autre groupe réagit avec les amines N-terminales et des lysines des peptides. Après mélange des peptides marqués des échantillons, ceux-ci sont analysés par LC-MS/MS. Le tag étant isobarique, un seul m/z pour le même peptide marqué sera détecté dans tous les échantillons. La quantification relative se fait en même temps que l'identification en MS/MS, d'une part en comparant l'intensité des ions rapporteurs et d'autre part en analysant les ions de séquence. Jusqu'à 8 échantillons peuvent être analysés en parallèle. Le TMT (Tandem Mass Tag, (Thompson et al., 2003)) qui est aussi un marquage des amines libres est basé sur le même principe que L'iTRAQ : chaque réactif isobarique contient un nombre différent d'isotopes lourds dans le groupe rapporteur et la quantification se fait par MS/MS. La TMT permet de comparer jusqu'à 10 échantillons.

Dans ces analyses de quantification relative, la normalisation par un run LC-MS de référence a toute son importance pour la suite de l'analyse différentielle. La quantification peut s'effectuer en faisant des tests statistiques par peptides marqués différentiellement dans chaque échantillon. Les tests statistiques peuvent aussi être effectués sur les protéines en prenant en compte leurs différents peptides marqués. Dans ce cas, chaque peptide est une variable qui représente la protéine, il existe différents modèles plus ou moins compliqués en particulier pour la prise en compte des peptides en commun à plusieurs protéines. Une valeur de quantification unique pour chaque protéine peut être calculée à partir de la somme ou la moyenne de ses peptides. Il s'agit des ratios obtenus à partir des aires de peptides marqués,

ou des 3 peptides marqués les plus abondants, ou des peptides prototypiques marqués, dépendamment du type d'analyse différentielle utilisée.

(2) Les approches « label free »

Les approches label free ne font appel à aucun marquage des peptides ou des protéines comme leur nom l'indique, mais se basent soit sur les attributs des pics MS des peptides (intensité ou aire, hauteur), soit sur le nombre de spectres MS/MS acquis pour une protéine, soit sur les deux types d'informations combinées (Dicker et al., 2010). La méthode du « Top 3 » s'appuie sur la moyenne des intensités (signal MS) des trois ions trypsiques les plus intenses pour une protéine, qui serait proportionnelle à la quantité de cette protéine dans une condition (Silva et al., 2006). Cette technique est aussi utilisée pour la quantification absolue. L'utilisation du Top 3 implique la maîtrise des paramètres de la reproductibilité: variations de temps de rétention d'un peptide entre les différentes conditions, stabilité du spray, problèmes de contaminants. Il est nécessaire d'aligner les temps de rétention sur la colonne pour retrouver les peptides entre différentes expériences. Une normalisation des intensités doit être réalisée. Enfin, une normalisation du bruit de fond est indispensable. L'utilisation d'un spectromètre de masse à très haute résolution est indispensable pour cette approche. Les approches « peptide count » ou « spectral count » se basent quant à elles sur le nombre de spectres MS/MS acquis pour une protéine, qui serait proportionnel à la quantité de cette protéine tout en tenant compte de la détectabilité des peptides (Lee et al., 2011).

Ces approches « label free » sont donc moins précises que les techniques utilisant un marquage de masse, en revanche elles permettent de quantifier une proportion plus importante de protéines d'un mélange complexe comparé à l'ICPL par exemple, qui ne marque que les Lysines, dont l'abondance est de 6%, et les N-termini des protéines qui peuvent échapper au marquage. Elles ne sont pas pertinentes pour les petites protéines ou les protéines peu abondantes (Zhou et al., 2010).

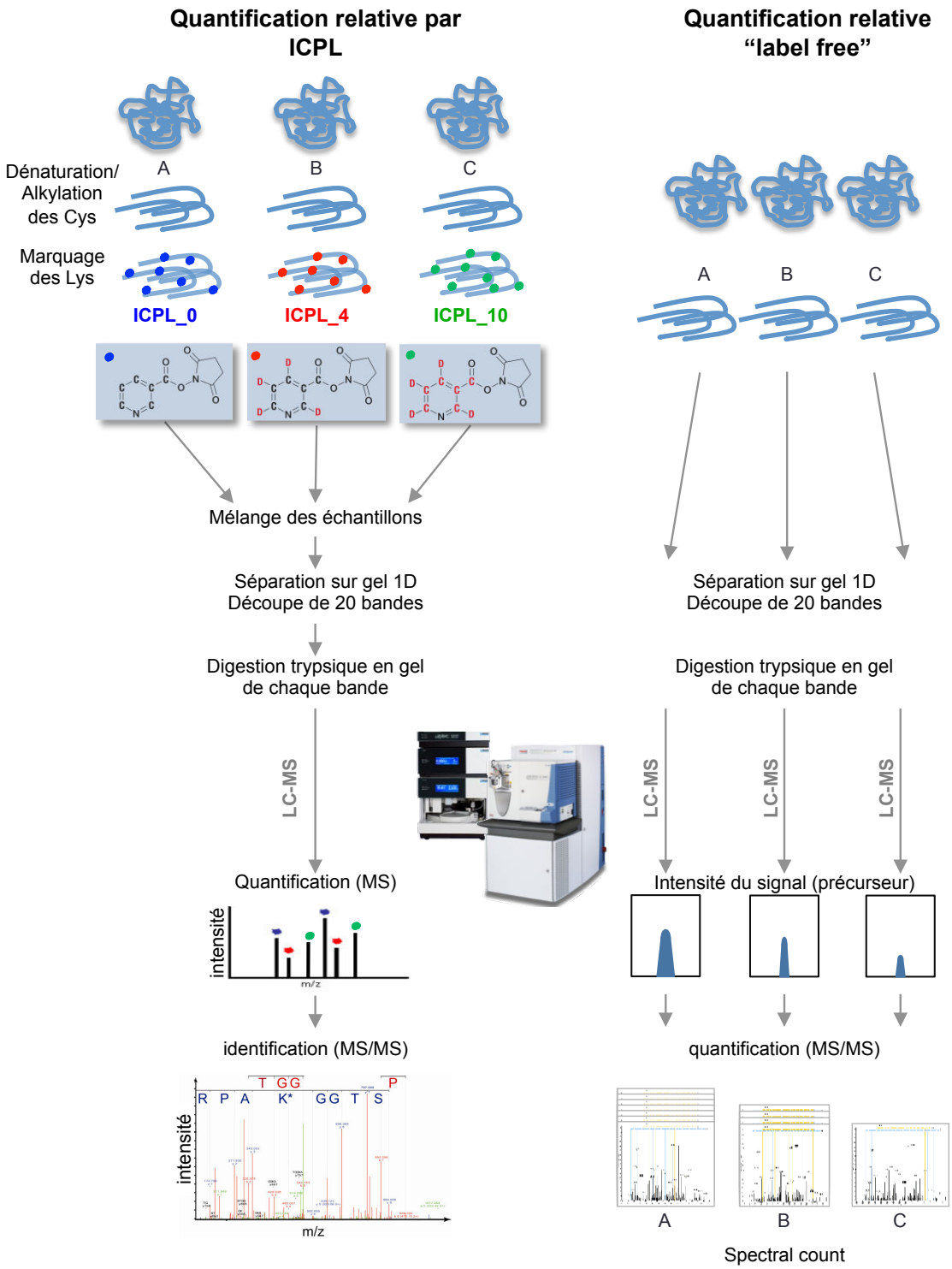

Figure 11. Comparaison schématique de deux approches protéomiques différentielles visant à quantifier les protéines dans différents échantillons

A gauche une approche utilisant un marquage isotopique (l'ICPL) et à droite une approche de type label free « spectral count ». Dans le cas de l'ICPL, la quantification se fait au stade MS avec l'intensité des pics des ions précurseurs.

D. Les bases de données de séquence

Sans bases de données de séquences protéiques ou nucléotidiques, pas d'identification de peptides ni de caractérisation des protéines possible lors d'études protéomiques! Ces banques sont maintenant essentielles à l'identification des protéines d'ailleurs facilitée si celles-ci sont les plus complètes possibles, et constituées de séquences de qualité. Les bases de séquences sont apparues au début des années 1980 (Grantham et al., 1981). En Europe, la banque de séquences EMBL data library a été créée et diffusée en 1986 (Hamm et Cameron, 1986). Cette équipe travaille au sein du Laboratoire Européen de Biologie Moléculaire l'EBI (European Bioinformatics Institute, Cambridge). Du côté américain, la banque d'acides nucléiques GenBank a été créée à Los Alamos (Bilofsky et al., 1986). Cette base de données est diffusée maintenant par le NCBI (National Center for Biotechnology Information). La collaboration entre l'EBI et le NCBI a commencé relativement tôt. Elle s'est étendue en 1987 avec la participation de la DDBJ (DNA Data Bank of Japan) pour donner naissance en 1990 à un format unique dans la description des caractéristiques biologiques qui accompagnent les séquences dans les banques de données nucléiques (Emmert et al., 1994; Overton et al., 1994).

Pour les protéines, deux banques principales ont été créées. La première, Protein Identification Ressource (PIR-NBRF) à Washington, qui produit maintenant une association de données issues du MIPS (Martinsried Institute for Protein Sequences), de la base Japonnaise JIPID (Japan International Protein Information Database) et des données propres de la NBRF (George et al., 1986). La deuxième, Swiss-Prot, constituée à l'Université de Genève à partir de 1986 regroupe entre autres des séquences annotées de la PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL (Bairoch et Apweiler, 1999; Bairoch et Boeckmann, 1993). Désormais, la ressource universelle de protéines UniProt (http://www.UniProt.org) a pour but de fournir à la communauté scientifique une approche globale de haute qualité et des ressources libres de séquences de protéines et annotations fonctionnelles. UniProt est produit par le UniProt Consortium, qui se compose de groupes de l'EBI (Institut européen de bioinformatique), du SIB (Institut Suisse de Bioinformatique) et de la PIR (Protein Information Resource) (The UniProt Consortium, 2014). Ses membres intègrent, interprètent et normalisent les données de la littérature (biocuration) et de nombreuses ressources afin de construire le catalogue le plus complet possible de l'information sur les protéines. UniProt est mis à jour et distribué toutes les 4 semaines. Les données de la PIR y sont intégrées depuis 2003. Ces banques et les services associés

(annotation manuelle), et les liens avec les autres bases de données, telle que la banque de mutualisation de données de protéomique : PRIDE, sont essentiels à l'annotation des génomes. Les utilisateurs peuvent soumettre des données dans ces banques, afin d'y contribuer (The UniProt Consortium, 2014).

a) Banques de séquences d'acides nucléiques (ENA, GeneBank, DDBJ)

Les banques ENA (EMBL-Bank), GenBank et DDBJ constituent des archives et des bases publiques des séquences primaires de séquences et des annotations associées fournies par les laboratoires qui les ont séquencées. Elles contiennent les séquences publiques dérivées des projets de séquençage des génomes, des centres de séquençage (cDNAs, ESTs…), de chercheurs individuels et de l'EPO (European patent office). Ces banques contiennent actuellement environ 266 millions de séquences provenant de plus de 300 000 espèces. Ces banques peuvent être cependant très redondantes pour certains loci. Il existe un échange quotidien entre ces banques de données et la banque NCBI.

b) Bases de données de séquences protéiques

(1) Finalités et « défis »

La finalité de ces bases de données est:

- de permettre l'identification des protéines par spectrométrie de masse pour laquelle leur « exhaustivité » et la qualité des séquences sont requises ;
- de permettre la recherche de similarités et prédictions fonctionnelles pour lesquelles la qualité de séquence (non redondance) et des annotations est requise ;
- de servir les outils de prédiction requérant séquences et annotations de qualité ;
- de permettre l'annotation des génomes requérant le maximum d'exhaustivité possible ainsi que la qualité des séquences et des annotations.

Les séquences protéiques qui se retrouvent dans les bases UniProt/Swiss-Prot et NCBI (NCBInr, refSeq) sont dérivées des séquences d'ADN codantes ou CDS (coding sequences), par la traduction des mRNAs transcrits de ces séquences d'ADN sur la base de prédictions des codons Start et Stop et des sites d'épissage. Si les séquences nucléotidiques soumises ne sont pas des CDS annotées, elles alimentent les banques de prédiction de gènes Ensembl et

RefSeq accessibles depuis les banques de données protéiques UniProt et NCBI. Etant donné les exigences des utilisateurs de bases de données pour les différentes applications que nous avons listées, ces derniers sont confrontés à plusieurs défis :

- la qualité des prédictions de gènes ;
- le nombre de bases de données de séquences différentes et des ensembles de données différents pour une même espèce ;
- de multiples identifiants pour une même protéine, par exemple la ferritine de rat possède les identfiants suivants : D3ZIZ8 (entrée UniProt); ENSRNOP00000044004; ENSRNOP00000044841 (identifiants ENSEMBL); GENSCAN00000044496; GENSCAN00000044514 (identifiants Genscan; Genscan étant l'algorithme de prédiction de gènes le plus populaire).

(2) Origine des séquences protéiques ?

UniProtKB est la plus grande base de données de protéines pour un grand nombre d'êtres vivants y compris pour les virus et les bactéries. Les séquences de Swiss-Prot, PIR, PRF sont aussi présentes dans la banque Protein NCBI qui regroupe aussi les séquences traduites de GenBank, RefSeq et TPA (Third Party Annotation). Dans la base de données UniProt, à peu près 98% des séquences protéiques sont dérivées de la traduction de séquences nucléotidiques : ARNm (cDNA, EST), gènes et génomes ; et seulement 1% du séquençage direct (séquençage d'Edmann, et MS/MS). La question de la qualité des séquences protéiques se pose alors, car ces dernières se basent sur des prédictions de gènes qui sont plus ou moins fiables. Les séquences protéiques présentes dans UniProt dérivent pour la plupart de:

- INSDC (International Nucleotide Sequence Database Collaboration) séquences traduites de CDS soumises (95,1%);
- prédictions à partir de prédictions de gènes de Ensembl (3,2%), et RefSeq (0,3);
- séquences issues de structures de PDB (ProteinDataBank, données 3D et séquences associées);
- séquences soumises directement par des utilisateurs ou scannées de la littérature.

(3) UniProtKB, d'où viennent les annotations?

La base de données de séquences protéiques UniProtKB se divise en deux sections; UniProt/Swiss-Prot et UniProt/Trembl, basées sur la nature de leurs annotations de séquences. Dans la mise à jour du 11 juin 2014:

- UniProt/Swiss-Prot contient 545.536 séquences ;
- UniProtKB/TrEMBL contient 69,014.937 séquences

(a) UniProt/Trembl

UniProt /Trembl contient des séquences dites « unreviewed » qui proviennent d'une annotation automatique de EMBL à Trembl. La qualité des séquences protéiques dépend de l'information fournie par le soumettant de l'entrée originale (CDS) ou bien de la « pipeline » de prédiction de gène (ex : ENSEMBL). Les séquences identiques à 100%, de même longueur, appartenant au même organisme sont fusionnées automatiquement. L'information biologique, source d'annotation, est fournie par le soumettant par le biais de diverses sources : PDB, TAIR (The Arabidopsis Information Ressource), EMBL,….D'où la question de leur cohérence comme ci dessous.

```
DE   4-aminobutyrate QUI SE DILATE aminotransferase (EC 2.6.1.19).
GN   PYRAB12830 OR PAB2386.
OS   Pyrococcus abyssi.
OC   Archaea; Euryarchaeota; Thermococci; Thermococcales; Thermococcaceae;
OC   Pyrococcus.
OX   NCBI_TaxID=29292;
```

Figure 12. Annotation erronée dans UniProtKB

(source : Institut Suisse de Bioinformatique, SIB).

L'information biologique peut aussi venir d'annotations automatiques à partir de règles d'annotation SAAS et UniRule. Les règles SAAS sont générées automatiquement. Elles sont basées sur l'algorithme de classification supervisée C4.5. Des problèmes peuvent survenir lors de l'annotation automatique, comme des sites d'initiation de la traduction des protéines non corrigés, des décalages dans le cadre de lecture, donc une séquence protéique fausse.

Une annotation automatique à partir de projets de génomes sans annotation manuelle peut être problématique. Si les prédictions sont faites automatiquement sans intervention manuelle, les protéines traduites peuvent être en grande partie fausses. A titre d'exemple, on peut citer l'annotation du génome de *Tetraodon nigroviridis*: chez qui plus de 90% des gènes modèles produisent des protéines incorrectes. Si des mises à jour ne sont pas faites, une certaine proportion de gènes modèles peuvent s'avérer être erronés. C'est le cas d'*Arabidopsis* pour qui un grand nombre d'annotations ont été produites au moment du séquençage, alors qu'aucune mise à jour n'a été faite depuis. Environ 20% des gènes modèles sont erronés. Au contraire, le génome de la Drosophile semble être un bon exemple grâce à Flybase, car les gènes modèles ne produisent des séquences différentes de ce qui est dans UniProt qu'à une faible proportion (environ 2%). Même pour les bactéries et les Archaea qui n'ont presque pas d'épissage, il reste des erreurs dans les prédictions de gènes (codons Start, petites protéines manquées <100a.a).

(b) UniProt/Swiss-Prot

Les séquences présentes dans Swiss-Prot apparaissant comme « reviewed » dans UniProt. Elles sont annotées manuellement par les biocurateurs, comme en témoignent des références bibliographiques associées plus nombreuses pour une séquence donnée, ainsi qu'une annotation bien plus complète dont la présence d'isoformes. Les annotations manuelles sont importantes dans la mesure où elles permettent de corriger par exemple une fonction validée expérimentalement, différente de ce qui avait été prédit pour une protéine donnée. Ceci permet de réutiliser ces nouvelles annotations dans les systèmes de prédiction de fonction et d'annotation automatique. L'annotation manuelle est donc essentielle à l'entretien de la connaissance et chaque mise à jour est rendue publique toutes les 4 semaines dans UniProt.

Les annotations manuelles proviennent de diverses sources. Elles viennent de l'analyse manuelle de séquences par des alignements multiples. Elles viennent aussi de l'utilisation des outils de prédiction de : topologie (domaines transmembranaires avec l'outil TMHMM, peptide signal avec SignalP, …), de modifications post-traductionnelles (avec par exemple GPI-predictor pour les sites d'ancrage GPI lipidiques), présence de domaines fonctionnels (avec par exemple ProSite ou InterPro). Les annotations manuelles viennent aussi de la littérature (PubMed) ou de contacts directs avec les experts, d'autres bases de données, ou bien de comités de nomenclature. La source de chaque annotation est traçable par des

attributions d'évidences (« étiquettes » à chaque annotation), telles que : « par similarité », « probable », « potentiel », ou par un numéro de référence pour indiquer une référence de la littérature. Les protéines correspondant aux entrées UniProt ont des « étiquettes d'évidence » PE (protein evidence) de 1 à 5 :

1- Protein level, ≈18% (caractérisé par spectrométrie de masse, western blot dans un tissu particulier, immunohistochimie avec une localisation sub cellulaire…);
2- Transcirpt level, ≈19% (molécule de type mRNA présente dans EMBL…);
3- Inferred by homology, ≈58% (au moins une prédiction de similarité de domaine, d'appartenance à une famille…);
4- Predicted, ≈5% (par défaut);
5- Uncertain, la plupart sont dans TrEMBL (produit d'une prédiction douteuse de CDS ou de gène, ou bien produit d'un pseudogène).

On peut consulter les règles d'attribution de ces « PE » mises à jour régulièrement et leur priorité d'attribution (http://www.uniprot.org/docs/pe_criteria). En effet, le fait qu'une séquence protéique soit « reviewed », donc manuellement annotée, ne signifie en aucun cas que l'existence de la protéine ait été démontrée d'une façon expérimentale, et elle peut très bien être annotée comme « uncertain » ou « inferred by homology ». Quand une entrée est annotée manuellement et entrée dans Swiss-Prot, celle-ci est supprimée de TrEMBL afin de minimiser la redondance au sein de UniProtKB.

E. Ontologies et bases de données biologiques

Aujourd'hui les approches protéomiques de type Shotgun permettent d'obtenir des listes de l'ordre de plusieurs milliers de protéines. Ces listes sont fastidieuses à traiter et à analyser de manière manuelle. L'intégration aux ensembles de données protéomiques des annotations telles que celles de la Gene Ontology, des données d'interactomique ou des voies de signalisation dans lesquelles les protéines sont impliquées, constitue depuis quelques années une aide à l'analyse de ces grandes listes de protéines. En effet, il est utile pour une analyse protéomique de découverte de pourvoir consulter les bases de données contenant un grand nombre d'informations biologiques en les intégrant aux données de protéomique pour orienter la recherche de protéines d'intérêt, comme ce qui a été réalisé dans mes projets de thèse. Les informations dérivées de ces bases, et les ontologies permettent d'établir des

critères de choix de candidats notamment si l'on veut restreindre notre champ d'investigation à une question biologique bien précise. Grâce à l'intégration de ces données biologiques *via* les identifiants (souvent le GeneID, identifiant du gène), il est possible d'exploiter de longues listes d'identification de protéines selon diverses approches dans le but d'extraire des protéines à priori pertinentes par rapport à l'objectif de l'étude, pour une analyse ultérieure approfondie de ces protéines dans le contexte expérimental.

a) Les bases de données biologiques

Ces bases de données biologiques sont nombreuses, elles vont des prédictions de domaines et de familles de protéines avec des bases telles que Pfam (Finn et al., 2014), aux bases d'interactions telles que BioGrid (Winter et al., 2011) ; MINT et MIntAct (Licata et al., 2012; Orchard et al., 2014). Sont également utilisables des bases telles que KEGG PATHWAYS pour la prédiction d'interaction ou de voies de signalisations dans lesquelles les protéines sont impliquées. Les liens des bases de données de séquence vers les autres bases de données d'informations biologiques sont présents *via* les « cross references », mais pour l'interprétation des listes entières de protéines générées par protéomique Shotgun en particulier, il est aussi nécessaire de pourvoir intégrer les informations biologiques (annotations) contenues dans les différentes bases de données, pour pouvoir orienter une analyse protéomique de manière à restreindre de champ d'investigation. En effet, de nos jours, les résultats d'identification particulièrement en protéomique Shotgun peuvent prendre la forme de listes de milliers de protéines identifiées dans un échantillon donné, rendant la fouille manuelle difficile voire impossible. Plusieurs outils sont disponibles tels que AMEN (Chalmel et Primig, 2008), DAVID, le plus utilisé (Huang et al., 2009), ou Ingenuity (Krämer et al., 2014). Ils permettent d'intégrer ces données et de réaliser un certain nombre d'analyses statistiques. L'une des fonctionnalités de ces outils est de réaliser des cartes d'enrichissements en termes de la Gene Ontology permettant d'en avoir une vue d'ensemble. La prédiction de réseaux d'interactions est également une option informative. Toutefois, il faut être prudent quant à l'utilisation de ces outils qui se basent sur des annotations d'origines diverses, et il convient de bien s'assurer que la stringence de l'analyse correspond à une question biologique donnée, sans compter qu'*a posteriori*, des validations biochimiques et des preuves expérimentales seront indispensables pour tirer des conclusions de ces analyses protéomiques.

b) La Gene Ontology

En biologie comme dans d'autres domaines, le même nom peut être parfois utilisé pour décrire différents concepts, et à l'inverse, un concept peut être décrit avec plusieurs noms. La comparaison est donc difficile entre les différentes espèces, domaines de recherche et même entre les différentes bases de données. Une solution intéressante vise à utiliser une nomenclature commune avec des noms uniques et une définition non ambiguë. Le projet Gene Ontology est une initiative communautaire de bioinformatique majeure, une ressource bioinformatique ayant pour but de standardiser la représentation de gènes et produits de gènes à travers les espèces et les bases de données (Gene Ontology Consortium, 2013). Une ontologie représente formellement la connaissance comme un ensemble de concepts dans un domaine ainsi que les relations entre ces concepts. Elle peut être utilisée pour modéliser un domaine et supporter le raisonnement entre les entités. Le projet Gene Ontology fournit un vocabulaire contrôlé de termes pour classifier les fonctions des produits des gènes et décrire leurs caractéristiques, mais pas les noms des produits de gènes. Des termes de l'ontologie existent pour décrire les **composants cellulaires,** c'est à dire les endroits au niveau des structures sub-cellulaires et des complexes macromoléculaires dans lesquels s'expriment un produit de gène situé dans un sous-composant d'un composant ou compartiment cellulaire particulier. Par exemple le terme « acrosomal vesicle » est une granule sécrétrice « secretory granule » qui fait partie de « endomembrane system ».

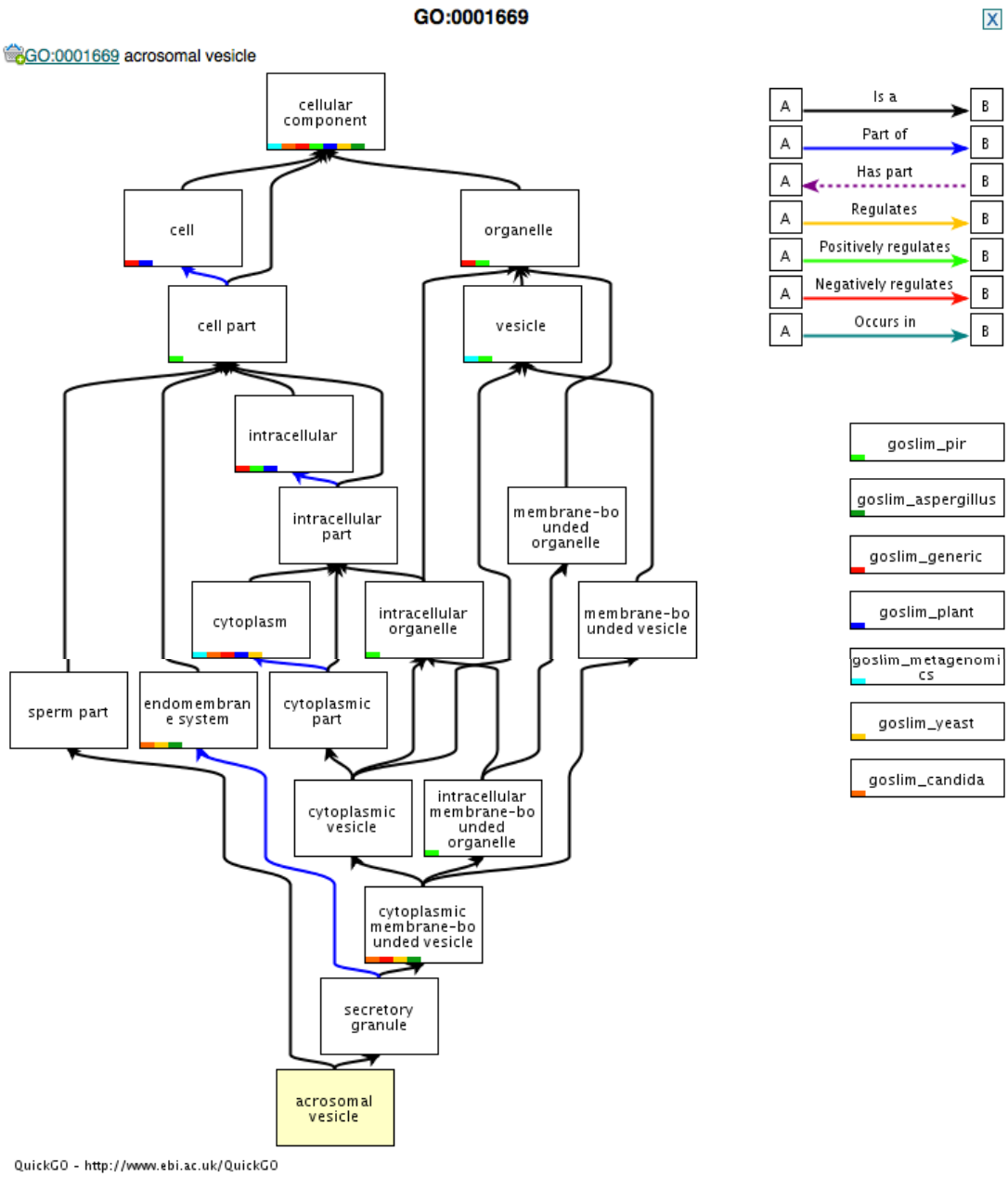

Figure 13. Parenté du terme de la Gene Ontology « acrosomal vesicle », vue de QuickGO (http://www.ebi.ac.uk/QuickGO).

Des termes désignant les **fonctions biologiques** d'un produit de gène sont les emplois qu'il a ou les capacités dont il dispose. Ceux-ci peuvent inclure le transport de molécules, la liaison, et la transformation. Ceci est différent des **procédés biologiques** dans lesquels sera impliqué un produit de gène, désignés aussi par des ontologies. On pourrait comparer cela avec une

organisation dans laquelle les individus (produits de gènes) ont des capacités différentes ou des tâches (fonctions) et travaillent ensemble pour atteindre des objectifs différents (processus). Les ontologies sont donc classées selon différents thèmes qui ont trait soit : 1) à un composant cellulaire, partie d'une cellule ou de son environnement extracellulaire où s'exprime un produit de gène; 2) à une fonction moléculaire élémentaire ou activité élémentaire d'un produit de gène au niveau moléculaire (telle que liaison ou catalyse); 3) à un processus biologique, ensemble d'événements moléculaires avec un début et une fin bien définis, pertinents pour le fonctionnement d'unités vivantes intégrées (cellules, tissus, organes et organismes) dans lequel il intervient. Ces termes sont hiérarchisés selon leurs liens de parenté. Par exemple, le terme « organelle » a deux enfants :

1) « mitochondrie» qui **est une** organelle,
2) « organelle membrane » qui **fait partie de** « organelle ».

Le terme « Mitochondrie » a deux parents :

1) c'**est un** « organelle » et
2) elle **fait partie de** « cytoplasme ».

Les relations « **is a** » et « **part of** » sont les deux relations les plus communément utilisées dans l'arbre de la Gene Ontology. La relation « **regulates** » est plus compliquée, car par exemple, si A est partie de B qui régule C, on ne peut pas conclure que A régule C.

Chaque terme possède donc un identifiant, un nom, une définition, un synonyme, un ensemble de données de référence et une parenté.

III. Les études protéogénomiques

Historiquement, les communautés de la génomique et de la protéomique ont travaillé en totale indépendance. Le rôle de la communauté de la génomique a été d'identifier les gènes et les séquences protéiques correspondantes. Cela a été fait souvent grâce à des efforts d'annotation à grande échelle, pendant et après le séquençage du génome. La collection de protéines dérivées a été considérée comme un ensemble fixe. Les scientifiques de la protéomique ont eu pour objectif de comprendre quelles protéines sont exprimées dans quels tissus ou cellules dans des conditions spécifiques et d'identifier les diverses modifications post-traductionnelles ainsi que d'autres maturations du protéome. En revanche, au sein de la protéogenomique, la protéomique et la génomique œuvrent ensemble afin de clarifier la structure des gènes.

Chez les eucaryotes, pour un transcrit donné, il peut y avoir différents motifs d'épissage alternatif dont chacun produit un ARNm mature différent et donc des protéines différentes (Figure 14).

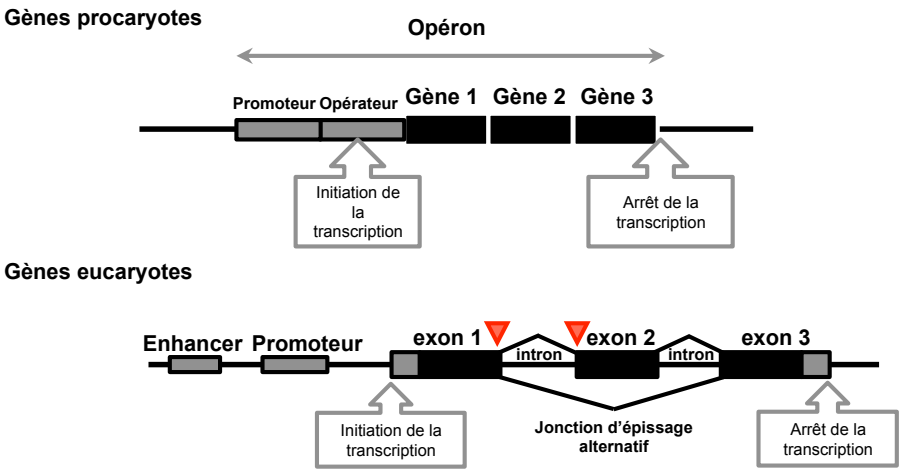

Figure 14. Structure d'un gène eucaryote comparée à celle d'un gène procaryote

Les gènes procaryotes peuvent être disposés dans un opéron, partageant le même promoteur, tandis qu'un gène eucaryote contient des régions codant pour des protéines appelées exons, séparés par des régions non codantes de protéines appelées introns. Une fois transcrit, les introns sont épissés pour ne garder que les exons dans l'ARN mature. Une jonction d'épissage alternatif est représentée, ainsi que trois exons (boîtes noires, les têtes de flèches rouges indiquent les régions d'épissage, les boîtes grises représentent les régions non traduites).

Bien que de nombreux programmes de prédiction de gènes soient disponibles, l'identification des gènes dans un organisme donné diminue en précision de façon drastique à mesure que l'échelle au niveau nucléotidique augmente: des exons aux structures entières de gènes (Zhang, 2002). Le plus souvent, les gènes courts ne sont pas prédits (Warren et al., 2010), et les isoformes d'épissage alternatifs sont également difficiles à prédire (Reese et al., 2000).

A. La protéogénomique

La protéogénomique à strictement parler est l'utilisation des données de la protéomique obtenues par spectrométrie de masse en tandem pour améliorer l'annotation des génomes grâce à la validation expérimentale que procurent les peptides identifiés. En général, l'annotation du génome se construit grâce à des «pipelines» en utilisant notre compréhension actuelle de l'annotation des gènes et sont donc sujettes à des erreurs et des incohérences. Le terme "protéogénomique" a été utilisé la première fois pour décrire l'alignement de peptides identifiés par spectrométrie de masse en tandem au cours d'une analyse du protéome, sur la séquence d'acides nucléiques codant pour ces peptides à leur locus spécifique (Jaffe et al., 2004). Dans ce cas, les spectres MS/MS sont attribués en utilisant une base de données comprenant tous les cadres ouverts de lecture ouverts possibles, c'est à dire la traduction des acides nucléiques dans les six cadres de lecture. Depuis une étude pionnière chez *Mycoplasma pneumoniae* (Jaffe et al., 2004), de nombreuses réannotations protéogénomiques de génomes ont été rapportées (voir pour des revues récentes: (Armengaud et al., 2013; Renuse et al., 2011)).

a) La protéogénomique chez les procaryotes

La plus grande étude de protéogénomique à ce jour a été réalisée par Venter et al., (Venter et al., 2011) qui a rapporté l'analyse protéogénomique de 46 organismes à la fois chez les archées et les bactéries. Au cours de cette étude, à peu près 700 nouvelles protéines ont été identifiées chez *Deinococcus radiodurans* dont le génome a été publié en 1999 et pour qui le taux d'annotations erronées était exceptionnellement élevé. A l'issue de différentes études chez les procaryotes, il est possible de dire que la protéogénomique permet de :

- valider des gènes prédits;
- détecter de nouvelles séquences codantes (gènes non prédits) et caractériser les protéines correspondantes;

- détecter des petites protéines qui n'auraient pas été prédites;
- détecter des nouveaux sites d'initiation de la traduction;
- corriger des sites d'initiation;
- détecter des protéines prédites à partir de prédiction de gènes « douteuses »;
- détecter des orientations inversées de gènes (correction de l'orientation de gènes).

Des exemples d'inversions d'orientation de gène ont été identifiées par protéogénomique sur les gènes ddrC et ddrH spécifiquement induits par les radiations chez *Deinococcus radiodurans* (et D. *geothermis*) dont le sens de la région codante est inversé comparé à *Deinococcus deserti* (de Groot et al., 2009). On peut aussi donner un exemple de nouveaux évènements codants importants dans un processus: deux nouvelles recA impliquées dans la réparation de l'ADN, exprimées après exposition aux UVs chez *Deinococcus deserti* (de Groot et al., 2009), ont pu être identifiées par protéogénomique. En plus de la découverte de nouvelles séquences codantes, ou de leur correction, le raffinement de l'annotation structurelle des gènes peut être obtenu par catalogage de peptides N-terminaux, de sorte que le codon d'initiation de traduction exact (Baudet et al., 2010) ou le peptide signal précis (sites de maturation) puissent être identifiés (Armengaud, 2009).

b) La protéogénomique chez les eucaryotes

L'annotation des gènes eucaryotes est bien plus compliquée par la prévalence de l'épissage alternatif des gènes qui jouent un rôle clé dans la production de la diversité du protéome. La détection fiable des jonctions d'épissage est difficile, et les algorithmes de recherche *ab initio* prédisent souvent un transcrit unique à un locus donné en ignorant complètement les variants d'épissage. Les outils reposant sur les ESTs (expressed sequence tags) pour prédire les variants d'épissage posent également des problèmes pour obtenir une annotation précise car les ESTs ne couvrent pas toute la séquence d'un gène et le séquençage n'est pas précis. De plus, il est difficile de savoir si deux transcrits alternatifs diffèrent de par leur région codante ou de par des UTRs.

Les signaux génomiques qui régissent le fonctionnement des prédicteurs *ab initio* conduisent à des erreurs dans la prédiction des ORFs et des frontières exons/introns. Or, par exemple, pour qu'un codon stop (« TGA », « TAA », « TAG ») soit bien prédit, il faut que le cadre de lecture du dernier exon soit correctement prédit. Mais les signaux de codage basés sur des hexamères ne sont pas suffisants pour déterminer le cadre de lecture des exons courts, et donc,

des exons courts peuvent être manqués par les prédicteurs, ainsi que des gènes courts (<100a.a.) qui peuvent pourtant coder pour des protéines (Oshiro et al., 2002). Des erreurs de prédiction peuvent aussi arriver pour des gènes longs dont la composition en G et C est différente (Burge et Karlin, 1997).

Dans le sillage des avancées technologiques dans le séquençage de l'ADN, le nombre de génomes eucaryotes séquencés a augmenté de façon spectaculaire au cours des 20 dernières années avec des génomes disponibles pour *Saccharomyces cerevisiae* (Goffeau et al., 1996), *Caenorhabditis elegans* (C. elegans Sequencing Consortium, 1998), *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), *Drosophila melanogaster* (Adams et al., 2000), *Homo sapiens* (Lander et al., 2001; Venter et al., 2001), *Mus musculus* (Mouse Genome Sequencing Consortium et al., 2002), *Anopheles gambiae* (Holt et al., 2002), *Rattus norvegicus* (Gibbs et al., 2004), et récemment, *Zea mays* (Schnable et al., 2009). Comme les séquences du génome de nombreux organismes modèles sont disponibles, les études protéogénomiques à grande échelle commencent à occuper une place importante, contribuant à l'amélioration de l'annotation de ces génomes.

Au cours des dix dernières années, des études protéogénomiques ont confirmé l'expression de 25% des ORFs chez la levure (Oshiro et al., 2002), ou de 224 protéines hypothétiques humaines (Tanner et al., 2007), les peptides identifiés dans ces études contribuant à la validation de gènes putatifs. Atteindre une large couverture du protéome donc est essentiel à la construction d'un catalogue complet et précis des gènes. Dans une étude récente, Brosh et collaborateurs ont validé par spectrométrie de masse l'aspect codant de 32% des gènes codants de 17% des exons et de 7% des jonctions introns-exons chez la souris, en interrogeant plus de 10 millions de spectres sur les prédictions de protéines contre l'ensemble du génome (Brosch et al., 2011). Chez la souris encore, une étude utilisant une base de jonctions exoniques et une base d'ORFs obtenues à partir du génome pour aligner les séquences des peptides obtenus par MS/MS a permis de valider 4471 gènes du point de vue de la traduction, 172 nouveaux évènements géniques, 52 évènements d'épissage et 120 ORFs intergéniques mais aussi intragéniques. Cette étude montre donc que de nouvelles régions introniques peuvent s'avérer codantes (Xing et al., 2011).

Ce qui peut se produire lors de l'alignement de peptides dans une région annotée est représenté (Figure 15). En effet, des peptides identifiés par protéogénomique peuvent ne correspondre à aucune région génomique connue, ils sont alors qualifiés de « nouveaux ». Ces nouveaux peptides alignés contre un génome peuvent se révéler soit intragéniques à un

locus connu à l'intérieur d'une structure de gène connu, soit intergéniques à l'extérieur d'un gène connu. Ils permettent de proposer différentes catégories de ré-annotation du génome.

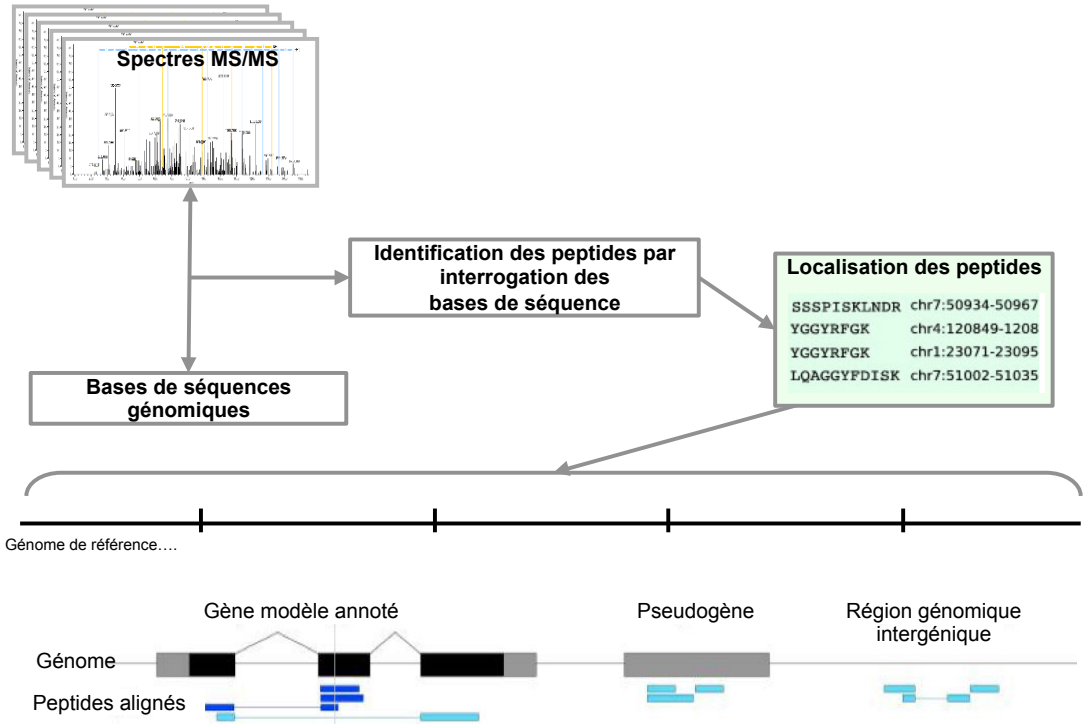

Figure 15. Exemple d'alignements de séquences de peptides identifiés par MS sur une région génomique contenant un gène annoté

Plusieurs peptides (bleu et bleu ciel) entrant dans un locus du gène annoté, valident la traduction de deux exons. Un peptide bleu ciel indique une nouvelle isoforme d'épissage qui saute l'exon interne. Des peptides supplémentaires relèvent d'un pseudogène annoté donnant une indication forte de sa traduction dans la cellule. Des peptides qui s'alignent dans la région intergénique indiquent probablement de nouveaux loci codant pour des protéines. Les peptides bleu ciel ne seraient probablement pas identifiés en utilisant une base de données protéomique canonique.

Chez les eucaryotes, un peptide découvert par protéogénomique et aligné contre le génome peut donc fournir une information unique quant à l'annotation de gène en permettant de:

- confirmer la traduction, et permettre de discriminer les pseudogènes des gènes codants (Lewis et al., 2003a);
- déterminer des ORFs et même des ORFs chevauchants;
- vérifier la localisation des sites d'initiation de la traduction;
- trouver des sites de clivage en jeu dans la maturation post-traductionnelle, comme le clivage du peptide signal;
- identifier les frontières exactes des exons si le peptide est partagé entre deux exons;

- identifier des variants d'épissage, incluant les variants d'épissage alternatif;
- identifier de nouveaux gènes codants.

Cinq améliorations d'annotation des gènes possibles chez les eucaryotes sont représentées (Figure 16).

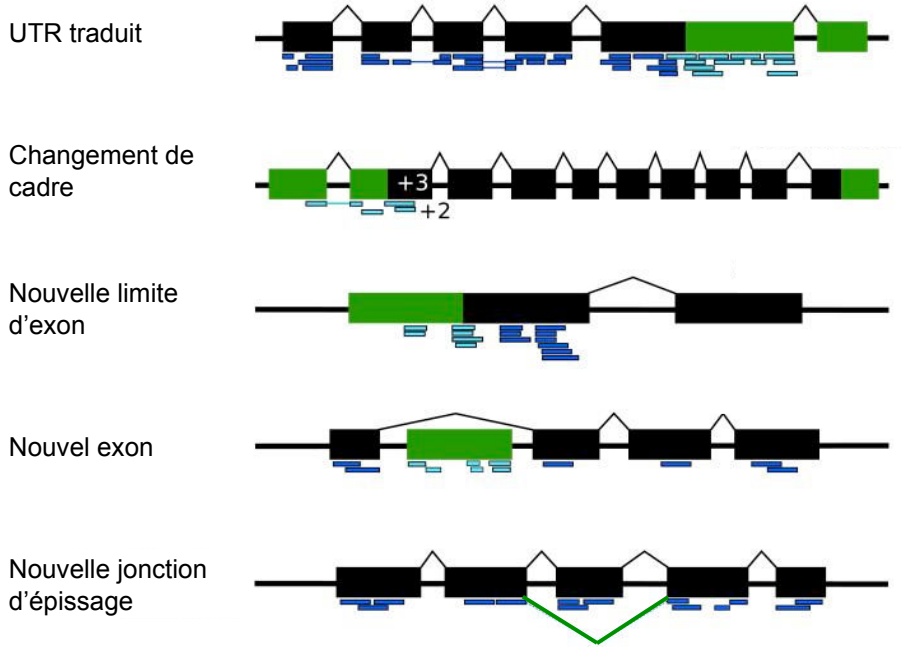

Figure 16. Cinq événements différents d'amélioration d'annotation d'un gène proposés par des peptides intragéniques

Les exons sont présentés dans des boîtes noires tandis que les nouvelles régions codantes ou les nouvelles formes d'épissage suggérées par l'alignement de peptides identifiés par MS sont indiquées en vert. Les peptides (bleu foncé et bleu clair) identifiés par MS sont représentés alignés aux gènes modèles (d'après Castellana et Bafna, 2010; Castellana et al., 2008).

Reconstruire la structure d'un gène à partir de peptides identifiés en protéomique Shotgun n'est pas trivial, parce que la qualité de l'information qu'ils fournissent sur une séquence génomique est limitée par leur couverture de séquence. Pour de nouveaux peptides intragéniques, il est difficile de savoir si la structure de gène doit être corrigée, ou bien si leur présence est due à une nouvelle forme d'épissage du transcrit. Les peptides jonctionnels « d'épissage » c'est à dire ceux qui sont à cheval sur deux exons, fournissent des informations sur les exons qui se raccordent mais ils ne sont pas informatifs sur un exon plus distal, et donc ne donnent pas accès au schéma d'épissage complet d'une isoforme. Il est donc nécessaire d'utiliser des données extrinsèques, telles que les séquences des transcrits ou

des régions génomiques homologues pour distinguer les deux cas (Baerenfaller et al., 2008; Castellana et al., 2008). Pour revues, lire (Armengaud et al., 2014; Castellana et Bafna, 2010; Renuse et al., 2011).

B. La protéogénomique au sens élargi

La protéogénomique au sens large comprend un ensemble de technologies qui permettent d'interroger des données de spectres de MS en tandem afin de découvrir de nouveaux gènes codants pour des protéines. Il semblerait donc que contrairement à la définition de la protéogénomique au sens strict qui consiste à aligner des peptides identifiés par MS/MS sur les génomes pour en améliorer l'annotation, la jonction entre protéomique et génomique se fasse plus intimement par l'intermédiaire de la construction de bases de séquences traduites des acides nucléiques nouvellement séquencés. La protéogénomique au sens large concerne donc les projets qui utilisent le séquençage à haut débit de l'ADN sans l'unique intention d'annoter le génome, mais plutôt de créer ces bases de données de séquences pour l'interprétation des spectres MS / MS.

Certaines approches récentes sont conformes à cette définition (Armengaud et al., 2014). Pour citer un exemple, en l'absence du génome complet du parasite protozoaire *Leishmania donovani*, Pawar et collaborateurs (Pawar et al., 2012) ont utilisé une stratégie inspirée de la protéogénomique en tenant compte des trois génomes les plus connexes, à savoir ceux de *Leishmania major*, *Leishmania infantum* et *Leishmania brasiliensis,* traduits dans les six cadres de lecture possibles, pour combler des manques dans chacun d'eux. Un total de 3711 protéines ont été identifiées par MS/MS grâce à la banque de séquences traduites qui a été utilisée pour interroger les données spectrales. L'ensemble de données a été utilisé pour affiner certains gènes dans les génomes respectifs de référence.

a) Protéogénomique et organismes non-modèles

L'annotation basée sur l'homologie est fastidieuse chez les organismes non-modèles, pour lesquels des génomes d'organismes proches par homologie ne sont pas disponibles. Les résultats sont en tout cas limités aux peptides les plus conservés des protéines les plus conservées. Par conséquent, séquencer le génome d'un organisme non modèle semble la stratégie la plus appropriée pour construire une base de données de séquences protéiques pouvant servir à l'interprétation des spectres MS/MS. Les efforts de séquençage sont

directement proportionnels à la complexité du génome et dépendent de la présence d'organismes apparentés dont le génome a déjà été séquencé, qui pourra servir de possible échafaudage pour l'assemblage du nouveau génome. Bien que l'annotation structurale et fonctionnelle des gènes ne commence habituellement qu'avec la séquence quasi-complète du génome, l'annotation par des données de protéomique peut finalement être effectuée sur des séquences génomiques incomplètement séquencées avec une faible couverture du génome. L'analyse protéomique peut être effectuée avec une base de données partielle comme cela a été fait avec *Mannheimia haemolytica* (Nanduri et al., 2005) et *Bacillus megaterium* (Sun et al., 2006). Cette approche «quick and dirty» permet des identifications rapides et fiables sur la base de séquences des ORFs traduits dans les six cadres de lecture, du génome «brouillon», ou des CDS extraits de ce génome (Rubiano-Labrador et al., 2014). Les CDS ou ORFs identifiés par spectrométrie de masse sont annotés par homologie *via* des recherches Blast. Cette approche est appropriée aux organismes non-modèles, particulièrement pour les premières espèces d'un embranchement afin d'éviter la propagation des erreurs d'annotation.

b) La protéomique informée par la transcriptomique

Nous l'avons vu dans la section traitant des bases de données de séquences protéiques, celles-ci contiennent une grande proportion de séquences dérivées de prédictions de gènes. De ce fait elles peuvent contenir des séquences erronées, dues à des erreurs de séquençage au départ ainsi que dans le «pipeline» de prédiction de gènes, et maintenir ces erreurs malgré la biocuration. Par ailleurs, ces banques ne sont pas exhaustives. L'utilisation de bases de données dédiées traduites de séquences dérivées du séquençage *de novo* des transcrits d'un organisme ou d'un type cellulaire a révolutionné les approches protéomiques exploratoires, car elle offre une solution aux problèmes évoqués plus haut en facilitant l'identification de protéines (Wang et al., 2012). Ces approches sont qualifiées de PIT (protéomique informée par la transcriptomique), ainsi nommées par Evans et collaborateurs (Evans et al., 2012).

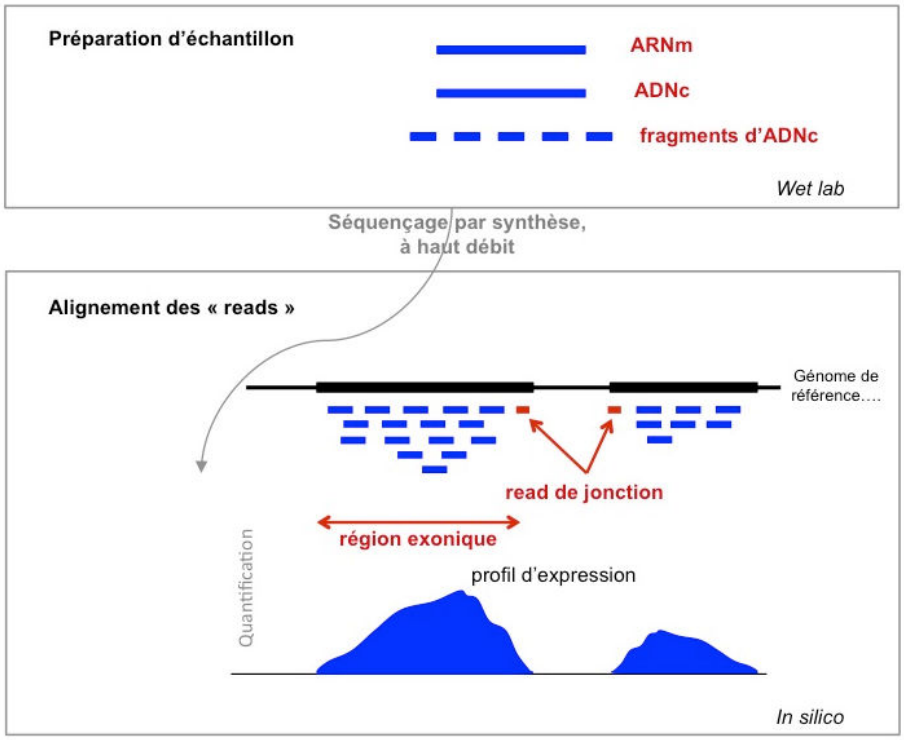

Figure 17. Principe du RNA sequencing

Schéma général du principe de séquençage *de novo* des ARNs, reconstruction du transcriptome et quantification des fragments (reads) de séquençage permettant d'obtenir un profil d'expression sur le génome de référence.

Les eucaryotes, contrairement aux procaryotes, contiennent comme nous l'avons vu une large quantité d'ADN intra génique et inter génique « non codant », avec un ratio important de « Junk DNA » contenant des séquences très répétées. En conséquence, seulement une petite partie du génome est considérée comme codante pour des protéines. De plus, à cause de l'épissage alternatif, la génération de bases de séquences protéiques s'avère plus difficile pour les organismes eucaryotes. Par conséquent, le séquençage des ARNs matures est une alternative pour générer rapidement des séquences de protéines par la traduction des transcrits (reverse transcrits ou cDNA) dans tous les cadres de lecture possibles, c'est à dire 3 quand le protocole utilisé est brin spécifique et 6 cadres quand le protocole utilisé n'est pas brin spécifique. L'utilisation de ces bases de données dédiées pourrait augmenter de manière significative la sensibilité de l'identification de peptides, réduire l'ambiguïté dans l'assemblage de la protéine, et permettre la détection de variants peptidiques connus et nouveaux (Wang et al., 2012, 2009). L'utilité de cette approche pour améliorer les annotations d'un génome a d'ailleurs été rapportée par la caractérisation de nouveaux gènes, de nouveaux exons, de nouveaux variants d'épissage, d'UTRs traduits, de décalages dans le

cadre de lecture et de transcriptions sur le brin anti-sens (Evans et al., 2012; Sheynkman et al., 2013; Wang et al., 2012; Woo et al., 2014). Cette approche permet également de quantifier les transcrits exprimés dans un échantillon donné en les alignant sur le génome de référence, et donc de quantifier les nouveaux évènements détectés (Figure 17).

Le RNA-seq a été appliqué avec succès à plusieurs espèces non-modèles, comme la morue de l'Atlantique *Gadus morhua* (Lanes et al., 2013) et le copépode marin *Calanus finmarchicus* (Lenz et al., 2014), parmi d'autres (pour une revue, lire (Armengaud et al., 2014)). Il a été utilisé pour améliorer l'annotation des génomes en révélant de nouveaux modèles de gènes et des événements d'épissage alternatif comme initié par Denoeud et al. (Denoeud et al., 2008). Essentiellement, l'identification des protéines peut être effectuée sur la base des informations de séquences d'acides nucléiques dérivées du RNA-seq sans l'inconvénient de passer par le séquençage et l'annotation du génome entier (Woo et al., 2014). Elle permet donc l'identification de protéines alors que le génome n'est pas séquencé. Par exemple, le RNA-seq sur des testicules totaux et la protéomique shotgun sur le spermatozoïde ont permis d'identifier de nouvelles protéines du spermatozoïde chez l'ormeau rouge *Haliotis rufescens*, qui est un modèle important pour l'étude des interactions moléculaires impliquées dans la fécondation. Les auteurs ont pu identifier 975 protéines du spermatozoïde dont la lysin et Sp6, deux protéines homologues aussi abondantes que la protéine majoritaire de l'acrosome chez l'ormeau. Sp6 aurait rapidement évolué et est probablement spécifique du spermatozoïde (Palmer et al., 2013). Une autre étude a été réalisée chez la salamandre, qui, pourtant considérée comme un organisme modèle connu pour ses capacités de régénération depuis plus de 200 ans, a été un peu oubliée. Ceci est peut être du à la complexité de son génome de l'ordre de 10^{10} bases, soit a peu près 10 fois plus gros que le génome humain, qui fait qu'aucun projet de séquençage n'a été envisagé jusqu'à présent. Pour *Notophtalamus viridescens* (triton vert à points rouges) seulement 164 transcrits annotés et 178 protéines sont disponibles dans la base NCBI. La reconstruction *de novo* du transcriptome et la validation du potentiel codant des transcrits par correspondance de peptides identifiés par spectrométrie de masse a donc été réalisée (Looso et al., 2013). Ces travaux ont confirmé que désormais, le transcriptome peut servir de base à l'identification de nouvelles protéines spécifiques des urodèles, par une stratégie PIT. Cette approche a déjà permis de définir un transcriptome dans des tissus particuliers et de déduire des familles de protéines inconnues jusqu'à présent impliquées dans la régénération tissulaire chez le triton (Looso et al., 2013).

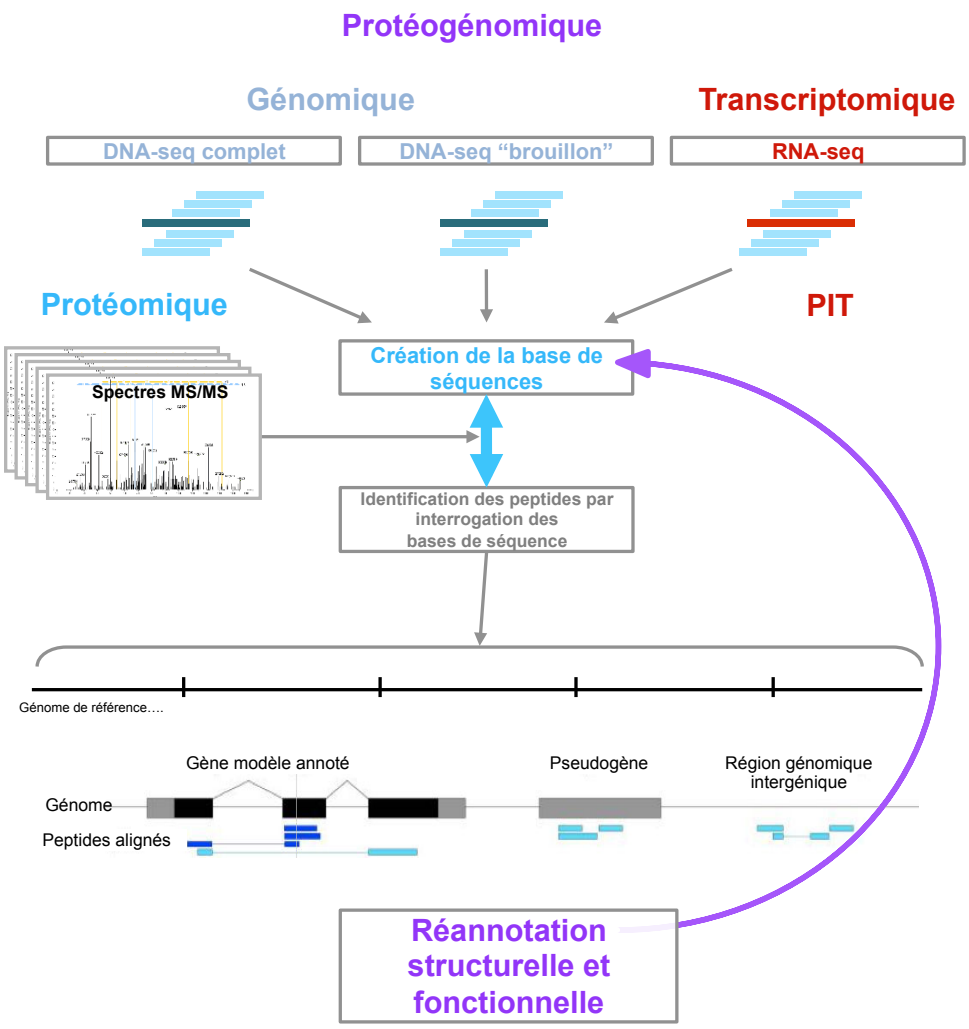

Figure 18. Principe de la protéogénomique

Typiquement, une traduction dans les six cadres de lecture est réalisée sur les acides nucléiques : séquence complète d'ADN ; "brouillon" du génome séquencé ou données de RNA-seq (transcriptome séquencé et reconstruit), pour générer une base de données de séquences protéiques et peptidiques théoriques. Les spectres MS/MS sont assignés aux séquences peptidiques de la base, et les protéines certifiées par MS peuvent être davantage caractérisées du point de vue de leur fonction et de leur structure, et implémenter / améliorer la base de données de séquences. Cette base peut être ensuite utilisée pour des stratégies de protéomique variées : exploratoire, ciblée, comparative. Inspiré de (Armengaud et al., 2014).

Le principal inconvénient de l'approche PIT est la grande taille de la base de données de transcrits résultant de l'assemblage des fragments séquencés. En effet, la grande redondance des fragments augmente l'incertitude nucléotidique et peut conduire à des erreurs d'assemblage qui se traduisent par des erreurs de séquence de protéines, changements de cadre de lecture pénalisants et arrêt prématuré des ORFs (stop dans les séquences protéiques). En outre, quand le sens de lecture des transcrits est inconnu (protocoles non « brin spécifiques »), une traduction dans les six cadres de lecture est nécessaire pour obtenir toutes

les séquences protéiques possibles, ce qui génèrera une très grosse base de données pour l'interrogation des spectres et induira une modification du « scoring » des peptides.

De nouvelles approches pour traiter ces bases de données sont actuellement en cours de développement, telle que celle proposée par Woo et collaborateurs (Woo et al., 2014). Ces derniers proposent, sur les données de *C.elegans;* la compression des données redondantes et la compilation des jonctions d'exons dans une base graphique des données d'épissage. Pour concevoir la base de données de séquences protéiques la plus appropriée à l'expérience, à partir des données de RNA-seq, les ARNms doivent représenter globalement l'espace de la protéine et donc, de préférence, les échantillons doivent être les mêmes à la fois pour l'analyse RNA-seq et l'analyse protéomique. Le développement des technologies de séquençage et les progrès de l'instrumentation, ainsi que des algorithmes d'identification à partir des données MS/MS et la construction de bases de données spécialisées présagent d'un avenir prometteur pour la protéogénomique. Le projet HPP (Human Proteome Projet) sur ses aspects de caractérisation chromosome centrique (C-HPP), profitera très certainement des avancées de la protéogénomique. En effet, l'intégration de techniques de séquençage des ARNs, telles que le séquençage des ARNms en cours de traduction ou « Ribosome-nascent chain complex mRNA » (RNC-mRNA) (Zhong et al., 2014), profiteront aux analyses protéomiques *via* la création de banques de séquences dédiées qui permettra de découvrir de nouveaux gènes s'exprimant spécifiquement dans un tissu ou dans un type cellulaire. La répartition des nouveaux évènements codants sur les 22 chromosomes et les chromosomes sexuels sera accessible. Au cours de cette thèse, deux groupes à l'origine de la ressource « The Human proteome Map » ont déjà tiré profit du couplage du RNA-seq et de la protéomique quantitative pour l'annotation du génome humain en découvrant des évènements codants nouveaux dont on pensait qu'ils ne codaient pas pour des protéines. L'équipe de Bernhard Küster (Wilhelm et al., 2014) a prouvé le caractère codant de 430 longs ARNs non codants intergéniques par la mise en évidence de leurs protéines, et le groupe de Akhilesh Pandey (Kim et al., 2014) a revu l'annotation de 808 gènes et prouvé le caractère codant de nombreux pseudogènes et ARNs non codants.

IV. Approches « Omiques » et spermatogenèse

Les études dans les domaines qualifiés de « Omiques » (néologisme se référant à « la totalité de… ») : génomique, transcriptomique et protéomique, ont eu un large apport dans notre compréhension de la spermatogenèse, et des larges jeux de données ont pu être générés depuis le début des années 2000. En effet, l'utilisation de la transcriptomique et la protéomique pour étudier la spermatogenèse est tout à fait appropriée, car le développement de gamètes mâle repose, comme nous l'avons vu dans cette introduction, sur une succession d'événements complexes et étroitement régulés. En outre, comme nous l'avons souligné précédemment, l'expression des ARNms et des protéines est particulièrement régulée et complexe, si bien que mesurer leurs niveaux d'expression dans un contexte testiculaire peut être trompeur sans une connaissance plus approfondie de l'expression des différents isoformes et produits de gènes spécifiques des cellules germinales ou d'un stade de la spermatogenèse en particulier. Il ne faut pas oublier par ailleurs que les ARNs non codants jouent un rôle particulièrement important, connu pour les sncRNAs, et très probable pour les lncRNAs, dans le processus de différenciation des cellules germinales mâles. De surcroît, démêler cette complexité par des approches de culture cellulaire est problématique en raison de la difficulté à cultiver les cellules les plus hautement différenciées de la lignée germinale mâle. Or, différentes approches dans les domaines de la transcriptomique et de la protéomique peuvent générer des clichés très précis des réseaux moléculaires séquentiellement impliqués dans le maintien de la spermatogenèse. Croiser ces approches s'avère alors absolument bénéfique pour la formulation d'hypothèses avant leur validation *in vivo* pour l'étude de la spermatogenèse.

A. Transcriptomique et étude de la spermatogenèse

Les progrès de la biologie moléculaire et de la génomique ont permis d'améliorer notre connaissance de la spermatogenèse, en permettant l'identification d'un grand nombre de gènes essentiels pour le développement des gamètes mâles fonctionnels (Matzuk et Lamb, 2002; de Rooij et de Boer, 2003). Depuis les années 2000, de très nombreux travaux d'analyse à grande échelle de la fonction testiculaire, reposant sur les rapides progrès dans le séquençage du génome et le développement de puces à ADN, ont conduit à l'identification de centaines de gènes spatialement et temporellement régulés au cours de l'ontogenèse du testicule (pour revue, voir (Wrobel et Primig, 2005)). Un grand nombre de gènes à la régulation spatiale et

temporelle au cours de l'ontogenèse des testicules, et essentiels pour le développement des gamètes mâles fonctionnels, ont pu être identifiés grâce également aux progrès dans l'utilisation de souris transgéniques (pour revues (Rolland et al., 2008; Calvel et al., 2010; O'Shaughnessy, 2014)).

Une analyse utilisant des puces à oligonucléotides (GenChip) a été réalisée pour l'étude du programme transcriptionnel de la méiose chez la souris, le rat et l'homme en utilisant des cellules testiculaires purifiées et des gonades entières. Le programme différentiel d'expression testiculaire chez les rongeurs et le transcriptome conservé de la méiose chez la souris, le rat et l'homme a été rapporté. Un groupe de 357 loci ayant un profil d'expression dans les cellules méiotiques /post méiotiques, qui est conservé entre les trois espèces a été mis en évidence (Chalmel et al., 2007a). Parmi ceux-ci certains sont critiques pour l'accomplissement de chacune des étapes de la spermatogenèse (Chalmel et al., 2007b). Peu après, Chalmel et collaborateurs ont établi les profils d'expression de gènes à l'échelle du génome chez le rat, la souris et l'homme, donnant des indices sur la machinerie régulatrice de la transcription dirigeant l'expression de gènes cibles au cours de la méiose. Ils ont mis en évidence les éléments régulateurs sur les promoteurs des gènes co-régulés. Plusieurs milliers de gènes ont été trouvés différentiellement exprimés entre les cellules de Sertoli et les cellules germinales. L'expression de ces gènes peut être plutôt somatique avec un profil d'expression similaire dans les contrôles somatiques, les cellules de Sertoli et dans les spermatogonies, tels que Lip1, Maged1, Maged2, par opposition aux gènes préférentiellement exprimés dans les cellules germinales méiotiques tels que Adam2 et Tdrd1, ou dans les spermatides (Cage1, Mageb5, Ssx2ip) (Chalmel et al., 2007b). Ces études fournissent un plan de l'expression testiculaire des gènes au niveau du type cellulaire. Une toute nouvelle étude transcriptomique de la différenciation spermatogoniale, de la division et de la méiose chez la souris suggère que certains gènes normalement exprimés à la méiose sont exprimés et traduits avant que les cellules germinales n'entrent en méiose (Evans et al., 2014). Cette étude a permis de dégager certains gènes candidats Asf1b et Esyt3 potentiellement impliqués dans la réorganisation de la chromatine spermatogoniale, les interactions cellules germinales/cellules de Sertoli et la formation de la BHT.

Jusqu'à récemment, l'analyse transcriptomique basée sur les puces à oligonucléotides très efficaces pour produire une vision globale de l'expression du génome se heurtent en revanche à la régulation des gènes au sein des cellules germinales, tant au niveau transcriptionnel que de la traduction (Eddy, 2002). Des décalages sont observables entre l'expression des

transcrits et des protéines ayant d'ailleurs été rapportés au cours de l'étude précédemment citée (Chalmel et al., 2007b). L'interprétation de la fonction d'un gène ne peut donc pas se résumer à l'étude du profil d'expression de son transcrit. Aussi, les analyses sur puces à oligonucléotides reposent sur les données d'annotation des génomes. L'analyse des régions codantes mais non annotées est donc impossible par ces approches. Or, ces régions subsistent sur des génomes pourtant déjà séquencés et annotés, tels que celui de l'homme et du rat.

Cette période pourrait être révolue car désormais les technologies du séquençage de l'ARN (RNA-seq) permettent de réaliser des études d'expression du génome entier, et peuvent faire progresser notre compréhension de la spermatogenèse normale et pathologique. Récemment, une étude importante du transcriptome de la spermatogenèse chez la souris a conduit à l'identification de plus d'un millier de nouveaux gènes méiotiques et de 5000 nouvelles isoformes potentielles de protéines (Margolin et al., 2014). Dans une autre étude importante utilisant le RNA-seq sur 26 tissus humains comparés au testicule, Djureinovic et collaborateurs (Djureinovic et al., 2014) ont classifié 2.050 gènes potentiels selon leurs profils d'expression. Ils montrent que de loin, le testicule est le tissu qui a le plus fort taux de gènes spécifiques. Plus de 1000 gènes sont spécifiques du testicule soit beaucoup plus que dans tous les autres tissus. Ces auteurs ont établi un « Top 50 » des gènes les plus exprimés au cours de la spermatogenèse et observé que 62 % des gènes identifiés « hautement enrichis » dans le testicule, étaient peu ou non caractérisés. L'expression des protéines correspondantes dans l'épithélium séminifère a été validée par immunohistochimie pour un certain nombre d'entre eux (Djureinovic et al., 2014).

De nombreuses études ayant exploré le transcriptome de cellules testiculaires isolées avec les nouvelles technologies de séquençage des ARNs ont mis en évidence des milliers de transcrits non annotés (Gan et al., 2013; Laiho et al., 2013; Soumillon et al., 2013; Djureinovic et al., 2014; Chalmel et al., 2014; Margolin et al., 2014; Meikar et al., 2014). Etant donné le manque d'informations les concernant, ils sont nommés longs ARNs non codants (lncRNAs) (ENCODE Project Consortium, 2004). Ces lncRNAs ont été trouvés s'accumulant aux stades méiotique et post méiotique de la spermatogenèse (Chalmel et al., 2014). Des petits ARNs non codants, piRNAs, bien connus pour avoir un rôle dans le spermatogenèse, ont été identifiés en même temps que des ARNms et des longs ARNs non codants dans le corps chromatoïde, à l'aide de techniques de séquençage (Meikar et al., 2014). Cette dernière étude nous éclaire sur le contenu de ces corps chromatoïdes pendant les étapes de différenciation des spermatides.

La fonction de ces transcrits nouvellement identifiés reste inexpliquée à ce jour. Il est donc plus que jamais fructueux de coupler les approches reposant sur le séquençage *de novo* des transcrits avec des approches de protéomique pour la découverte d'évènements codants au cours de la spermatogenèse, afin de nous éclairer sur le rôle de ces transcrits mystérieux et leur implication potentielle dans la régulation de l'expression protéique. De plus, nous l'avons vu, l'interprétation de la fonction d'un gène ne peut absolument pas se résumer à l'étude du profil d'expression de son transcrit au cours du processus de différenciation des cellules germinales mâles. Il est donc essentiel d'intégrer les approches « Omiques » pour l'étude de la spermatogenèse.

B. La protéomique pour comprendre la spermatogenèse

Notre équipe s'est attachée depuis le début des années 2000 à décrypter le protéome des cellules germinales testiculaires chez le rat afin d'identifier de nouvelles protéines germinales qui pourraient jouer un rôle important dans la spermatogenèse normale ou pathologique. Les analyses du protéome des spermatogonies de rat ont été réalisées par électrophorèse bidimensionnelle (2DE) combinée à la spectrométrie de masse (Guillaume et al., 2000; Com et al., 2003). Ces études ont permis de caractériser un répertoire de protéines exprimées par les spermatogonies et d'en tirer bénéfice en étudiant plus tard la distribution spatio-temporelle de quelques unes de ces protéines : TCTP (Guillaume et al., 2001b); Stathmin (Guillaume et al., 2001a) et MCM7 (Com et al., 2006) au sein de l'épithélium séminifère. Une stratégie plus intégrée a ensuite été utilisée, dans laquelle les niveaux de protéines particulières ont été comparés entre les différentes catégories de cellules germinales purifiées. Une analyse du protéome différentiel de la spermatogenèse chez le rat basée sur la technologie DIGE (Rolland et al., 2007) a été effectuée. Cette étude a permis d'identifier 123 protéines dont l'abondance relative diffère significativement entre les trois types de cellules germinales à différents stades de maturation et a conduit à l'identification de nouvelles protéines. L'une de ces protéines est la CLPH. Cette protéine cytoplasmique spécifique des spermatides et des corps résiduels est associée à la membrane interne mitochondriale aux derniers stades de différenciation des spermatides. C'est une protéine conservée chez les mammifères uniquement. Cette protéine s'est révélée être une protéine désordonnée dotée d'une forte capacité à lier le calcium. Elle est capable d'être phosphorylée par la caséine kinase 2 qui est une enzyme cruciale pour l'élongation des spermatides. C'est pour cette

raison qu'elle est considérée comme importante pour les réarrangements que subissent les cellules germinales de la spermiogénèse (Calvel et al., 2009).

D'autres équipes ont réalisé des études similaires pour déchiffrer les modes d'expression des protéines dans les cellules germinales isolées ou tout au long de la spermatogenèse, notamment dans les modèles de rats et de souris, mais aussi dans d'autres espèces (pour revue, voir (Calvel et al., 2010; Chocu et al., 2012; Macleod et Varmuza, 2013)). Parce que les approches basées sur la 2DE comportent des limitations, les études plus récentes sur la spermatogenèse ont fait usage de la démocratisation et de la résolution de la spectrométrie de masse en tandem, en adoptant les approches de type Shotgun LC-MS/MS. Parmi celles-ci, les travaux de Huang et Sha (Huang et Sha, 2011) ont ainsi utilisé la protéomique Shotgun pour définir les profils d'expression de protéines spécifiques des cellules germinales tétraploïdes et haploïdes de testicules de souris adultes purifiées par cytométrie de flux. Ils ont détecté plus de 3500 protéines dans les cellules germinales tétraploïdes, 216 qui ont été montrées pour avoir des homologues chez la levure, connus pour être impliqués dans la méiose. Les autres études citées ci-après et qui utilisent cette technologie ont été très informatives pour l'étude de la spermatogenèse. Chez le macaque, cette approche a permis d'identifier 9.078 protéines (correspondant à 8.662 gènes) dans le testicule. Parmi eux, 3.010 gènes furent validés au niveau de l'expression de la protéine pour la première fois (Wang et al., 2014a). Toujours chez le macaque, Skerget et collaborateurs (Skerget et al., 2013) ont caractérisé 1.247 protéines du spermatozoïde, dont 10% sont sous exprimées dans le testicule, ce qui reflète une spécificité acquise pendant leur maturation dans l'épididyme. Trois de ces protéines ADAMS (A-Disintegrin and Metalloprotease proteins), ADAM18-, 20- and 21-like, semblent ne pas être conservées au cours de l'évolution jusqu'à l'humain. Chez l'homme, un certain nombre de chemokines et de facteurs de croissance ont été mis en évidence par ce même type d'approche dans le sécrétome des cellules péritubulaires, cellules considérées comme contribuant au maintien de la niche des cellules souches spermatogoniales *via* la sécrétion de protéines (Flenkenthaler et al., 2014). Ces auteurs ont caractérisé 660 protéines du sécrétome des cellules péritubulaires par LC-MS/MS, parmi lesquelles 263 sont sur-représentées dans le milieu conditionné comparé aux lysats de cellules péritubulaires. Des analyses utilisant les termes de la Gene Ontology et des prédictions de peptide signal ont été réalisées, et permettent d'obtenir des éléments supplémentaires en faveur du caractère sécrété de certaines de ces protéines. Une étude récente par LC-MS/MS a été réalisée dans notre laboratoire sur du plasma séminal humain, fluide provenant de sécrétions à partir du testicule, de

l'épididyme, des vésicules séminales et de la prostate, et nécessaire au transport et à la nutrition des gamètes mâles. Elle a permis d'identifier 2.545 protéines et de caractériser plusieurs marqueurs fonctionnels dont les gènes correspondants sont préférentiellement exprimés dans le testicule (83 gènes), l'épididyme (42), les vésicules séminales (7) et la prostate (17). Des marqueurs potentiels ont été identifiés dont TKTL1, LDHC et PGK2, qui pourraient être utilisés pour faire la distinction entre le sperme des hommes fertiles et infertiles (Rolland et al., 2013). Une autre étude récente utilisant cette fois la 2DE et électrophorèse 1D couplée avec la LC-MS/MS rapporte l'identification de 7.346 protéines à partir de biopsies de testicule humain, et la découverte d'un ensemble de protéines exprimées dans le testicule et associées au cancer (TMPRSS12, TPPP2, PRSS55, DMRT1, PIWIL1, et HEMGN). Les auteurs ont vérifié leur expression dans le testicule à l'aide d'anticorps disponibles *via* le "Human Protein Atlas". Ils ont par ailleurs développé Human Testis Proteome database (HTPD : http://reprod.njmu.edu.cn/htpd/) afin de rendre leurs données disponibles pour la communauté (Liu et al., 2013).

Cette liste d'études n'est pas exhaustive. Elle vient en complément de la revue intitulée « La protéomique, un outil puissant pour comprendre la spermatogenèse normale et pathologique » (Chocu et al., 2012), jointe à ce manuscrit. Les études protéomiques citées utilisent une variété de stratégies de protéomique et de validations biochimiques de protéines d'intérêt qui permettent de répondre à des questions biologiques variées et importantes concernant la spermatogenèse. Les résultats obtenus à ce jour ont d'ailleurs fourni de précieuses informations et permis l'amélioration de notre compréhension des mécanismes régissant la spermatogenèse. Toutefois, les stratégies sur lesquelles elles sont fondées souffrent de limitations expérimentales et / ou d'analyse : la qualité de l'échantillon (étant donné que les cellules germinales isolées subissent rapidement l'apoptose après leur isolement ce qui peut altérer l'analyse protéomique), et le temps requis pour la collecte des données.

C. Génomique intégrative et spermatogenèse.

Bien que la protéomique et la transcriptomique soient désormais largement équivalentes en termes de cohérence des résultats, de toute évidence, la diversité des protéines ne peut pas être caractérisée seulement par l'analyse de l'expression des gènes. En effet, la complexité bien connue des mécanismes de régulation de l'expression des gènes chez les mammifères est en partie responsable des écarts souvent signalés entre l'abondance des ARNms et celles des protéines (Schwanhausser et al., 2011). De tels écarts sont particulièrement répandus dans la spermatogenèse, et la corrélation entre les taux de protéine et d'ARNm est plus faible dans le testicule que dans d'autres tissus (Cagney et al., 2005). De nombreux gènes présentent un retard apparent de la traduction par rapport à la transcription pendant la spermiogénèse (pour revue, voir (Eddy, 2002)). La combinaison des approches protéomiques et transcriptomiques est alors fructueuse pour améliorer notre connaissance de la spermatogenèse. Une telle intégration des approches est réalisée dans l'étude de Govin et ses collègues qui ont utilisé une stratégie protéomique basée sur de la chromatographie TP (transition protéines) permettant d'isoler les protéines nucléaires, suivie d'une analyse en spectrométrie de masse pour étudier le protéome nucléaire de cellules germinales mâles post-méiotiques (spermatides en allongement et allongées) chez la souris. Ces auteurs ont ensuite comparé leurs données de protéomique avec celles du transcriptome de tissu normal de la souris ou des cellules germinales mâles disponibles à partir du référentiel GEO (Gene Expression Omnibus), puis effectué des analyses fonctionnelles basées sur les termes de la Gene Ontology. Ils ont produit une liste de facteurs impliqués dans l'empaquetage et la programmation du génome mâle post-méiotique (Govin et al., 2012). Plus récemment, une étude importante utilisant des approches transcriptomiques et protéomiques quantitatives intégrées a été utilisée pour évaluer des mécanismes de régulation post-transcriptionnelle au cours de la différenciation de cellules germinales mâles (Gan et al., 2013). Leur approche protéomique basée sur l'iTRAQ et la LC-MS/MS leur a permis d'identifier 2.008 protéines dans les spermatogonies, les spermatocytes pachytène, les spermatides rondes et les spermatides allongées de souris. Ils ont pu faire ressortir en corrélant leur profil d'expression protéique (par iTRAQ) à celui de leur transcrit (par l'analyse de résultats de microarrrays) au cours des différents stades de la spermatogenèse. D'abord, quatre groupes des protéines différentielles ont été identifiés reflétant 1) l'amplification mitotique des spermatogonies, 2) des évènements pré-méiotiques dans les spermatocytes, 3) la régulation des ARNms dans les jeunes spermatides et 4) la

différentiation des spermatides en allongement. Puis cinq groupes de transcrits / protéines qui évoluent de manière dynamique différente ont été montrés. Cinq mécanismes majeurs de régulation des transcrits ont ainsi été mis en évidence en fonction des variations du niveau de la protéine par rapport à l'évolution des taux d'ARNm. Ces cinq groupes sont nommés **1) « transcrit seul »** pour lesquels le niveau d'expression de la protéine est fonction du niveau de son transcrit, **2) « répression de la traduction »** pour lesquels le niveau du transcrit augmente plus que le niveau de la protéine, **3) « dégradation du transcrit»,** pour lesquels les niveaux de transcrits chutent par rapport aux niveaux d'expression de la protéine, **4) « dé-répression de la traduction** » pour lesquels le niveau d'expression de la protéine augmente plus que celui du transcrit et **5) « dégradation de la protéine »** pour lesquels le niveau de la protéine baisse drastiquement par rapport au niveau de transcrit pendant la spermatogenèse. Ces auteurs constatent par ailleurs que ces mécanismes de régulation post-transcriptionnels sont liés à la génération de piRNAs et de transcrits anti sens. Ils ont apporté des éléments de compréhension des mécanismes de régulation post-transcriptionnels de l'expression génique dans la spermatogenèse chez les mammifères, et fournissent un inventaire précieux des protéines produites au cours de la spermatogenèse chez la souris (Gan et al., 2013). Cette étude démontre bien la puissance des approches « Omiques » intégrées pour l'étude de la spermatogenèse chez les mammifères.

Nombre d'auteurs d'études protéomiques sur la spermatogenèse croisent leur données fraîchement générées avec des données de transcriptomique disponibles, comme les bases d'EST par exemple chez le macaque (Wang et al., 2014a), ou les données de puces à oligonucléotides comme l'ont fait Gan et collaborateurs (Gan et al., 2013), ainsi que Rolland et collaborateurs (Rolland et al., 2013) dans l'étude intégrative consistant à trouver des marqueurs fonctionnels pour tous les organes participant à la production du plasma séminal. De nouvelles perspectives sur les événements moléculaires au cours de la spermatogenèse, et par extension sur l'explication de troubles de la reproduction humaine, sont désormais ouvertes. En effet, une grande série de jeux de données «omiques» du testicule a été générée, collectée, et peut être rendue disponible pour les chercheurs de la discipline (Com et al., 2014). Il existe une compilation de données de transcriptomique rassemblant un grand nombre d'études transcriptomiques dans le domaine de la reproduction: la base de données GermOnline (Lardenois et al., 2010), (http://www.germonline.org), qui est un lien rassemblant des études pertinentes pour le cycle cellulaire, la gamétogenèse et la fertilité. Cette base incorpore un navigateur inter espèces permettant de fournir des annotations, des

séquences d'ADN, des relations d'évolution et des annotations fonctionnelles. Une autre base plus récente rassemble maintenant les données de plusieurs études transcriptomiques et de nos études protéomiques récentes disponibles dans un système référentiel également orienté sur la reproduction : le ReproGenomicsViewer (RGV), (http://rgv.genouest.org/). Les données y sont représentées graphiquement sur un navigateur inspiré de l'UCSC genome browser (Meyer et al., 2013). RGV permet de visualiser chez le rat des données récentes de RNA-seq des cellules testiculaires isolées (Chalmel et al., 2014) et des données issues d'expériences antérieures d'expression sur différents types cellulaires testiculaires et sur différents tissus chez le rat (données de Exon Array). Des données sont aussi disponibles dans ce navigateur pour l'homme et la souris. Ces outils sont utiles pour l'aide à la décision et l'élaboration d'hypothèses en biologie de la reproduction, tout comme les bases de données biologiques brièvement présentées dans la partie traitant des bases de données, telles que KEGG pathways, MINTAct, BioGriD ainsi que SPD (Secreted Protein Database). Ces dernières sont une aide à la formulation d'hypothèses en amont de validations « wet lab » (études fonctionnelles d'interactions protéiques, validation de l'expression des transcrits et des protéines) et d'études *in vivo*.

OBJECTIFS

Ces travaux de thèse ont eu pour objectif d'identifier des protéines de la lignée germinale, qui de par leur niveau d'expression et leur expression différentielle aux différents stades de différenciation des cellules germinales mâles pourraient jouer un rôle dans le déroulement de la spermatogenèse normale. Ces travaux se sont divisés en quatre projets permettant d'aller vers cet objectif principal, et qui seront présentés respectivement dans les quatre chapitres de ce manuscrit:

Des transcrits testiculaires non annotés et des lncRNAs ont été montrés en proportion importante dans les cellules méiotique et post-méiotique de rat par Chalmel et collaborateurs (Chalmel et al., 2014). Or, seule une approche intégrative faisant appel à la protéomique permet de trancher entre des évènements transcrits inconnus mais qui codent pour des protéines, et des transcrits non codants exprimés de manière stade-spécifique au cours de la spermatogenèse. C'est dans ce but que nous avons décidé de nous focaliser sur la recherche de nouveaux évènements codants différentiellement exprimés au cours des stades méiotiques et post-méiotiques de la spermatogenèse, par une approche de protéomique qui intègre les données de la transcriptomique par RNA-seq. La première étude présentée intitulée « Découverte de nouveaux loci codants, par une approche de type protéomique informée par la transcriptomique (PIT) dans les cellules germinales de rat » a donc eu pour objectif de valider le potentiel codant de transcrits testiculaires non annotés, et d'identifier de nouvelles protéines / de nouveaux gènes, de la lignée germinale potentiellement impliqués dans la spermiogénèse. Ce travail sera présenté dans le premier chapitre de ce manuscrit. Une deuxième étude sera présentée en continuité de ce travail, et qui traitera de la découverte de nouvelles isoformes spécifiques des protéines germinales par la même approche PIT. L'objectif de ce deuxième projet intitulé « Découverte de nouvelles isoformes spécifiques des cellules germinales chez le rat » a eu pour objectif de mettre en évidence de nouvelles isoformes de protéines connues, spécifiques des stades méiotique et post-méiotique, potentiellement impliqués dans la spermiogénèse. Ce travail sera présenté en un second chapitre de ce manuscrit.

Le dialogue entre les cellules germinales et les cellules de Sertoli essentiel à la spermatogenèse est encore mal compris et met potentiellement en jeu des protéines exprimées à la membrane des cellules germinales au cours de la différentiation. L'analyse protéomique différentielle des protéines membranaires dans les cellules méiotiques et post-méiotiques, et dans les corps résiduels chez le rat a eu pour objectif d'identifier des protéines membranaires dont l'expression est différentielle aux stades méiotique, post-méiotique et dans les corps résiduels, potentiellement impliquées dans le dialogue membrane/membrane entre les cellules de Sertoli et les cellules germinales et qui de par leur expression stade-spécifique seraient importantes pour le spermatogenèse. Ce travail sera présenté dans le troisième chapitre de ce manuscrit.

Il est connu que la phagocytose des corps résiduels déclenche une nouvelle vague spermatogénétique juste après la spermiation, mais ce mécanisme est encore mal connu. Le devenir du corps résiduel au sein de la cellule de Sertoli est également mal connu. Afin d'identifier des protéines susceptibles d'avoir un rôle dans le devenir des corps résiduels, nous avons réalisé le protéome du corps résiduel et celui des cellules de Sertoli en utilisant une approche protéomique shogun itérative. Avec pour objectif de mieux comprendre les mécanismes en jeu dans la formation du corps résiduel, sa phagocytose et son devenir dans la cellule de Sertoli, nous avons initié l'étude intitulée « Identification de protéines impliquées dans le devenir des corps résiduels par protéomique shotgun chez le rat » qui sera présentée dans le quatrième chapitre de ce manuscrit.

MÉTHODOLOGIES

Dans cette partie, les méthodologies de protéomique utilisées au cours de mes travaux de thèse sont décrites, et les étapes communes et spécifiques à chaque projet sont précisées. Les procédures expérimentales utilisées dans chaque projet sont plus détaillées dans les sections « Matériels et méthodes » de l'article 1 et de l'article 2.

I. Préparation des cellules germinales et des corps résiduels

Les animaux sur lesquels les cellules germinales ont été prélevées sont des rats mâles Sprague-Dawley de plus de 90 jours provenant de l'élevage Janvier (Le Genest Saint-Isle, France). Les rats ont été sacrifiés par asphyxie au gaz carbonique, puis les testicules ont été prélevés pour l'isolement des spermatocytes pachytènes, des jeunes spermatides et des corps résiduels par le principe de l'élutriation centrifuge, selon la méthode décrite par Pineau et collaborateurs, mais en utilisant la dissociation mécanique des tubules et non la digestion trypsique (Pineau et al., 1993). La suspension de cellules germinales est préparée comme suit. L'artère testiculaire est perfusée au PBS et l'albuginée est retirée. Les tubules séminifères sont ensuite soumis à dissociation mécanique au scalpel pendant 10 minutes. Le surnageant riche en cellules germinales est filtré sur membrane de nylon avec des pores de100µm en présence de DNAse à 1µg/mL. Puis le filtrat est filtré sur laine de verre pour éliminer les spermatozoïdes. Ce filtrat est centrifugé 10 minutes à 100xg à 4°C. Les culots sont lavés trois fois dans du tampon d'élutriation (PBS complémenté avec CaCl2 à 0,8mM; MgCl2 à 0,5mM, BSA à 5g/L; glucose à 1,6mM; pyruvate de sodium à 7mM) puis filtrés sur membrane de nylon avec des pores de 20µm en présence de DNAse. La suspension de cellules germinales totales subit ensuite une élutriation centrifuge à 4°C dans du tampon d'élutriation. L'élutriation centrifuge permet la séparation de différents types cellulaires en fonction de leur taille (Onada et al, 1991). La suspension cellulaire germinale totale est introduite dans une chambre d'élutriation associée au rotor JE-6B (Beckman Coulter), dans la centrifugeuse J2-21 B (Beckman Coulter). Cette technique permet la séparation des cellules essentiellement en fonction de leur taille, car elles sont soumises à deux forces opposées dans la chambre d'élutriation. Elles sont d'une part la force centrifuge exercée par la rotation du rotor qui tend à entraîner les cellules vers le fond de la chambre, et la force centripète exercée par le flux du tampon d'élutriation qui tend à chasser les cellules hors de la chambre. La modification du

débit à l'aide d'une pompe péristaltique et de la vitesse de centrifugation permettent une sortie sélective des différentes populations de cellules. Pour une suspension de cellules hétérogènes, l'augmentation du débit permet la sortie sélective des cellules des plus petites aux plus grandes. Les différentes populations de cellules germinales sont séparées selon les conditions décrites dans le tableau 3.

Débit (mL/min)	Volume (mL)	Fraction	Pureté
1ère élutriation à 2000rpm			
26	250	2ème élutriation	-
31	100	Spermatocytes pachytènes	86%
36	100	Spermatocytes pachytènes	95%
45	125	Spermatocytes pachytènes	89%
2ème élutriation à 2500rpm			
13	250	3ème élutriation	-
25	100	Spermatides rondes	84%
3ème élutriation à 3360rpm			
22 ,5	200	Corps résiduels	>80%

Tableau 3. Conditions d'élutriation des cellules germinales de rat adulte.

Les fractions d'intérêt sont centrifugées 10 minutes à 1.000g à 4°C, puis lavées trois fois au PBS. Les pourcentages (tableau 3) correspondent à la pureté des fractions évaluée par une analyse du contenu en ADN des différentes fractions de cellules germinales par cytométrie de flux (Pineau et al., 1993). Les culots cellulaires sont conservés à -80°C en vue d'une analyse protéomique.

II.Préparation des échantillons en vue des analyses protéomiques

A. Préparation de protéines

a) Extraction des protéines totales

En vue d'une analyse protéomique Shotgun, nous avons préparé des extraits protéiques à partir de culots cellulaires de cellules germinales (50 millions de spermatocytes pachytène,

100 millions de spermatides rondes) et de corps résiduels (300 millions). Les culots de cellules germinales sont repris dans du tampon d'extraction (PIPES 100mM; NaCl 70mM; $MgCl_2$ 2mM; ph7,4) additionné extemporanément de cocktail d'inhibiteurs de protéases (EDTA 1mM; DTT 0,5mM; AEBSF 1mM et E-64 10µM); de nucléase à 0.6 U/mL et de Nonidet-P40 à 2%. Les suspensions cellulaires sont soumises à sonication sur glace (6 pulses de 10s à 40% d'amplitude) avec 30s d'attente entre chaque phase de sonication, afin d'éviter la dégradation des protéines. Les suspensions cellulaires sont ensuite laissées 1 heure sur la glace. Les lysats sont centrifugés à 1.000 g pendant 10 min à 4°C pour éliminer les débris et les cellules non lysées. Les surnageants sont ensuite centrifugés à 105.000g (ultra-centrifugeuse Sorvall Discovery M120 SE, rotor S120AT 3 dans des tubes en polyallomer (Beckman); pendant 1 h à 4°C. Les surnageants contenant les protéines solubles sont conservés à 4°C, et les culots contenant les protéines membranaires sont repris dans une solution de Na_2C0_3 (100mM), puis soumis à sonication dans les mêmes conditions que précédemment. Ces suspensions sont ensuite centrifugées à 105.000g pendant 45 minutes à 4°C. Les surnageants sont rassemblés avec les protéines solubles de la première extraction. Les protéines sont dosées par la méthode colorimétrique de l'acide bicinchonique (BCA) puis stockées à -80°C en vue d'une analyse en LC-MS/MS.

a) Extraction des protéines membranaires

Pour l'extraction des protéines membranaires, les culots de cellules (50 millions de spermatocytes, 100 millions de spermatides rondes) sont repris dans le tampon d'extraction (PIPES 100mM ; NaCl 70mM ; $MgCl_2$ 2mM ; pH7,4), additionné de nucléase et du même cocktail d'inhibiteurs de protéases que pour les protéines totales, mais sans ajout de détergent. Les culots de corps résiduels (300 millions) sont repris dans un tampon d'extraction Tris (Tris 10mM; $MgCl_2$ 10mM; pH 7,4). Les lysats sont réalisés par sonication, et les débris éliminés de la même façon que décrit pour les protéines totales. Une première centrifugation à 105.000g permet de séparer les protéines solubles des protéines membranaires. Les culots de protéines membranaires sont lavés dans une première solution (Na_2C0_3 100mM), puis soumis à sonication, centrifugés à 105.000g pendant 45 minutes à 4°C, et lavés dans une deuxième solution saline (Na_2C0_3 100mM; Na Cl 1M), soumis à sonication et centrifugés à 105.000g à 4°C, 45 minutes. Les culots sont finalement repris dans du tampon adapté au marquage ICPL (6M guanidine HCl; pH8,5), les protéines y sont dosées et ces préparations sont conservées au –80°C.

b) Marquage des protéines membranaires

Pour quantifier de manière relative les protéines membranaires des spermatocytes pachytène (pSPC), des spermatides rondes (rSPT) et des corps résiduels (CR), nous avons utilisé un marquage de type isotopique, le marquage ICPL, qui se fait sur les résidus lysine et N-termini des protéines entières. Après réduction des ponts disulfure avec du tris(2-carboxyethyl)phosphine 0.2 M, et alkylation des cystéines avec de l'iodoacétamide 0,4mM, 50µg de protéines de chaque extrait ont été marqués avec les réactifs ICPL. Ce marquage a été réalisé avec des étiquettes de masse légère, le ^{12}C-nicotinoyloxysuccinimide (ICPL_0); lourde, le ^{13}C-nicotinoy-loxysuccinimide (ICPL_6 avec 6^{13}C); et super lourde, le ^{13}C-^{2}D-nicotinoyloxysuccinimide (ICPL_10 avec 6^{13}C et 4 deutériums); sur les extraits membranaires des pSPC, rSPT et CR de manière inversée dans trois triplex techniques comme indiqué dans le tableau de la Figure 19. Pour un triplex, les protéines des trois échantillons marquées différemment sont séparées par électrophorèse monodimensionnelle classique en présence de SDS sur gel à 12% de polyacrylamide (SDS-PAGE). Les protéines sont digérées dans le gel et les peptides sont extraits comme décrit dans la prochaine section, mais sans alkylation des cystéines.

B. Préparation d'extraits peptidiques

Les protéines dénaturées par chauffage et réduites par du DTT 50mM sont séparées par éléctrophorèse monodimensionnelle classique en présence de SDS sur gel à 12% de polyacrylamide (SDS-PAGE). Pour chaque type cellulaire analysé, 100µg de protéines sont séparées sur gel. Pour l'analyse des protéines membranaires en protéomique différentielle, 50µg de protéines de chaque triplex contenant chacun les protéines de trois types cellulaires marquées et mélangées sont chargées sur une même piste du gel. Les bandes de gel sont découpées sur toute la longueur de la piste de migration des protéines visibles après coloration du gel au bleu de Coomassie. Chaque bande de gel est découpée en plus petits morceaux afin de faciliter la digestion trypsique des protéines dans le gel. Les morceaux de gel sont d'abord lavés pour éliminer les colorants et le SDS susceptibles d'inhiber l'activité de la trypsine, dans du bicarbonate d'ammonium (NH_4HCO_3) à 50mM dans l'eau avec 50% d'acétonitrile. Les morceaux de gel sont ensuite déshydratés avec de l'acétonitrile 100%, et réhydratés dans du NH_4HCO_3 100mM. Ils sont lavés et déshydratés ainsi une deuxième fois. Les ponts disulfure des protéines dans le gel sont ensuite réduits par du DTT 65mM à 37 °C, puis les cystéines sont alkylées en présence d'iodoacétamide 135mM à température ambiante

dans l'obscurité. Les morceaux de gel sont ensuite lavés dans du NH_4HCO_3 100mM dans l'eau avec 50% d'acétonitrile, puis déshydratés avec de l'acétonitrile 100%. Pour la digestion protéolytique, les morceaux de gel sont réhydratés par une solution enzymatique de trypsine à 12,5 ng/µL, à 37°C pendant la nuit. Les peptides trypsiques sont ensuite extraits du gel avec l'utilisation séquentielle des solutions d'acétonitrile 70%/acide formique 0,1%, puis d'acetonitrile 100%, et de nouveau d'acétonitrile 70% acide formique 0,1%. L'acétonitrile est évaporée avant l'analyse LC-MS/MS.

III. Analyses protéomiques

Le schéma général des analyses protéomiques menées au cours des différents projets est montré Figure 19. Trois approches protéomiques différentes sont utilisées. Une approche intégrative visant à découvrir de nouveaux évènements codants dans les cellules germinales. Elle fait appel à une banque de séquences protéiques personnalisée contenant des séquences traduites de transcrits de cellules testiculaires isolées de rat obtenues par RNA-seq lors d'une précédente étude (Chalmel et al., 2014). Ce travail est présenté dans le premier chapitre. Une autre approche de protéomique différentielle visant à quantifier de manière relative des protéines membranaires des cellules germinales fait appel à une technique de marquage isotopique (ICPL), dans une étude présentée dans le troisième chapitre. Une autre stratégie de protéomique systématique visant à établir le protéome le plus exhaustif possible des corps résiduels afin de mieux comprendre les mécanismes de leur formation, et leur devenir au sein des cellules de Sertoli, concerne le projet présenté dans le quatrième chapitre. Toutes ces approches ont en commun l'analyse LC-MS/MS décrite dans l'introduction générale, mais font appel à des stratégies différentes d'analyse des données de protéomique. L'analyse LC-MS/MS sera détaillée et les différences de stratégies entre les approches utilisées dans les différents projets seront exposées.

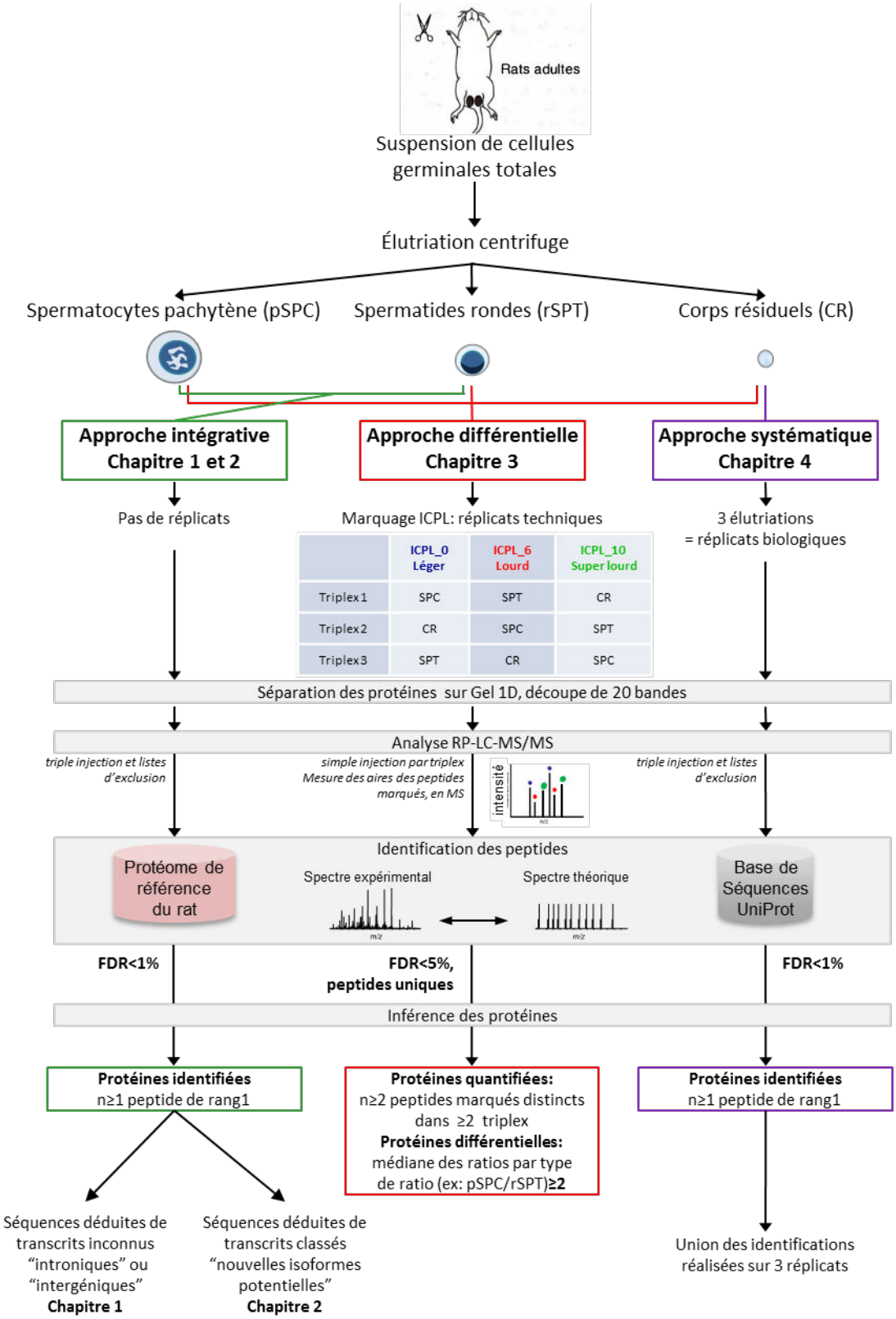

Figure 19. Schéma d'ensemble des différentes stratégies de protéomique utilisées

A. Acquisition des données par analyse LC MS/MS

Pour acquérir les données de spectrométrie de masse, un spectromètre de masse hybride constitué d'une trappe linéaire et d'une trappe orbitale (LTQ-Orbitrap XL, Thermo Fisher Scientific) a été utilisé. Celui-ci est couplé avec une nano chromatographie liquide haute performance (HPLC), (Ultimate 3000, Dionex Thermo Scientific) connectée au spectromètre de masse *via* une source nano electrospray (ESI), (New objective). Les échantillons à analyser sont maintenus à 8°C pour réduire leur évaporation et la colonne de séparation des peptides est maintenue à 30°C. La phase mobile A est composée de 99.9% d'eau pure et 0.1% d'acide formique, la phase mobile B est composée de 99.9% d'acétonitrile et 0.1% d'acide formique. Ces phases sont distribuées par une nano pompe en direction de la colonne HPLC à phase inverse (RP-HPLC) de type C18 (La phase stationnaire apolaire est donc composée de silice greffée de chaînes linéaires à 18 atomes de carbones). Les peptides sont adsorbés sur la phase stationnaire puis désorbés de la silice par un gradient d'acétonitrile. Dans un premier temps, les peptides sont retenus et concentrés sur la phase en tête d'une pré-colonne, les sels sont élués, puis les peptides sont entraînés vers la colonne analytique qui permettra ensuite leur séparation au fur et à mesure d'un gradient d'élution. Le gradient d'élution est composé d'Acétonitrile (phase B) de 2 à 35% pendant les 60 première minutes, de 35 à 60% pendant les minutes 60-85, et de 60 à 90% pendant les minutes 85-100. La colonne est ensuite lavée pendant 16mn avec de la phase B à 90%, et finalement remise à l'équilibre à 2% de phase B pendant 19mn dans le but de préparer l'injection de l'échantillon suivant. Les peptides arrivent donc au niveau de la source suivant un ordre croissant d'hydrophobicité. Un voltage de 1,5 kV induit le chargement des peptides en sortie de l'HPLC et l'évaporation du solvant les entourant. Ainsi seuls les peptides chargés pénètrent dans le spectromètre de masse. L'acquisition des spectres se fait en deux temps. Tout d'abord, un scan MS (masse et charge) est réalisé au niveau de l'Orbitrap tandis que le LTQ réalise la fragmentation des peptides par collision (CID) ainsi que leur détection (MS/MS). La gamme de masse de sélection des ions s'étend des m/z = 400 aux m/z = 2000 avec une résolution de 60000 à m/z=400. Dans le cas d'analyses en triple injection, l'objectif est d'analyser le même échantillon de peptide trois fois en LC-MS/MS afin d'optimiser l'identification de peptides. Une liste d'exclusion des m/z est établie entre chaque injection pour permettre au spectromètre de masse d'ignorer les peptides déjà identifiés à l'injection précédente.

B. Analyse des données

Les données acquises par le spectromètre de masse sont analysées avec le logiciel Proteome Discoverer 1.2 (Thermo Scientific). Ce logiciel permet l'interrogation des données de spectres *via* les moteurs de recherche SEQUEST et /ou MASCOT contre la base de séquences protéiques UniProt restreinte à l'espèce *Rattus Norvegicus* en ce qui concerne l'approche systématique, ou contre la base de séquences personnalisée « Protéome de référence du rat » pour les deux autres approches. L'objectif de cette étape est de caractériser les peptides et d'identifier les protéines à partir des spectres expérimentaux par comparaison avec les spectres de digestion théoriques obtenus à partir des séquences contenues dans ces bases. La base de séquences « Protéome de référence du rat » contient des séquences traduites à partir de transcrits qui ont été identifiés par séquençage à haut débit (RNA-seq) (Chalmel et al., 2014). Les séquences nucléotidiques des 99.438 isoformes de transcrits nouvellement assemblés dans cette étude ont ainsi été traduites dans les six cadres de lecture possibles. Les séquences d'au moins 10 résidus d'acides aminés entre deux codons stop ont été définies comme des séquences de protéines potentielles (environ 3 millions de séquences prédites). Ces séquences ont été fusionnées avec les bases UniProt (37.175 séquences de protéines et isoformes, version 10_2012) et Ensembl (32.971 séquences de protéines connues et 44.993 séquences de protéines prédites, version 3.4.68), pour constituer la banque de séquences personnalisée « Protéome de référence du rat ». Pour chaque expérience, les interrogations sont réalisées sur les données de spectre obtenues à partir des 20 extraits peptidiques correspondant à 20 bandes de gel en même temps (un triplex dans le cas de l'expérience différentielle). Dans le cas d'analyses en triples injections, pour générer les listes d'exclusion de masses, les interrogations sont faites « bande par bande » afin d'exclure les m/z pour le prochain passage de chacun des 20 échantillons dans le spectromètre de masse. Au final dans les expériences en triple injection, les interrogations sont faites sur l'ensemble des données de spectre.

Pour les expériences sans marquage des protéines, la sélectivité de l'enzyme a été fixée à une digestion complète à la trypsine avec un clivage manquant autorisé. Les modifications de masse peptidiques fixes prises en compte sont la carbamidométhylation des cystéines (dues à la préparation des échantillons), et les modifications variables sont les oxydations des méthionines, l'acétylation des lysines et des N-ter, et les phosphorylations des sérines, thréonine et tyrosines. Toutefois, pour les données de l'expérience différentielle avec un marquage ICPL, étant donné que la modification des résidus lysine par les étiquettes ICPL

empêche leur clivage par la trypsine, l'arginine C a été choisie comme enzyme avec un clivage manquant autorisé. Les modifications de masse variables prises en compte en plus sont le marquage des résidus lysine et du N-ter par les étiquettes de masse ICPL légères, lourdes, ou super-lourdes. La tolérance de masse des ions parents et fragments est fixée à 10 ppm et 0,5 Da, respectivement.

L'échantillon peptidique analysé par LC-MS/MS en triple injections permet d'augmenter le taux d'identification d'environ 20% par rapport à une simple injection. Après la première injection, une liste d'exclusion des m/z correspondant aux peptides détectés est établie afin d'ignorer ceux-ci et d'identifier de nouveaux peptides lors des injections suivantes.

Les peptides identifiés pris en compte dans l'analyse sont ceux ayant un score Mascot individuel au dessus du seuil d'identité tel que la p-value, probabilité que la correspondance spectre théorique /spectre expérimental soit due au hasard, soit inférieure à 0.05 (p-value < 0.05). Les peptides identifiés sont ensuite filtrés différemment selon les analyses. Ils sont filtrés sur la base de leur score afin d'obtenir un taux de faux positifs par rapport à des identifications sur une banque leurre fixé à 5% pour l'analyse différentielle ICPL, et fixé à 1% maximum pour les autres analyses (Figure 19). Les peptides partagés par plusieurs protéines sont pris en compte, et les protéines sont groupées automatiquement, puis rangées au sein d'un groupe en fonction de leur score, qui prend en compte leur couverture de séquence.

C. Traitement des données de protéomique

a) Dans l'approche intégrative PIT

Dans l'approche de protéomique informée par la transcriptomique (PIT) visant à découvrir des évènements nouveaux, il est important de pouvoir sélectionner le plus possible de protéines identifiées dans les types cellulaires étudiés (pSPC et rSPT). Nous n'avons donc pas imposé de filtres sur le nombre de peptides ayant permis d'identifier une protéine. En effet, les sélections strictes des candidats potentiels selon des critères génomiques et transcriptomiques qui suivent la détection en spectrométrie de masse des protéines éliminent un certain nombre d'évènements correspondant à de fausses identifications. En revanche, seulement les peptides correspondant à des correspondances de rang 1 (meilleure

correspondance entre un spectre expérimental et un spectre théorique) sont conservés dans l'analyse. Le taux de faux positifs par rapport à une base leurre est fixé à 1% (FDR1%) pour ne garder que les peptides de « haute confiance ». Pour la recherche de nouveaux loci codants exprimés dans les cellules méiotiques et post-méiotiques (étude présentée au chapitre 1), seules les identifications qui correspondent à des séquences traduites de transcrits inconnus, c'est à dire introniques ou intergéniques non annotées, et celles qui correspondent à des longs ARNs non codants (Chalmel et al., 2014) sont prises en compte. La sélection des candidats selon des critères génomiques et transcriptomiques se fait selon un certain nombre d'étapes qui sont détaillées dans l'article « Forty-Four Novel Protein-Coding Loci Discovered Using a Proteomics Informed by Transcriptomics (PIT) Approach in Rat Male Germ Cells » (Chapitre 1, page 132).

Pour la recherche de nouvelles isoformes (étude présentée au chapitre 2), seules les identifications qui correspondent à des séquences déduites de transcrits classés comme des nouvelles isoformes potentielles sont prises en compte. Pour faciliter la mise en évidence de nouvelles isoformes potentielles identifiées avec des peptides spécifiques correspondant à de nouveaux exons, des UTRs ou une nouvelle jonction d'épissage, les protéines identifiées avec un peptide non partagé par d'autres protéines sont prises en compte.

b) En protéomique différentielle

La quantification des protéines est obtenue par le calcul automatique, par Proteome Discoverer 1.2 des ratios lourd/léger, super-lourd/léger et super-lourd/lourd par la comparaison des aires relatives des intensités des signaux m/z correspondant aux peptides marqués, observés sur les spectres MS. Seuls les peptides uniques (non partagés par plusieurs protéines), et marqués par une étiquette de masse sont considérés dans l'analyse. Dans un échantillon donné, une protéine est considérée comme significativement quantifiée si elle est quantifiée par au moins deux peptides de séquence distincte distribués dans au moins deux réplicats techniques ou triplex.

Pour les peptides redondants dans un même triplex, la valeur médiane de chacun des ratios (en log-2) est considérée. La quantification de chaque protéine au sein d'un triplex est estimée grâce à la médiane des ratios peptidiques de tous les peptides (non-redondants) assignés à cette protéine. La valeur des ratios au sein des protéines est ensuite normalisée entre les trois triplex par la méthode « quantile-quantile ». Les protéines quantifiées qui

présentent un ratio supérieur ou égal à 2, et dans au moins deux triplex, sont considérées comme différentielles. Un test statistique (test de LIMMA, p-value ajustées par la correction de Benjamini et Hochberg) est ensuite utilisé pour identifier les protéines qui sont différentielles de manière significative. Les protéines différentielles sont ensuite classifiées selon leur profil d'expression relative (Figure 20).

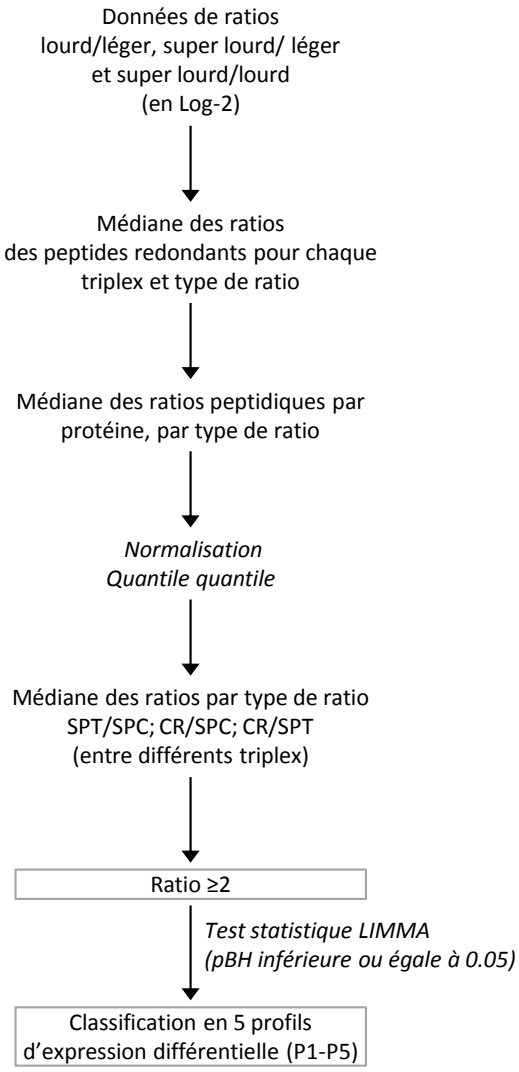

Figure 20. Stratégie d'analyse des protéines quantifiées par ICPL

D. Utilisation de la Gene Ontology

Pour chacun des groupes de protéines identifiées par protéomique, une analyse utilisant les annotations de la Gene Ontology (GO) est réalisée dans les chapitres 2, 3 et 4. Dans le chapitre 3, des enrichissements en termes de la GO dans les groupes P1 à P5 décrits dans l'étude, en utilisant le Gene ID comme référence, ont été calculés grâce à l'outil AMEN (Chalmel et Primig, 2008) en utilisant la probabilité exacte de Fisher. Un terme a été considéré comme sur-représenté par rapport au protéome théorique du rat lorsque le nombre de gènes dans le groupe portant cette annotation est ≥ 3 et lorsque la p-value associée est ≤ 0.01. Les enrichissements en termes GO de la fraction membranaire des spermatides (m_rSPT) comparés à leur extrait total (t_rSPT) sont calculés entre les deux groupes et non par rapport au protéome théorique total, et un terme est considéré comme sur-représenté quand le nombre de gènes dans un groupe portant cette annotation est supérieur à $\geqslant 15$, et quand la p-value associée est $\leqslant 0.01$. Dans l'analyse du protéome du corps résiduel présentée au chapitre 4, les termes GO sont utilisés pour trier des groupes de protéines annotées par un terme en particulier, mais sans réaliser d'enrichissements statistiques sur ces termes.

RÉSULTATS

Chapitre 1

Découverte de nouveaux loci codants par une approche protéomique informée par la transcriptomique (PIT) dans les cellules germinales de rat

I. Contexte et objectifs de l'étude

Dans une optique de découverte de nouveaux gènes codant pour des protéines spécifiquement exprimées à certains stades clés de la spermatogenèse et donc potentiellement impliquées dans ce processus, nous avons utilisé une approche combinant des données de transcriptomique de séquençage d'ARN à haut débit (RNA-seq) à une approche protéomique de type Shotgun. Ce type d'approche nommé Protéomique Informée par la Transcriptomique (PIT) a été décrit par les travaux pionniers de Evans et collaborateurs (Evans et al., 2012). Appliquée aux différents types cellulaires testiculaires, une approche de ce type a été utilisée avec pour objectif d'identifier de nouveaux gènes spécifiquement exprimés dans des cellules germinales mâles. Le profilage à haute résolution de l'expression de nouvelles régions transcrites dans les cellules testiculaires chez le rat a été récemment réalisé dans l'unité en utilisant la technologie Illumina de séquençage nouvelle génération (Chalmel et al., 2014). Des milliers de nouveaux transcrits testiculaires non annotés (TUTs) et de longs transcrits non codants (lncRNAs) ont été mis en évidence dans cette étude. Les auteurs ont montré que ces transcrits s'accumulent aux stades méiotique et post-méiotique de la différenciation des cellules germinales. Bien que les longs ARNs non codants soient définis dans les cellules germinales comme une nouvelle catégorie de lncRNAs et que les TUTs partagent la plupart de leurs caractéristiques génomiques (plus grande longueur d'exons pour les TUTs ou lncRNAs méiotiques par exemple), nous ne savons pas quels sont ceux codant pour des protéines. Dans l'étude de Chalmel et collaborateurs, les TUTs ont été prédits comme ayant un faible voire aucun potentiel codant, et il a été suggéré que la majorité des TUTs consistait en de potentiels nouveaux lncRNAs. Pourtant, nous avons supposé que certains de ces événements de transcription pouvaient correspondre à de nouveaux gènes codant pour des protéines, étant donné leur expression tissu-spécifique et cellule-spécifique, ainsi que leur accumulation aux stades méiotique et post-méiotique.

L'objectif principal de notre étude visait à caractériser et valider au niveau protéique le potentiel codant de ces TUTs ou lncRNAs exprimés dans deux types cellulaires représentatifs des stades méiotique et post-méiotique. La spectrométrie de masse utilisée en mode shotgun permet de répondre au moins partiellement à cette question, à partir du moment où la base de données de séquence permettant d'interroger des données spectrales contient les séquences dérivées de la traduction des transcrits reconstruits après assemblage des données de RNA-seq. Pour réaliser cette étude, nous avons donc intégré les données de RNA-seq des cellules

testiculaires de rat récemment obtenues (Chalmel et al., 2014), aux jeux de données protéomiques obtenus par analyse shotgun des extraits de spermatocytes pachytène (pSPC) et spermatides rondes (rSPT). Une banque de données de séquences dérivées des transcriptomes des cellules testiculaires a été produite par traduction dans les six cadres de lecture des séquences de transcrits reconstruits par RNA-seq. Elle a ensuite été fusionnée avec les bases de données canoniques classiquement utilisées en protéomique : UniProt et Ensembl, pour constituer la base de séquences « Protéome de référence du rat ». Sur la base des identifications protéiques réalisées sur cette banque, et d'une stratégie de sélection des évènements identifiés sur des critères génomiques et transcriptomiques, 44 nouveaux loci ont été identifiés dans les cellules méiotiques et post-méiotiques. Nous validons expérimentalement le fait que des TUTs, ou des lncRNAs, codent véritablement pour des protéines. Pour démontrer la pertinence de cette approche, deux candidats émergeant d'une série de sélections manuelles supplémentaires sont étudiés expérimentalement. Nous validons l'expression de leur transcrit dans le testicule chez le rat, la souris et l'homme, ainsi que dans les cellules germinales méiotiques et post-méiotiques chez le rat. Pour l'un des deux candidats, un anticorps a pu être produit et nous a permis de valider l'expression de la protéine au sein de l'épithélium séminifère.

II. Résultats et discussion

A. Découverte de nouveaux gènes codants exprimés pendant la spermiogénèse

L'approche PIT et notre stratégie de sélection nous ont permis de mettre en évidence 44 nouveaux évènements codants dans les spermatocytes pachytène et les spermatides rondes chez le rat. L'analyse par LC MS/MS des extraits de protéines totales issues des spermatocytes et des spermatides et l'interrogation contre notre banque de données dédiée nous a permis d'identifier au total 19,966 peptides correspondant à 4.999 protéines dans ces deux types cellulaires. Les peptides identifiés pris en compte dans l'analyse sont ceux ayant un score Mascot individuel au dessus du seuil d'identité tel que la p value, probabilité que la correspondance spectre théorique /spectre expérimental soit due au hasard, soit inférieure à 0.05 (p-value < 0.05). Seuls les peptides de rang 1 (meilleure correspondance des spectres parmi plusieurs possibilités de peptides) sont considérés, et taux de faux positifs par rapport à

une banque leurre est fixé à 1% maximum (FDR 1%). En cas de peptides partagés par plusieurs protéines, les protéines ont été groupées automatiquement, et la protéine qui obtient le meilleur score du groupe (calculé en fonction de sa couverture de séquence entre autres), est considérée comme correctement identifiée. Les protéines identifiées avec un seul peptide dont l'identification est acceptée, sont considérées dans notre étude car notre but ici est de découvrir le maximum d'évènements codants nouveaux, et la sélection des candidats ultérieure se veut stringente, et élimine les identifications erronées. Parmi ces identifications, nous nous sommes intéressés aux nouvelles protéines et plus particulièrement à celles qui sont dérivées de TUTs ou de transcrits testiculaires annotés comme des lncRNAs. Les transcrits de cette catégorie s'accumulent pendant les stades méiotique et post méiotique de la spermatogenèse (Chalmel et al., 2014).

a) Sélection de 69 transcrits (44 loci) codants, exprimés dans les spermatocytes ou les spermatides

Ces identifications protéiques obtenues par LC MS/MS avec des peptides « de haute confiance » (Rang 1; FDR1%) qui correspondent à ces transcrits représentent 131 évènements non connus (TUTs) correspondant à (97 loci); et à 29 lncRNAs correspondant à (17 loci). Nous avons choisi de sélectionner, parmi ces TUTs et lncRNAs validés par MS, ceux dont le transcrit est exprimé dans l'un ou l'autre de ces 2 types cellulaires. En effet, les technologies de séquençage peuvent générer un certain nombre d'artéfacts dus au bruit de fond transcriptionnel, à la présence d'ADN génomique dans les échantillons ou bien à des erreurs dans l'alignement des fragments nouvellement séquencés et dans l'assemblage des transcrits (Prensner et al., 2011; Chalmel et al., 2014). Le taux de faux positifs (FDR) des peptides doit également être considéré en protéomique Shotgun (Nesvizhskii, 2010; Serang et al., 2013). Cette sélection des transcrits déjà connus pour être exprimés dans ces types cellulaires permet de limiter la découverte d'évènements artéfactuels par MS. Pour cela, nous avons aussi sélectionné en priorité les protéines correspondant aux longues isoformes, avec une longueur supérieure à 200 nucléotides. Ce critère est généralement posé en transcriptomique pour éliminer les possibilités d'artéfact. Ceci nous a amené à sélectionner 129 TUTs (96 loci) et 29 lncRNA (17loci). Parmi ceux-ci, 69 transcrits (44 loci), comprenant 48 TUTs (30 loci) et 21 lncRNAs (14 loci), sont exprimés au niveau du transcrit dans les pSPC et/ou dans les rSPT. Il est à noter que 15 TUTs (12 loci) et 16 lncRNAs (12 loci) ont

des profils d'expression préférentiellement méiotique ou post-méiotique au regard des données quantitatives de RNA-seq obtenues par Chalmel et collaborateurs.

Comparé aux TUTs et lncRNA exprimés dans les pSPC et /ou dans les rSPT, mais qui n'ont pas été détectés par MS/MS, ces 69 transcrits « MS-identified » présentent des caractéristiques génomiques différentes, comme le montre l'analyse statistique réalisée dans cet article. Ils ont aussi été comparés aux ARNms dont la protéine a été identifiée par MS/MS, et qui ont une annotation connue. Les 69 TUTs et lncRNAs « MS-identified » présentent une meilleure conservation des exons, et des ORFs plus longs que ceux qui ne sont pas identifiés par MS, mais qui sont exprimés dans ces types cellulaires. Ils sont aussi à une distance des gènes connus similaire de celles des mRNAs connus identifiés par MS. Les TUTs et lncRNAs identifiés par MS possèdent deux fois plus de variants d'épissage que ceux non identifiés par MS. Ces 69 TUTs et lncRNAs ont des caractéristiques génomiques qui se situent entre celles des ARNms connus, et celles des lncRNAs. Il est important de noter que la distance des 69 « MS-identified » aux plus proches gènes codant pour des protéines n'est pas significativement différente de la distance des ARNms identifiés par MS aux plus proches gènes codants Ces observations vont dans le sens du caractère codant de ces 69 transcrits, et qui confirme que ce ne sont pas de nouveaux longs ARNs non codants comme suggéré précédemment (Chalmel et al., 2014). Notons que les peptides de haute confiance et de rang 1 ont été alignés sur le génome rn4 du rat, et que ces résultats sont visibles graphiquement sur le site de RGV (http://rgv.genouest.org/). Il faut toutefois garder à l'esprit que parmi ces peptides, même s'ils sont alignés, la quantité de faux positifs est estimée à 1%.

b) Vamp9 et XLOC_001949 : nouveaux gènes codant pour des protéines exprimées pendant la spermiogénèse.

Dans cette étude, un de nos objectifs a été de démontrer la pertinence d'une approche PIT pour mettre en évidence de nouvelles protéines dont le rôle dans la biologie des cellules germinales mériterait une analyse approfondie. Nous avons donc cherché parmi ces 69 nouveaux transcrits ceux qui étaient les plus enclins à avoir une fonction dans la spermatogenèse chez les mammifères. En considérant le fait qu'une fonction conservée chez les mammifères nécessite une séquence génomique conservée, nous avons étudié de plus près ceux qui ont le score de conservation le plus élevé (score de conservation PhastCons défini comme la conservation base par base calculée entre les génomes de 9 vertébrés fournis dans le navigateur UCSC (Meyer et al., 2013)). Suite à ce choix des candidats aux séquences les

plus conservées, et de préférence multiexoniques, nous avons poussé notre investigation sur deux candidats dont nous avons évalué l'expression au niveau du transcrit par RT-PCR dans le testicule comparé à plusieurs autres tissus (moelle osseuse, cerveau, rein, foie, poumon et muscle) ou types cellulaires isolés (cellules de Sertoli, spermatogonies, spermatocytes pachytène et spermatides rondes) chez le rat. Nous avons également comparé l'expression de ces deux candidats dans le testicule par rapport à d'autres tissus chez la souris (rein, foie et poumons) et chez l'homme (épididyme, prostate et vésicule séminale). Ces analyses ont montré une expression testicule-spécifique de ces deux transcrits candidats chez le rat, et une expression préférentielle dans le testicule chez la souris et l'homme. Les profils d'expression de ces transcrits dans des cellules germinales méiotiques ou post-méiotiques évalués par RNA-seq ont été confirmés par qPCR, et par hybridation *in situ* sur des testicules de rats adultes. Notons que pour le candidat XLOC_001949 exprimé dans les spermatocytes pachytène et plus fortement dans les spermatides, l'expression n'est pas visible en ISH dans les spermatocytes pachytène, alors que celle-ci est détectable par qPCR et RNA-seq. Nous pouvons mettre en cause une plus grande sensibilité de ces deux dernières techniques comparées à l'ISH, et dans une moindre mesure, l'existence de contaminations possibles lors de la purification des cellules germinales.

Les alignements multiples des protéines orthologues prédites pour les deux candidats, retrouvées chez 26 espèces de mammifères pour XLOC_001949, et 13 espèces de mammifères pour XLOC_013843 avec l'outil Blast (NCBI) contre UniProt, RefSeq et les bases d'EST ont été instructifs. Ils révèlent un domaine énolase conservé pour la protéine XLOC_001949 et un domaine SNARE-like caractéristique des protéines de la famille des longines pour la protéine XLOC_013843. Cette dernière présente une forte homologie avec la protéine VAMP7, son paralogue potentiel, qui possède également un domaine « longin ». Nous avons donc appelé cette nouvelle protéine VAMP9 (vesicle-associated membrane protein 9), puisque cette famille de protéines s'étend actuellement de VAMP1 à VAMP8. Nous avons nommé T-ENOL (pour Testicular Enolase) la protéine XLOC_001949 qui possède un domaine énolase conservé prédit.

Ces résultats étaient particulièrement intéressants pour le candidat XLOC_001949 (nouvelle énolase), à savoir une forte conservation de la séquence chez les mammifères avec une prédiction de domaine fonctionnel ainsi que la présence d'une majorité d'EST retrouvés exprimés dans le testicule chez 7 espèces de mammifères. Au niveau expérimental, ce transcrit présente une expression différentielle et restreinte dans les pSPC et les rSPT où il est

le plus exprimé (jeunes spermatides à spermatides au stade 18), comme en témoignent les résultats de qPCR et RT-PCR ainsi que d'ISH présentés dans cet article. J'ai donc cloné l'ORF XLOC_001949 et produit la protéine recombinante T-ENOL contre laquelle un anticorps polyclonal a été produit. Cet anticorps utilisé en immunohistochimie a permis de confirmer l'expression méiotique et post-méiotique de la protéine T-ENOL chez le rat, depuis le stade jeune spermatides, et s'intensifiant jusqu'aux stades spermatides allongées et même dans les corps résiduels. Rappelons que cette protéine avait été identifiée par MS/MS avec un peptide dans les pSPC et les rSPT, et avec deux peptides dans les corps résiduels (jeu de données non présenté dans cet article, voir Tableau 4).

Grâce à une approche PIT, nous avons donc découvert deux protéines susceptibles de jouer un rôle dans la spermatogenèse au vu de leur expression spécifiquement testiculaire et germinale. La protéine VAMP9 est particulièrement intéressante pour l'étude de la spermatogenèse dans la mesure ou les données de la littérature indiquent que cette protéine est associée à d'autres protéines, formant des complexes associés à plusieurs processus cruciaux pour la spermatogenèse ou la reproduction. VAMP9 contient en effet un domaine SNARE-like, (soluble N-ethylmaleimide-sensitive factor attachment protein receptor), et doit appartenir à la famille des Longines comme son proche paralogue VAMP7 (Filippini et al., 2001). Les protéines VAMP sont impliquées dans la régulation de trafic membranaire et les complexes SNARE sont connus pour être impliqués dans les voies d'endocytose et de sécrétion (Chaineau et al., 2009). VAMP3 (Cellubrevin) est sans doute impliquée dans la spermatogenèse, car elle est associée au transport de TEX101 à la surface des cellules germinales (Tsukamoto et al., 2006). TEX101 est considérée comme un marqueur de fertilité chez le mâle (Drabovich et al., 2011), nécessaire à la bonne migration du spermatozoïde dans l'oviducte chez la souris (Li et al., 2013). D'autres protéines de la famille des SNARE sont impliquées dans des fonctions importantes pour la formation des spermatozoïdes, notamment pendant la biogenèse de l'acrosome (Ramalho-Santos et al., 2001) et la capacitation, comme la syntaxine 2 (Hutt et al., 2005) ainsi que VAM6P et SNAP (Brahmaraju et al., 2004). La syntaxine 17 pourrait être impliquée dans la stéroïdogenèse (Katafuchi et al., 2000). Il est aussi intéressant de constater que VAMP7 est impliquée dans des processus de différenciation tels que la croissance des neurites (Sato et al., 2011a) et la ciliogenèse (Szalinski et al., 2014), et définirait également une nouvelle voie de circulation à la surface cellulaire dans les cellules neuronales et non neuronales (Flowerdew et Burgoyne, 2009). La ciliogenèse est un processus particulièrement intéressant dans l'étude de la spermatogenèse,

car les flagelles et les cils ont de nombreuses caractéristiques en commun (Gottardo et al., 2013). La protéine T-ENOL que nous avons découverte est fortement conservée et très exprimée dans les spermatides en allongement et allongées ainsi que dans les corps résiduels. En accord avec de précédentes observations chez d'autres espèces de mammifères, cette protéine pourrait être une nouvelle énolase spécifique des cellules germinales mâles matures.

L'énolase (2-phospho-D-glycérate hydrolase; EC 4.2.1.11) catalyse la conversion du 2-phosphoglycérate en phosphoénolpyruvate, le second des deux produits intermédiaires de haute énergie qui génèrent de l'ATP dans l'avant-dernière étape de la glycolyse. Les isoenzymes de l'énolase comprennent les énolases 1 (α), énolase 2 (γ) et énolase 3 (β) codées par 3 gènes différents (Tracy et Hedges, 2000). Cependant, une énolase spécifique des spermatozoïdes a été détectée chez l'homme, le bélier, et la souris (Edwards et Grootegoed, 1983). En outre, l'énolase a été immunolocalisée dans le flagelle chez le rat (Gitlits et al., 2000) et dans la pièce principale du flagelle chez l'homme (Force et al., 2004). L'activité enzymatique de l'énolase a été détectée pendant l'élongation des spermatides chez la souris (Edwards et Grootegoed, 1983) et dans les corps résiduels chez le rat (Gitlits et al., 2000). Chez l'homme, l'activité de l'énolase est associée avec les spermatozoïdes matures portant du matériel cytoplasmique (Force et al., 2004). Les énolases spécifiques qui sont exprimées dans les spermatozoïdes chez la souris ont une expression relativement tardive au cours de la différenciation des cellules germinales mâles (Edwards et Grootegoed, 1983). Des énolases spécifiques du spermatozoïde sont aussi identifiées chez l'homme : Eno-S associée à des spermatozoïdes normaux (Force et al., 2002; 2004). Elles ont un rôle clé dans la spermatogenèse, car elles sont impliquées dans la glycolyse, utilisée pour générer de l'ATP indispensable à la motilité des spermatozoïdes. Une nouvelle énolase qui ne présente pas d'homologie avec T-ENOL a été découverte tout récemment chez la souris, codée par le gène 64306537H0Rik, présente comme « non caractérisée » dans les bases de données, et re-nommée ENO4 (Nakamura et al., 2013). L'invalidation du gène Eno4 conduit à l'infertilité en raison d'une réduction de la motilité des spermatozoïdes qui présentent un assemblage anormal de la gaine fibreuse (Nakamura et al., 2013). Les travaux de ces auteurs sont similaires aux nôtres dans le sens de la découverte d'une nouvelle protéine par une approche de protéomique (ces auteurs ont utilisé la 2D-MS). Ils ont toutefois le mérite de fournir la preuve que cette nouvelle ENO4 est indispensable à une spermatogenèse normale, par des expériences d'invalidation de gène. La nouvelle énolase que nous avons découverte, T-ENOL, a une expression qui augmente graduellement de la méiose à la fin de la

spermiogénèse, et jusque dans les corps résiduels. Ceci nous amène à penser que T-ENOL pourrait avoir chez le rat un rôle important dans la motilité des spermatozoïdes. Ensemble, ces éléments montrent qu'une approche de type PIT couplant RNA-seq et protéomique Shotgun permet la découverte de nouveaux éléments codants intéressants pour l'étude de la spermatogenèse, et qu'il peut en être de même dans d'autres tissus et processus biologiques.

B. L'approche PIT et la réannotation du génome du rat

L'intérêt de cette approche n'est pas simplement la découverte de nouveaux évènements codants impliqués dans un processus biologique particulier. Elle permet aussi de pointer du doigt des erreurs d'annotation concernant des gènes connus ou prédits voire des problèmes d'assemblage du génome. En particulier, notre étude a permis de mettre en évidence une erreur d'assemblage sur le génome rn4 du rat. En effet, la séquence correspondant au transcrit TCONS_00003700 correspondant au locus XLOC_003503 qui a été dentifiée par MS dans les pSPC ainsi que dans les corps résiduels, a passé nos filtres de sélection selon la stratégie de filtration utilisée pour les TUTs et les lncRNA présentée au début de ce chapitre.

TCONS_00003700 (chr1:85221796..85222201)

3'5' cadre 2

```
gctggctacaggcggctgttaagaagcgtaacggacgctggtctccgacagcatgatggct
 L  A  T  G  G  C  -  E  A  -  R  T  L  V  S  D  S  M  M  A
gccttagcggccggaggttacgcgcggagtgacacgatagaaaagctgtcctctgtcatg
A  L  A  A  G  G  Y  A  R  S  D  T  I  E  K  L  S  S  V  M
gcgggagttccggcgcggagaaaccagtcctccccgcctcctgccccaccgctctgcctc
A  G  V  P  A  R  R  N  Q  S  S  P  P  P  A  P  P  L  C  L
cggcggcggacgcgactcgcggcggctcccgaggacactgtgcagaaccgggtgagaggc
R  R  R  T  R  L  A  A  A  P  E  D  T  V  Q  N  R  V  R  G
tggtcgccctgtttatccaactccaggcttctctgtgcctccttggttcctcccatcact
W  S  P  C  L  S  N  S  R  L  L  C  A  S  L  V  P  P  I  T
tcttgttccctcaggcggcttcccaggctcttctggtccggaagccacgcccccttggca
S  C  S  L  R  R  L  P  R  L  F  W  S  G  S  H  A  P  L  A
aaggctccctctacgttcgtgctggcttcagagtttggtgccgcc
K  A  P  S  T  F  V  L  A  S  E  F  G  A  A
```

Figure 21. Séquence traduite du transcrit TCONS_00003700

Séquence protéique traduite du transcrit TCONS_00003700. Les résidus sur lesquels s'alignent les peptides identifiés par MS/MS dans les pSPC (Max.E. Value=4.10^{-4} et Max.E. Value 9.10^{-4}) et dans les corps résiduels (données non présentées dans l'article) (Max.E. Value=$2,2.10^{-2}$), sont indiqués en vert. La méthionine initiatrice est indiquée en rouge.

Ce transcrit est exprimé dans les pSPC, et son « gène », monoexonique, situé sur le chromosome 1 du génome du rat (chr1:85,221,796..85,222,202), présente une région très conservée parmi les vertébrés (Figure 22A), et est isolé sur le génome du rat. Nous avons détecté sa protéine en spectrométrie de masse dans les spermatocytes et les corps résiduels, et retenu le transcrit TCONS_00003700 (XLOC_003503), car ce dernier est annoté comme un TUT. Il est de surcroit très exprimé à un des stades évalués (pSPC) (Figure 22A,B). Nous avons donc décidé d'évaluer l'expression de ce transcrit sélectionné dans les cellules germinales isolées en comparaison avec d'autres types cellulaires et d'autres tissus. Le transcrit est fortement exprimé dans les spermatocytes pachytène, et, dans une moindre mesure, dans les autres types cellulaires testiculaires (cellules de Sertoli, spermatogonies et

spermatides rondes) ainsi que le montrent les données de RNA-seq (Figure 22A,B) et le confirme l'expérience d'ISH (Figure 22C). Les spermatides rondes et les spermatides en allongement expriment plus faiblement le transcrit TCONS_00003700 à ces stades comme le montre l'ISH (Figure 22C). Les résultats de RT-PCR visant à détecter l'expression du transcrit dans le testicule et les différentes cellules germinales isolées, en comparaison avec d'autres tissus, montrent que ce transcrit est préférentiellement exprimé dans le testicule et les cellules germinales (Figure 22D).

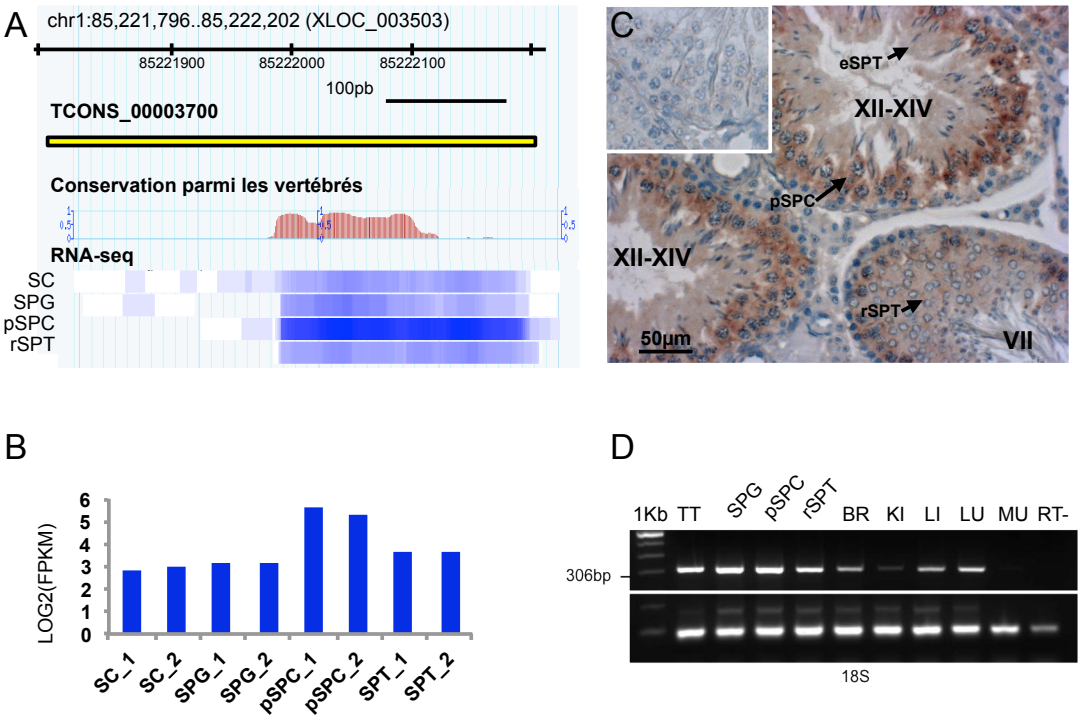

Figure 22. Profil d'expression du transcrit XLOC_003503 (TCONS_00003700) dans les cellules testiculaires isolées de rat, et dans le testicule comparé à d'autres tissus

A, structure du gène (boîte jaune) le long du chromosome 1 du génome du rat, conservation de la séquence entre neuf génomes de vertébrés fournis dans le navigateur UCSC (score phastCons en rouge), et, abondance du transcrit déterminée par RNA-seq dans quatre populations cellulaires testiculaires différentes (SC=cellules de Sertoli, SPG=spermatogonies, pSPC=spermatocytes pachytène et rSPT=spermatides rondes) représentée par un code couleur bleu. **B,** expression quantitative du transcrit TCONS_00003700 évaluée par le nombre de reads obtenus par RNA-seq (Chalmel et al., 2014). **C,** Localisation *in situ* du transcrit TCONS_00003700, les chiffres romains indiquent les stades de l'épithélium séminifère selon Leblond et Clermont (Leblond et Clermont, 1952). **D,** détection du transcrit TCONS_00003700 par RT-PCR dans le testicule total (TT), et les cellules germinales isolées: SPG, pSPC et rSPT comparé à d'autres tissus: cerveau (BR), rein (KI), foie (LI), poumon (LU), et muscle (MU). (RT= contrôle négatif). L'hybridation *in situ* a été réalisée sur des sections de paraffine de testicule de rat adulte avec une sonde anti-sens (ARN polymérase T7 avec le vecteur linéarisé par BamH1) ou la sonde sens (ARN polymérase SP6 avec le vecteur linéarisé par Xho1) pour le contrôle négatif (dans l'encart, en C), marquées à la dioxygénine et spécifiques du transcrit TCONS_00003700. La barre d'échelle représente 50µm. Amorces utilisées pour la détection du transcrit TCONS_00003700 en RT-PCR : Forward, 5'_GCGGCTGTTAAGAAGCGTAA ; Reverse, 5'_GCCTGAGGGAACAAGAAGTG ; amorces utilisées pour la synthèse des sondes destinées à l'ISH : Forward, 5'-GCGGCTGTTAAGAAGCGTAA ; Reverse :5'_AGAGGGAGCCTTTGCCAA.

La séquence protéique traduite de TCONS_00003700 (XLOC_003503) détectée par MS (Figure 21), s'aligne (en utilisant l'algorithme Blastp, NCBI) en avec le N-ter de la protéine prédite: F1M5K2 (WD repeat-containing protein 62) d'une longueur de 1551 a.a. En revanche, la séquence XLOC_003503, locus du transcrit TCONS-00003700 situé sur le chromosome 1 à la position (chr1:85,221,796..85,222,201), ne correspond pas à la

localisation du gène Wdr62 qui est localisé en amont sur le chromosome 1 (chr1:85,448,728-85,482,510), sur le génome rn4 du rat (Baylor 3.4/rn4). Elle se trouve d'ailleurs isolée de gènes connus, comme le montre la copie d'écran du navigateur RGV (Figure 23A). En effet, mise à part la prédiction de gène «rergla» à plus de 2kb, il n'y a aucun gène connu dans cette région à moins de 10kb, c'est pourquoi cette séquence correspondait à nos critères de sélection des candidats. L'utilisation de l'outil BLAT (UCSC) avec la séquence de la protéine WDR62 (F1M5K2) sur le génome rn4 laisse apparaitre que le premier exon codant pour les 60 premiers a.a. de cette protéine n'est pas aligné à la position du gène Wdr62 (chr1:85448849-85482423). Par contre, il est localisé en sens inverse à la position chr1:85221967-85222149 correspondant à XLOC_003503 (Figure 23B). Ce premier exon code en sens inverse du gène Wrd62 pour le N-ter de WDR62 pour lequel nous avons détecté deux peptides dans les spermatocytes. En effet, la protéine TCONS_00003700 est codée en 3'-5'cadre 2 (Figure 21). Or, chez l'homme, en utilisant l'outil BLAT avec la séquence de la protéine homologue WDR62 (O43379) sur le génome GRch38/hg38 (décembre 2013), il apparaît que le premier exon du gène codant pour cette protéine se trouve dans le bons sens, localisé dans la même région que les 31 autres exons du gène Wdr62 (chr19:36054972-36105025). Donc deux hypothèses se présentent: 1) une modification réelle au niveau du gène Wdr62 conduisant à une perte du N-ter-de la protéine WDR62 s'est produite chez le rat, ou bien 2) une erreur d'assemblage du génome rn4 du rat s'est produite, où l'exon 1 (XLOC_003503) aurait été inversé et éloigné des autres exons du gène Wdr62 de plus de 226,5kb (Figure 23B).

Pour tester la seconde hypothèse, nous avons généré des amorces sens sur l'exon 1 (XLOC_003503), en tenant compte de son inversion (c'est à dire sens inverse de Wdr62, tel qu'il est assemblé sur le génome rn4), et des amorces antisens complémentaires d'une partie de la séquence de l'exon 3 de Wdr62. Une absence d'amplicons serait en faveur de la première hypothèse et une présence d'amplicons à la taille attendue serait en faveur de l'hypothèse 2, c'est à dire attesterait que ces deux régions codent dans le même sens et sont liées sur l'ADN génomique. Les résultats de PCR montrent un amplicon à la taille attendue entre l'exon 1 et 3 du gène Wdr62 (255pb) (Figure23D), ce qui indique bien que cette séquence XLOC_003503 appartient au gène Wdr62 et que le transcrit TCONS_00003700 est le produit de la transcription du premier exon de ce gène et non un transcrit isolé. Le séquençage de cet amplicon confirme qu'il s'agit bien d'une portion de Wdr62. Une erreur d'assemblage du génome rn4 s'est donc produite où la séquence XLOC_003503, premier

exon du gène Wdr62 est éloignée de 226kb du reste du gène et inversée par rapport à lui. En revanche, sur le nouvel assemblage du génome du rat RGCS5.0/rn5 (rn5), le premier exon codant pour WDR62 est assemblé dans le même sens que les 31 autres exons codant pour cette protéine dans la région chr1:89852321-89890241 en brin inverse, ou dans la région chr1:92132020-92171687 en brin sens (Figure 23C).

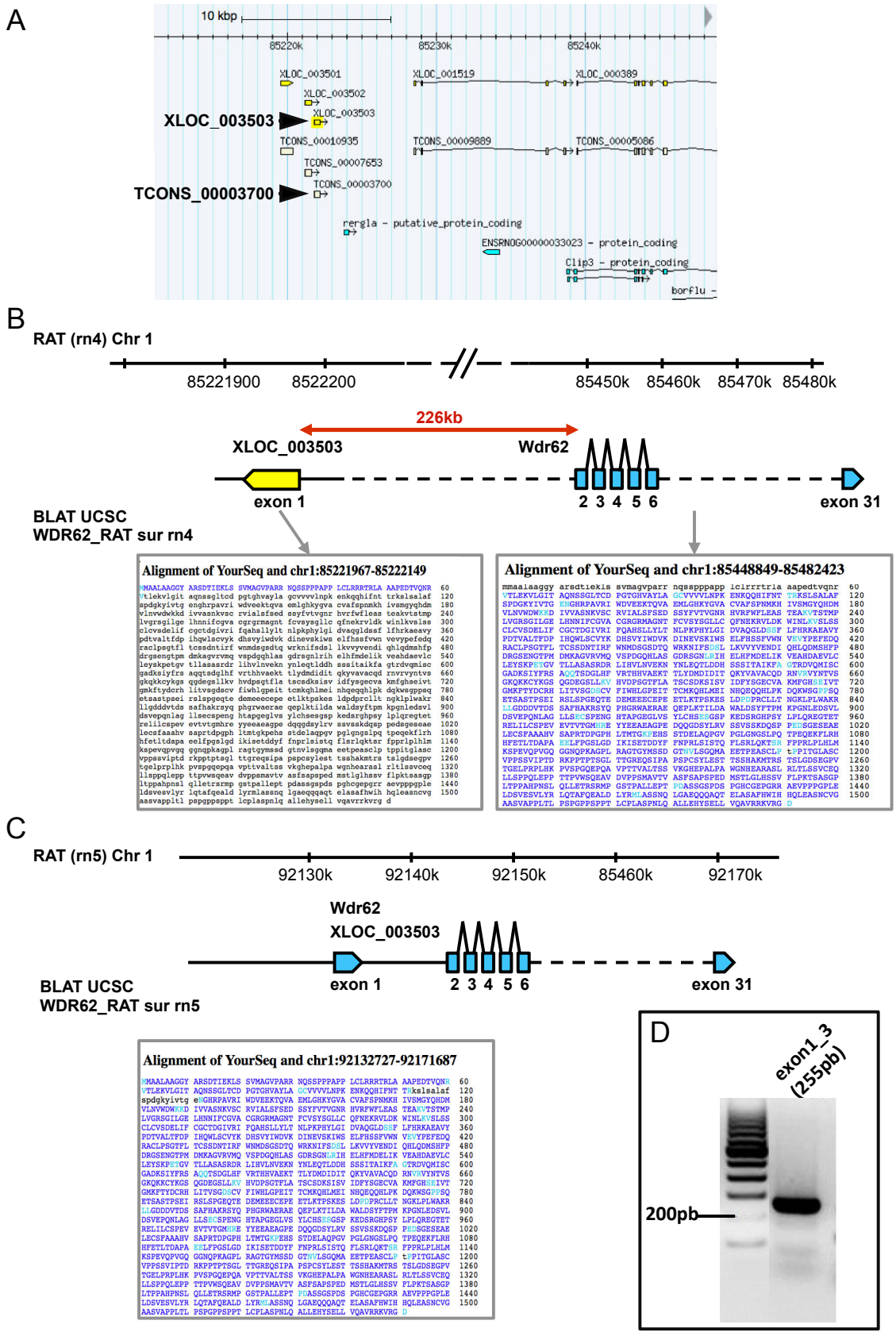
A
10 kbp
85220k
85230k
85240k
XLOC_003501
XLOC_001519
XLOC_000389
XLOC_003502
XLOC_003503
XLOC_003503
TCONS_00010935
TCONS_00009889
TCONS_00005086
TCONS_00007653
TCONS_00003700
TCONS_00003700
rergla - putative_protein_coding
ENSRNOG00000033023 - protein_coding
Clip3 - protein_coding
borflu -
B
RAT (rn4) Chr 1
85221900
8522200
85450k
85460k
85470k
85480k
226kb
XLOC_003503
Wdr62
exon 1
2 3 4 5 6
exon 31
BLAT UCSC
WDR62_RAT sur rn4
Alignment of YourSeq and chr1:85221967-85222149
Alignment of YourSeq and chr1:85448849-85482423
C
RAT (rn5) Chr 1
92130k
92140k
92150k
85460k
92170k
Wdr62
XLOC_003503
exon 1
2 3 4 5 6
exon 31
BLAT UCSC
WDR62_RAT sur rn5
Alignment of YourSeq and chr1:92132727-92171687
D
exon1_3
(255pb)
200pb

Figure 23. Erreur sur le génome rn4 du rat au niveau du 1er exon du gène Wdr62

A, vue de la position du transcrit TCONS_00003700 identifié par MS, correspondant au locus XLOC_003503 sur le génome rn4 du rat, sur le navigateur RGV (http://rgv.genouest.org/). Cette vue élargie montre l'isolement de TCONS_00003700 du reste du gène Wdr62 et des gènes connus (boîtes bleues). **B, C**, vue schématique de la localisation des séquences codant pour la protéine WDR62 chez le rat. B, sur le génome rn4 du rat, et en C, sur l'assemblage rn5 du génome du rat. Les copies d'écran des résultats du Blat UCSC pour les séquences des protéines WDR62 chez le rat, sur l'un ou l'autre des assemblages du génome rn4 ou rn5, indiquent les portions de séquence codantes alignées sur le génome. Les résidus colorés en bleu sont ceux correspondant à des bases alignées sur le génome, et les résidus bleu clair correspondent aux frontières entre les exons. **D,** Une RT-PCR réalisée avec 1µL d'ADNc de rat avec des amorces sur l'exon1 (Sens : aagctgtcctctgtcatggc), et l'exon 3 (antisens : gttcaagaccaccaccacac) du gène Wdr62 montre qu'un amplicon de 255 pb est obtenu, confirmant le lien entre XLOC_003503 et le reste du gène Wdr62 ainsi que l'inversion de XLOC_003503 sur l'assemblage rn4 du génome.

L'approche PIT nous a permis de détecter cette erreur, et comme notre stratégie de sélection de candidats potentiels est très stringente, celle-ci nous est apparue comme évidente. Cependant, d'autres erreurs peuvent être cachées derrière les identifications de transcrits moins exprimés dans les cellules méiotiques et post-méiotiques, et que nous n'avons pas analysés manuellement.

L'approche PIT permet d'améliorer l'annotation du génome du rat comme le prouve cet exemple. De plus, un autre exemple s'est présenté lors de notre étude, avec la nouvelle énolase T-ENOL. En effet, par inférence, nous pouvons fournir une nouvelle annotation pour son homologue humain sur le locus LOC440356 qui était annoté de façon ambiguë comme un évènement non codant « non-coding CDIPT antisense RNA 1 » et comme produit d'une prédiction de gène douteuse dans la base Uniprot (accession Q0VD67). L'approche PIT permet donc de mettre le doigt sur des problèmes ou des incertitudes dans l'annotation du génome du rat en plus de permettre la découverte de nouveaux évènements codants. En ce sens elle est donc une approche puissante, mais nous verrons que telle que nous l'avons menée elle mérite des améliorations.

C. Une limite de l'approche PIT

Des séquences traduites écourtées suite à des erreurs de séquençage ou d'assemblage des transcrits peuvent être identifiées avec de meilleurs scores que les séquences protéiques réelles plus longues. Un exemple a été facilement identifiable dans l'étude présentée dans ce chapitre sur XLOC_001949, locus sur lequel plusieurs transcrits ont été assemblés, dont TCONS_00010279 et TCONS_00012662. La séquence traduite de TCONS_00012662 de 90 a.a. a été identifiée dans les pSPC et les rSPT avec une meilleure couverture de

séquence (15,56%) que ne l'aurait été TCONS_00010279 longue de 116 a.a. car sa couverture de séquence aurait en effet été de 12%. Or, le peptide ayant permis d'identifier TCONS_00010279 dans les corps résiduels n'a pas été détecté par MS/MS dans les autres types cellulaires. C'est donc la séquence TCONS_00012662 qui a été identifiée avec la meilleure couverture de séquence dans les pSPC et les rSPT (Tableau 4,5).

Cell.	Accession	# a.a.	MW	Couv.	ΣPept.	SequencePeptide	ScoreMax	E.ValueMax
pSPC	TCONS_00012662_6_8	90	10,28	15,56	1	**ISTELTDEALFTAR**	65,79	$8,41.10^{-5}$
rSPT	TCONS_00012662_6_8	90	10,28	15,56	1	**ISTELTDEALFTAR**	101,35	$2,35.10^{-8}$
CR	TCONS_00010279_6_6	116	12,94	25,86	2	**ISTELTDEALFTAR**	105,81	$8,42.10^{-9}$
CR	TCONS_00010279_6_6	116	12,94	25,86	2	**DTWPIQAAASLGGGQK**	59,12	$3,82.10^{-4}$

Tableau 4. Identification des séquences TCONS_00012662 et TCONS_00010279 (XLOC_001949)

Identification des séquences TCONS_00012662 et TCONS_00010279 (XLOC_001949) dans les spermatocytes pachytène (pSPC), les spermatides rondes (rSPT), et les corps résiduels (CR). Les peptides identifiés par MS/MS correspondant à ces différentes séquences sont indiqués pour les pSPC, les rSPT et les CR. Les corps résiduels ont été obtenus par la même technique d'élutriation centrifuge que les cellules germinales isolées (Pineau et al., 1993), la préparation et l'analyse en LC-MS/MS des CR a été réalisée de la même façon que pour les cellules germinales (Chocu et al., 2014). L'accession des séquences des protéines déduites des transcrits assemblés est affichée telle qu'elle figure dans la base de données « Non redundant Rat proteome ». La longueur de la séquence protéique en nombre de résidus d'acides aminés (#a.a.), le poids moléculaire (MW) en kDa, la couverture de séquence en pourcentage (Couv .), le nombre de peptides non redondants identifiés correspondant à chaque séquence, leur séquence en a.a., leur score maximal (ScoreMax) et leur E.value maximale (EValueMax) sont indiqués. Un grouping de protéines a été appliqué de la même façon pour chaque analyse, prenant en compte les peptides de rang 1, avec un FDR de 1% maximum. L'accession affichée ici est celle de la « master protein ». Dans les pSPC et rSPT, c'est la séquence la plus courte qui est identifiée, en l'absence du deuxième peptide.

La protéine TCONS_00010279 est cependant correcte car cette protéine est conservée et possède un domaine énolase prédit, tandis qu'un Blast sur Uniprot de la partie N-ter de TCONS_00012662 qui diffère entre ces deux séquences de protéines ne donne pas de résultat. De plus, la traduction de TCONS_00012662 dans un autre cadre de lecture permet d'obtenir la partie N-ter de TCONS_00010279 (MASTSARSGDKKDTWPIQAAASLGGGQK). On peut en déduire que la séquence TCONS_00012662 résulte d'un changement de cadre de lecture délétère.

Locus	Séquence prédite
XLOC_001949	>TCONS_00012662_3'-5' Frame3 *MGDGIHICQEWGQKGHLANSSCCLLRWRAASLSRSEEFLTRISTELTDEALFTARSHMNPMPDKEKQTKDQGTQISRHVFFTKTRGTDTR*
	>TCONS_00010279_3'-5': Frame 3 *AKQRVFFKDRTETVGCLSFSRGLNKWGMASTSARSGDKKDTWPIQAAASLGGGQKASLSRSEEFLTRISTELTDEALFTARSHMNPMPDKEKQTKDQGTQISRHVFFTKTRGTDTR*

Tableau 5. Séquences des protéines traduites de transcrits issus du même locus XLOC_001949

Séquences de transcrits issus de XLOC_001949 identifiées dans les sSPC, les rSPT et les CRs. Les peptides identifiés par MS/MS sont indiqués en vert sur la séquence, et la méthionine initiatrice en rouge. Les séquences traduites constituant la base de données « PIT » comprennent tous les résidus entre deux codons stop (*), c'est pourquoi, la longueur considérée pour la première séquence est de 90 a.a., et le seconde de 116 a.a. (cf Tableau 5).

III. Conclusion

Nous l'avons vu, le plus grand avantage du séquençage *de novo* en protéomique est qu'il permet l'identification de spectres pour lesquels le peptide exact n'est pas présent dans les bases de données canoniques. De ce fait, l'approche PIT a permis de découvrir de nouveaux évènements codants sur la base du transcriptome reconstruit dans les cellules testiculaires isolées et dont le rôle potentiel pendant la spermatogenèse reste à confirmer. Les TUTs mis en évidence, et qui s'accumulent aux derniers stades de la spermiogenèse, sont supposés être de potentiels nouveaux lncRNAs (Chalmel et al., 2014). Parmi eux nous prouvons qu'un certain nombre (69 transcrits) sont codants pour des protéines détectées par MS avec un ou plusieurs peptides de haute confiance. Cette approche nous a permis par ailleurs de mettre en évidence de nouvelles isoformes potentielles, qui représentent 75 à 78% des identifications dans les différents types cellulaires étudiés, et qui feront l'objet d'études ultérieures au laboratoire en tenant compte des biais mentionnés plus haut. En outre, cette approche peut permettre de ré annoter le génome du rat comme décrit pour les deux exemples: celui du transcrit XLOC_001949 dont l'homologue humain est annoté comme ARN non codant et qui finalement code pour une nouvelle énolase, et celui du gène WRD62 sur le génome du rat. L'alignement des peptides identifiés par MS/MS sur le génome du rat et visibles *via* le

navigateur RGV est un bon moyen de vérifier visuellement qu'un transcrit est codant ou non, ou qu'une région codante est nouvelle (s'alignant sur un 3' ou un 5'UTR de transcrit connu, ou bien sur une région intergénique, ou encore sur une région intronique de gène connu). Il faut cependant garder en mémoire que l'analyse protéomique peut générer un certain taux de faux positifs. Les faiblesses de cette approche telles que la détection de séquences erronées ou raccourcies due à des erreurs de séquençage, d'alignement, ou de changement de cadre de lecture lors de la traduction auront certainement tendance à disparaître avec l'expérience sur ce type d'approches et l'adaptation des protocoles de séquençage. Par exemple, adopter un protocole de RNA-seq brin spécifique, ce qui réduirait la taille des banques de données de séquences personnalisées, pourrait contribuer à limiter le nombre de séquences erronées dans ces banques. On peut donc dire que l'approche PIT est puissante dans la mesure où elle permet de découvrir de nouveaux gènes et même de ré annoter le génome, mais que son potentiel est encore sous réserve d'améliorations de celle-ci.

ARTICLE 1

Forty-four novel protein-coding loci discovered using a PIT approach in rat male germ cells

BIOLOGY OF REPRODUCTION (2014) **91**(5):123, 1–14
Published online before print 10 September 2014.
DOI 10.1095/biolreprod.114.122416

Forty-Four Novel Protein-Coding Loci Discovered Using a Proteomics Informed by Transcriptomics (PIT) Approach in Rat Male Germ Cells[1]

Sophie Chocu,[5,6] Bertrand Evrard,[6] Régis Lavigne,[5,6] Antoine D. Rolland,[6] Florence Aubry,[6] Bernard Jégou,[6] Frédéric Chalmel,[3,4,6] and Charles Pineau[2,4,5,6]

[5]*Proteomics Core Facility Biogenouest, Inserm U1085, IRSET, Campus de Beaulieu, Rennes, France*
[6]*Inserm U1085, IRSET, Université de Rennes 1, Rennes, France*

ABSTRACT

Spermatogenesis is a complex process, dependent upon the successive activation and/or repression of thousands of gene products, and ends with the production of haploid male gametes. RNA sequencing of male germ cells in the rat identified thousands of novel testicular unannotated transcripts (TUTs). Although such RNAs are usually annotated as long noncoding RNAs (lncRNAs), it is possible that some of these TUTs code for protein. To test this possibility, we used a "proteomics informed by transcriptomics" (PIT) strategy combining RNA sequencing data with shotgun proteomics analyses of spermatocytes and spermatids in the rat. Among 3559 TUTs and 506 lncRNAs found in meiotic and postmeiotic germ cells, 44 encoded at least one peptide. We showed that these novel high-confidence protein-coding loci exhibit several genomic features intermediate between those of lncRNAs and mRNAs. We experimentally validated the testicular expression pattern of two of these novel protein-coding gene candidates, both highly conserved in mammals: one for a vesicle-associated membrane protein we named VAMP-9, and the other for an enolase domain-containing protein. This study confirms the potential of PIT approaches for the discovery of protein-coding transcripts initially thought to be untranslated or unknown transcripts. Our results contribute to the understanding of spermatogenesis by characterizing two novel proteins, implicated by their strong expression in germ cells. The mass spectrometry proteomics data have been deposited with the ProteomeXchange Consortium under the data set identifier PXD000872.

proteomics, RNA profiling, spermatogenesis, testicular unannotated transcripts, transcriptome

INTRODUCTION

Spermatogenesis is a specialized and dynamic process facilitating the transmission of genetic inheritance [1]. It involves an intricate program of germ cell development, still poorly documented, that is dependent upon the successive activation and/or repression of thousands of gene products [2–4]. Consistent with the complexity of the process, the testis is one of the most complex organs in the body [5].

A large number of genes have been identified as being spatially and temporally regulated during postnatal testicular ontogenesis and germ cell differentiation by genome-wide transcriptional expression studies [2, 6–12]. Several groups recently launched a massive reexploration of the testicular transcriptome using next-generation sequencing technologies and discovered thousands of novel unannotated loci possibly important for spermatogenesis [13–20]. However, as little is known about the associated transcriptional events, these RNAs are usually annotated, arbitrarily, as being long noncoding RNAs (lncRNAs) [21–24]; they indeed share many traits with lncRNAs, including being relatively short and having few exons, a low GC content, only weak sequence conservation (comparable to that of introns), and a low abundance [25–29]. However, some of these transcripts may be translated to give proteins. It has been demonstrated that a comprehensive integration of Shotgun proteomics and next-generation sequencing data is informative about this possibility [30].

Shotgun mass spectrometry now routinely involves liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). Recent technological developments in mass spectrometry have led to a new generation of instruments with unprecedented resolution and sensitivity. As a consequence, it is now possible to establish extensive, indeed near-exhaustive, protein repertoires that unsurprisingly include several thousand nonredundant proteins from a total cell lysate and in a single run [31, 32]. Recently, Evans and collaborators described a novel approach that significantly improves the power of proteomic exploration: they termed this strategy ''proteomics informed by transcriptomics'' (PIT) [30].

The PIT approach is based on the assumption that established sequence databases used for protein identification are incomplete. The query of mass spectrometry data using sequence databases such as UniProt [33] inevitably leads to a loss of information deriving from the MS/MS data itself. The importance of such losses cannot be anticipated: if the sequence from which they are derived is not present in databases, MS/MS spectra that cannot be assigned to any theoretical peptide derived from virtual trypsin digestion of these sequences will not lead to any protein identification.

[1]The Proteomics Core facility Biogenouest is supported by Infrastructures en Biologie Santé et Agronomie (IBiSA), Région Bretagne, Fonds Européen de Développement Régional and Conseil Régional de Bretagne structural funding awarded to C.P. Aspects of this work were supported by l'Institut national de la santé et de la recherche médicale (Inserm); l'Université de Rennes 1; l'Ecole des hautes études en santé publique (EHESP); the grant INERIS-STORM awarded to B.J. (grant number N 10028NN); and the Rennes Métropole grant "Défis scientifiques émergents-2011" and the PNR EST 2013 grant (Anses, grant number DBI20131228558) awarded to F.C.
[2]Correspondence: Charles Pineau, Proteomics Core Facility Biogenouest, Inserm U1085-IRSET, Université de Rennes 1, 263 av. du Général Leclerc, 35042 Rennes cedex, France.
E-mail: charles.pineau@inserm.fr
[3]Correspondence: Frédéric Chalmel, Inserm U1085-IRSET, Université de Rennes 1, 263 av. du Général Leclerc, 35042 Rennes cedex, France.
E-mail: frederic.chalmel@inserm.fr
[4]These authors contributed equally to this work.

Received: 18 June 2014.
First decision: 8 July 2014.
Accepted: 11 August 2014.
© 2014 by the Society for the Study of Reproduction, Inc.
This is an Open Access article, freely available through *Biology of Reproduction's* Authors' Choice option.
eISSN: 1529-7268 http://www.biolreprod.org
ISSN: 0006-3363

Downloaded from www.biolreprod.org.

However, mass spectrometry-based protein identification using customized theoretical translations of de novo assembled transcripts from RNA-seq experiments can greatly improve the sensitivity of peptide identification [34, 35]. The usefulness of PIT strategies for improving genome annotation by characterizing novel genes, novel exons, novel splicing events, translated UTRs, frame shifts, and reverse strands has now been demonstrated [30, 34, 36, 37]. However, this approach has not been applied to the discovery and identification of novel germ cell proteins in any model organism.

We therefore used a PIT strategy that combined recently published RNA-seq data from isolated rat germ cells [13] with a Shotgun proteome analysis of rat spermatocytes and spermatids. We discovered 44 novel protein-coding loci expressed in meiotic and postmeiotic germ cells, whose corresponding proteins were identified by mass spectrometry. As a proof of concept, two of the candidates selected on the basis of interesting features were further validated experimentally. Our study significantly improves the genome annotation of a model organism, and it reveals novel players, possibly central to germ cell physiology and male fertility.

MATERIALS AND METHODS

Ethics Statement

Experimental procedures reported here were performed in conformity with the principles for the use and care of laboratory animals in compliance with French and European regulations on animal welfare and were approved by the Rennes Animal Experimentation Ethics Committee. The investigators were appropriately authorized by the French ''Direction des Services Vétérinaires'' to conduct or supervise experimentation on live animals. Human materials were obtained at Rennes University Hospital from patients seronegative for HIV-1: normal testis and epididymis samples were collected at autopsy. Normal seminal vesicles were collected from patients who underwent radical prostatectomy and had not received hormone treatment; prostate tissues were obtained from otherwise healthy men who underwent prostatic adenomectomy for BPH. The local ethics committee approved the study protocol, ''Study of Normal and Pathological Human Spermatogenesis,'' registered under No. PFS09-015 at the French Biomedicine Agency, and informed consent was obtained from all donors as appropriate.

Animals

Male Sprague-Dawley rats of various ages were used as sources of tissue samples and for testicular cell isolation, in situ hybridization, and immunohistochemical experiments. Animals were purchased from Elevage Janvier.

Isolation of Testicular Cells

Pachytene spermatocytes (pSPC) and early spermatids were prepared by centrifugal elutriation with a purity greater than 90% according to a previously described method [38] except that enzymatic dissociation of cells was replaced by mechanical dispersion. Spermatogonia were isolated from 9-day-old rat testes according to sedimentation velocity at unit gravity [39]. Sertoli cells (SC) were isolated from 20-day-old rat testes according to previously described methods [40, 41]. For RNA extraction, protein extraction, and Western blot experiments, cells were gently pelleted, snap frozen in liquid nitrogen upon isolation, and stored at −80°C until use. Rat tissues (testis, bone marrow, brain, kidney, liver, lung, and muscle) and human tissues (testis, epididymis, prostate, and seminal vesicles) used for RT-PCR were frozen in liquid nitrogen and stored at −80°C until analysis.

RNA-Sequencing Analysis

Comprehensive database of known transcripts. Transcript annotations from public databases (Ensembl [42]; National Center for Biotechnology Information [NCBI] release RGSC3.4 [43]; AceView [44]; and mRNA data from University of California at Santa Cruz [UCSC] rn4 [45]) were merged into a combined set of nonredundant known transcript annotations using Cuffcompare [46].

Read mapping. Briefly, and as previously described [13], RNA-seq-derived reads from each sample replicate were aligned independently to the *Rattus norvegicus* genome (rn4, downloaded from the UCSC genome browser website [45] with TopHat (version 1.4.1) [47] using published approaches [29, 46]. The database of known transcripts (see above) and expressed sequence tag (EST) alignments (from UCSC) was used to define an additional junction set (AJS) for each TopHat run. The junction outputs from individual TopHat runs were pooled and added to the AJS to allow TopHat to use junction information from all samples. TopHat was run again for each sample using the resulting AJS. The output of this second run comprised the final alignment.

Ab initio transcriptome assembly. Individual sample alignments for each testicular cell type were pooled. The transcriptome of each individual cell type was assembled with Cufflinks (version 1.2.0) by finding a parsimonious allocation of reads of the transcripts within a locus using default settings [46, 48]. Next, the Cuffcompare program was used to merge the individual transcript fragments (transfrags) into a combined set (nonredundant ''union'' of all transfrags that share all introns and exons) of 99 438 assembled transcripts; this set was named the rat nonredundant reference transcriptome.

Transfrag quantification. The isoform-level abundances (expression level) were assessed using Cuffdiff [46, 48] for each sample with upper quantile normalization. Abundance was measured in fragments per kilobase of exon model per million reads mapped (FPKM). A matrix of FPKM values was then prepared from the results of transcriptome quantification. The data were quantile normalized to reduce systematic effects and allow direct comparison between the individual samples.

Transfrag classification. The Cuffcompare program [46, 48] was used to classify the 99 438 transfrags belonging to the rat nonredundant reference transcriptome according to the known transcript annotation database. All long (cumulative exon length ≥200 nucleotides [nt]) transcripts that were annotated automatically as complete match (Cuffcompare class ''='') or potentially novel isoform (''j'') of annotated noncoding genes and all novel intronic (''i,'' i.e., falling entirely within a reference intron and without exon-exon overlap with another known locus) or intergenic (''u'') loci were selected, and this led to 31 582 genes (34 458 transfrags) being included in the analysis.

Creation of the Rat Nonredundant Reference Proteome Database

The nucleotide sequences of the 99 438 assembled transcript isoforms were translated into the six possible reading frames using the Transeq program (EMBOSS suite of tools) [49]. Deduced amino acid sequences of at least 10 residues between two stop codons were defined as potential protein sequences; there were 3 348 184 such predicted protein sequences. We assembled a rat nonredundant reference proteome by merging the UniProt (37 175 canonical and isoform sequences; release 2012_10) [33] and Ensembl (32 971 known and 44 993 predicted protein sequences; release 3.4.68) [50] proteome databases with the set of predicted protein sequences.

Mass Spectrometry Analysis

Protein extraction. Frozen cell pellets of rat pSPC and round spermatids (rSPT) were resuspended in extraction buffer (100 mM PIPES, 70 mM NaCl, 2 mM $MgCl_2$, pH 7.4). A cocktail of protease inhibitors with 1 mM EDTA, 0.5 mM dithiothreitol (DTT), 1 mM 4-(2-aminoethyl) benzenesulfonyl fluoride hydrochloride, 10 mM *trans*-epoxysuccinyl-leucylamido(4-guanidino)butane, 0.6 U/ml nuclease, and 2% (v/v) Nonidet P-40 (Sigma-Aldrich) was added to the extraction buffer just before use. Cell suspensions were subjected to sonication on ice. The resulting lysates were centrifuged at 1000 × *g* and 4°C for 10 min to remove cellular debris. The supernatants were then centrifuged at 105 000 × *g* at 4°C for 1 h. The supernatants containing the soluble proteins were kept on ice, and the pellets were retrieved in 100 mM Na_2CO_3 and sonicated as described above. These suspensions were again centrifuged at 105 000 × *g*, 4°C for 45 min and the supernatants pooled with those containing the soluble proteins from the first extraction. The protein concentration was determined using the Bradford colorimetric assay (Bio-Rad), and protein extracts were stored at −80°C until use.

Protein prefractionation and digestion. Aliquots of 100 μg of total proteins from spermatocytes and spermatids were denatured at 70°C for 10 min in a LDS NuPage Sample buffer (Invitrogen) with 50 mM DTT. Proteins were separated by 12% SDS-PAGE (NuPage Novex Bis Tris Mini Gel; Invitrogen), in MES SDS Running Buffer. The gels were stained with Coomassie blue (EZBlue; Sigma-Aldrich) for 45 min, and destained overnight. Each gel lane was manually cut into 20 (for spermatid extracts) or 21 (for spermatocyte extracts) slices of approximately the same size. The proteins in the gel slices were reduced, alkylated, and digested with modified trypsin (Promega) and the peptides extracted as previously described [51].

Data acquisition. MS measurements of peptide extracts were performed with a nanoflow HPLC system (Ultimate 3000; Thermo Scientific Dionex)

Downloaded from www.biolreprod.org.

connected to a hybrid LTQ-OrbiTrap XL (Thermo Fisher Scientific) mass spectrometer equipped with a nano electrospray ion source (New Objective). The MS instrument was operated in its data-dependent mode by automatically switching between full-survey-scan MS and consecutive MS/MS acquisition. Survey full-scan MS spectra (mass range 400–2000) were acquired in the OrbiTrap section of the instrument with a resolution of r = 60 000 at m/z 400; ion injection times were calculated for each spectrum to allow for accumulation of 106 ions in the OrbiTrap. The seven most intense peptide ions in each survey scan with an intensity above 2000 counts (to avoid triggering fragmentation too early during the peptide elution profile) and a charge state ≥2 were sequentially isolated at a target value of 10 000 and fragmented in the linear ion trap by collision-induced dissociation. Normalized collision energy was set to 35% with an activation time of 30 milliseconds. Peaks selected for fragmentation were automatically put on a dynamic exclusion list for 120 sec with a mass tolerance of ±10 ppm to avoid selecting the same ion for fragmentation more than once. The repeat count was set to 1, the exclusion list size limit was 500, singly charged precursors were rejected, and the maximum injection time was set at 500 and 300 ms for full MS and MS/MS scan events, respectively. For an optimal duty cycle, the fragment ion spectra were recorded in the LTQ mass spectrometer in parallel with the OrbiTrap full scan detection. For OrbiTrap measurements, an external calibration was used before each injection series, ensuring an overall mass accuracy error below 5 ppm for the detected peptides. MS data were saved in RAW file format (Thermo Fisher Scientific) using XCalibur 2.0.7 with Tune 2.4.

Data processing. Three successive LC-MS/MS runs and dynamic exclusion were employed to prevent repetitive selection of the same peptide. Proteome Discoverer software (version 1.2; Thermo Fisher Scientific) supported by the Mascot (Matrix Science) search engine was used for peptide and protein identification. MS/MS spectra were used to search our rat nonredundant reference proteome database (number of residues: 161 273 757; number of sequences: 3 428 361) and also the randomized version of this database (decoy) to determine the false-positive rate, defined as the number of validated decoy hits/(number of validated target hits + number of decoy hits) ∗ 100, using the Mascot algorithm (Mascot server v2.2.07). Mass tolerance for MS and MS/MS was set at 10 ppm and 0.5 Da, respectively. Enzyme selectivity was set to full trypsin with one missed cleavage allowed. Carbamidomethylation of cysteines was considered as a fixed protein modification, whereas oxidation of methionine, acetylation of lysine, and phosphorylation of serine, threonine, and tyrosine were considered as variable modifications. Peptide identifications extracted from Mascot result files were validated at a final peptide false-discovery rate (FDR) of 1%. During the dynamic exclusion process, lists of peptides not filtered out were exported as a text file containing uncharged and accurate mass values to four decimal places and a retention time window of approximately 1 min. The mass spectrometer was configured to work with uncharged masses and automatically to calculate a peptide mass based on its exact mass and charge state. A mass tolerance of ±10 ppm was used to reject previously identified peptides within the specified retention time window [51].

Protein identification. All MS/MS spectra were used to search our rat nonredundant reference proteome database and the decoy database in a single Mascot query to generate one compiled search file (.msf file) per run and per sample. Identified peptides were filtered according to the Mascot score to obtain a FDR of 1%. Peptide identifications were accepted if the individual ion Mascot scores were above the identity threshold (the ion score is $-10 * \log(P)$, where P is the probability that the observed match is a random event, P value $<$ 0.05). In the case of peptides shared by different proteins, proteins were automatically grouped. Only the best matches of the peptides (rank 1) were considered. The proteins within a group were ranked according to their protein score. The protein reported in the protein table (Supplemental Data S1; Supplemental Data are available online at www.biolreprod.org) corresponds to the top score protein. Each search file (.msf) was exported as a protXML format file. The peptide and final protein lists, together with the associated description, are reported in Supplemental Data S1 and S2.

Refinement of Transfrag Selection

To select high-confidence novel protein-coding loci and thus to eliminate artifacts due to errors in read mapping, transcript assembly, and protein identification, we applied an additional filtering step. Briefly, we defined a background expression cutoff (BEC = 3.72 FPKM), calculated as the overall median of FPKM values for the assembled transcripts that completely match (Cuffcompare class "=") RefSeq curated mRNAs (RefSeq category "NM") [43]. This allowed the selection of "expressed" or "detectable" transfrags for which FPKM values in both replicates of a given cell type were ≥BEC.

Statistical Filtration and Cluster Analysis

The transfrags expressed differently in four testicular cell types (SC, spermatogonia [SPG], pSPC, and rSPT) were filtered statistically using the AMEN (Annotation, Mapping, Expression and Network) suite of tools [52] according to the following criteria: Transfrags that exhibited a ≥3-fold difference in expression between averaged cellular conditions (pairwise comparisons) were selected first. Transfrags with significant differential expression were then identified using a LIMMA statistical test [53] and an F value adjusted with the FDR method: $P \leq 0.01$. Selected transfrags were then grouped into six expression patterns (P1–P6) using the Partitioning Around Medoids algorithm. The P4–P6 patterns were those involving preferential expression in spermatocytes and/or rSPT, as described in our previous study [13].

Multiple Alignments of Protein Sequences

Protein sequences predicted from selected candidates were used as probes to search UniProt, Ensembl and RefSeq protein databases using the BlastP program [54]. Additional orthologous sequences were retrieved and predicted by querying the Ensembl genome, the EMBL nucleic sequence databases [55], and EST databases using the TBlastN program. Multiple sequence alignments were then generated using the MAFFT module [56] implemented in the JalView editor [57] with default parameters. The CD-search program was used to predict protein domains for the two protein sequence candidates studied in greater detail [58, 59], and the JNet algorithm implemented in JalView was used to predict secondary structures [60].

Experimental Validation

RT-PCR and real-time quantitative RT-PCR. Complementary DNA was obtained from aliquots of 4 μg of DNase-treated RNA (DNase I; Promega) using random hexamers and Moloney murine leukemia virus reverse transcriptase (Invitrogen). Two sets of primer pairs were synthesized for PCR or real-time PCR. All primers were purchased from Sigma-Aldrich (Supplemental Data S3 lists the primers and the expected sizes of PCR products and efficiency of the primer pairs for qPCR). Conventional PCR was performed using Taq polymerase (Qiagen), or high-fidelity HotStar HiFidelity DNA Polymerase (Qiagen) for cloning, in a Peltier thermocycler (Labgene). PCR products were then resolved on 1.5% agarose gels. Real-time PCR was performed using the ABI 7500 Fast Real-Time PCR System (Applied Biosystems) in the presence of SYBR green. All samples were studied in triplicate. Specificity of the product amplification was confirmed by melting curve analyses and agarose gel electrophoresis. A stable gene in testicular cells (Snx17) was selected for normalization based on its low coefficient of variance in microarray data of germ and somatic cells. Relative expressions were calculated according to the delta Ct method.

Gene cloning and recombinant protein production. The XLOC_001949 PCR product was resolved by electrophoresis on a 1.2% agarose gel, excised, and purified using the QIAquick Gel Extraction Kit (Qiagen). The purified DNA was digested with *Bam*H1-HF in the appropriate CutSmart Buffer (New England Biolabs) to obtain the sequences of interest as *Bam*H1 fragments; these fragments were inserted into a pQE-30 vector (Qiagen) using the T4 DNA ligase in the Quick ligation buffer (New England Biolabs), according to the manufacturer's instructions. Integrity and insertion orientation of the cDNA within the pQE-30 vector were checked by single-read sequencing of the two strands (Eurofins mwg Operon). Supercompetent bacteria of strain XL1-blue (Stratagene) were transformed with the pQE-30/XLOC_001949 DNA and cultured in Luria Broth medium supplemented with 100 μg/ml ampicillin and 1 mM isopropyl β-D-1-thiogalactopyranoside to induce recombinant protein production. The bacterial cells were harvested and lysed, and the soluble XLOC_001949 6 His-tagged recombinant protein ($_{rec}$XLOC_001949) present in the supernatant was purified by affinity on Ni-NTA agarose slurry (Qiagen), using a 250 mM imidazole elution buffer according to the manufacturer's protocol. The purity of $_{rec}$XLOC_001949 was evaluated by SDS-PAGE, and the nature of the recombinant protein was checked by tryptic digestion and nano LC-MS/MS on a HCT Ultra PTM Discovery (Bruker Daltonik, GmbH) ion trap mass spectrometer.

Antibody production. Antibodies against the $_{rec}$XLOC_001949 protein were raised in rabbit, using the 28-Day Super Speedy Polyclonal Antibody Protocol (Eurogentec).

In Situ Expression Analyses

In situ hybridization and immunohistochemistry experiments were performed with testes from adult male Sprague-Dawley rats fixed in Bouin

Downloaded from www.biolreprod.org.

fixative and embedded in paraffin, as previously described [61]. Adult male rats under pentobarbital anesthesia were perfused via the left ventricle with PBS containing heparin (10 U/ml) for 5 min and then with Bouin solution (Microm Microtech) for 20 min. Testes were isolated and immersed in the same fixative for 6 h. The specimens were dehydrated in a graded series of ethanol concentrations, in butanol, and then embedded in paraffin. Sections 5 μm thick were cut and mounted onto poly-L-Lysine-coated slides.

In situ hybridization. RT-PCR products corresponding to XLOC_001949 and XLOC_013843 were gel purified using the Qiaquick Gel Extraction Kit (Qiagen), inserted into the pCR II-TOPO vector (Life Technologies), and used to transform competent XL1 blue bacteria. The constructs were screened by PCR and sequenced. Sense (T7 RNA polymerase on *Bam*H1 linearized vector) and antisense (Sp6 RNA polymerase on *Xho*I linearized vector) riboprobes were generated and labeled with digoxigenin-UTP (Boehringer Mannheim). The abundance of transcripts of the XLOC_001949 and XLOC_013843 constructs was then evaluated by in situ hybridization with antisense or sense riboprobes at 0.8 ng/μl. The hybridization signal was detected with an alkaline phosphatase-conjugated anti-digoxigenin antibody at 1:500 (Boehringer Mannheim) and visualized by incubation for 16 h at room temperature with S-bromo-4-chloro-3-indolyl phosphate (50 mg/ml) and nitroblue tetrazolium (75 mg/ml) as substrates (Boehringer Mannheim).

Immunohistochemical experiments. Tissue sections were incubated for 2 h at room temperature with the anti-$_{rec}$XLOC_001949 antibody used at a final dilution of 1:4000. After several washes in TBS, sections were incubated for 1 h with the Peroxidase/DAB Rabbit/Mouse EnVision solution (Dako), and developed with a diaminobenzidine solution (Sigma-Aldrich). The sections were counterstained with Masson hemalun, dehydrated, and mounted in Eukitt (Labnord). Rabbit preimmune serum was used as a negative control at a 1:2000 dilution.

Data Access

The RNA-seq data files were submitted to NCBI's Sequence Read Archive and to the NCBI Gene Expression Omnibus under accession numbers SRP026340 and GSE48321, respectively. The mass spectrometry proteomics data were deposited with the ProteomeXchange Consortium with the data set identifier PXD000872. Sequences of the identified peptides were mapped on the rat genome using BLAT [62] and subsequently deposited in the ReproGenomics Viewer (http://rgv.genouest.org). Selected unannotated transfrags were deposited with the GenBank Transcriptome Shotgun Assembly Sequence Database as bioproject no. PRJNA209702.

RESULTS

Experimental Design and PIT Workflow

A large set of lncRNAs and testicular unannotated transcripts (TUTs) has previously been characterized by high-resolution expression profiling of male germ cells in the rat using next-generation sequencing [13]. We used a PIT strategy to identify those transcripts that code for proteins, as summarized in Figure 1.

A RNA-seq analysis was conducted previously to characterize the transcriptome of four testicular cell types: SC, SPG, pSPC, and rSPT [13]. Briefly, the paired-end reads resulting from the sequencing of RNAs were aligned on the rat genome sequence and assembled into nonredundant transfrags. Reconstructed transcripts were then translated into protein sequences and a database of predicted open reading frames (ORFs) was generated. This data set was merged with a canonical list of rat proteins from UniProt (37 175 canonical and isoform sequences; release 2012_10) and Ensembl (32 971 known and 44 993 predicted protein sequences; release 3.4.68), to generate a rat nonredundant reference proteome. In parallel, comprehensive rat proteomes from isolated pSPC and rSPT were generated by shotgun proteomic analyses. Briefly, protein extracts were sequentially prefractionated, digested with trypsin, and analyzed by LC-MS/MS. The spectral data were then used to search the rat nonredundant reference proteome with the Proteome Discoverer software using the Mascot algorithm. A highly stringent refinement strategy was developed to select those protein identifications corresponding to novel unannotated or noncoding transcriptional events (TUTs and lncRNAs) that were unambiguously detected in meiotic and/or postmeiotic germ cells.

The Rat Nonredundant Reference Transcriptome Contained About 100 000 Transcripts and the Proteome 3 000 000 Predicted ORFs

Of the 140 million paired-end reads resulting from the Illumina sequencing experiment [13] about 80% were properly aligned on the rat genome (Fig. 2). These reads were subsequently assembled into a unique set of 99 438 transfrags, termed the rat nonredundant reference transcriptome (Fig. 2): 32 024 of these (29 668 loci) were classified as novel intronic (cuffcompare class code ''i'') or intergenic (class code ''u'') long (cumulative exon length $\geq$ 200 nt) TUTs, and 1274 (795 loci) appeared to correspond to known (class code ''=''') or novel (class code ''j'') isoforms of annotated, long, noncoding genes (lncRNAs) (Fig. 2).

We then predicted about 3.3 million ORFs ($\geq$10 amino acids [aa]) from all six frames of each reconstructed transcript, and the deduced aa sequences were combined with UniProt (release 2012_10) and Ensembl (release 3.4.68) public proteome databases, leading to a unique set of 3 428 361 proteins that we used as our customized rat nonredundant reference proteome.

The PIT and Refinement Strategies Revealed 44 Novel Potentially Protein-Coding Loci

Mascot and a target decoy strategy were used for searches with the MS/MS spectra data from pSPC and rSPT protein extracts against the rat nonredundant reference proteome. This led to the identification of 19 966 nonredundant peptides corresponding to 4999 nonredundant proteins in pSPC (16 611 peptides; 4056 proteins) and/or rSPT (11 755 peptides; 3061 proteins) (Supplemental Data S1 and S2).

To identify the most likely protein-coding candidates associated with a germ cell line expression pattern, an additional refinement step based on transcript abundance was applied. This resulted in a final set of 5379 long, nonredundant transcripts (4065 loci) significantly expressed in pSPC and/or rSPT: they included 4458 TUTs (3559 loci) (Fig. 2, left) and 921 lncRNAs (506 loci) (Fig. 2, right). Mass spectrometry identified translation products for 69 of these transcripts (44 loci), including 48 TUTs (30 loci) and 21 lncRNAs (14 loci), with at least one high-confidence peptide (rank 1; 1% FDR) in isolated pSPC and/or rSPT; these 44 loci were therefore qualified as MS identified. They included 15 TUTs (12 loci) and 16 lncRNAs (12 loci) that were preferentially expressed in SPC and/or SPT (patterns P4–P6) (Fig. 2).

The Features of MS-Identified Transcripts Diverge Significantly from Those Typical of Known lncRNAs

TUTs share many genomic characteristics with known lncRNAs in vertebrates [13], including relatively short length, low exon number, low GC content, low sequence conservation (comparable to that of introns), low abundance, and highly temporally and spatially restricted expression patterns [25–29, 63–65]. To determine whether or not the 69 MS-identified TUTs and lncRNAs share features with nonidentified transcripts expressed in meiotic and postmeiotic germ cells, we compared a list of genomic traits between transcript populations (Fig. 3). This analysis included the 5355 MS-identified mRNAs, those annotated protein-coding transcripts assembled

Downloaded from www.biolreprod.org.

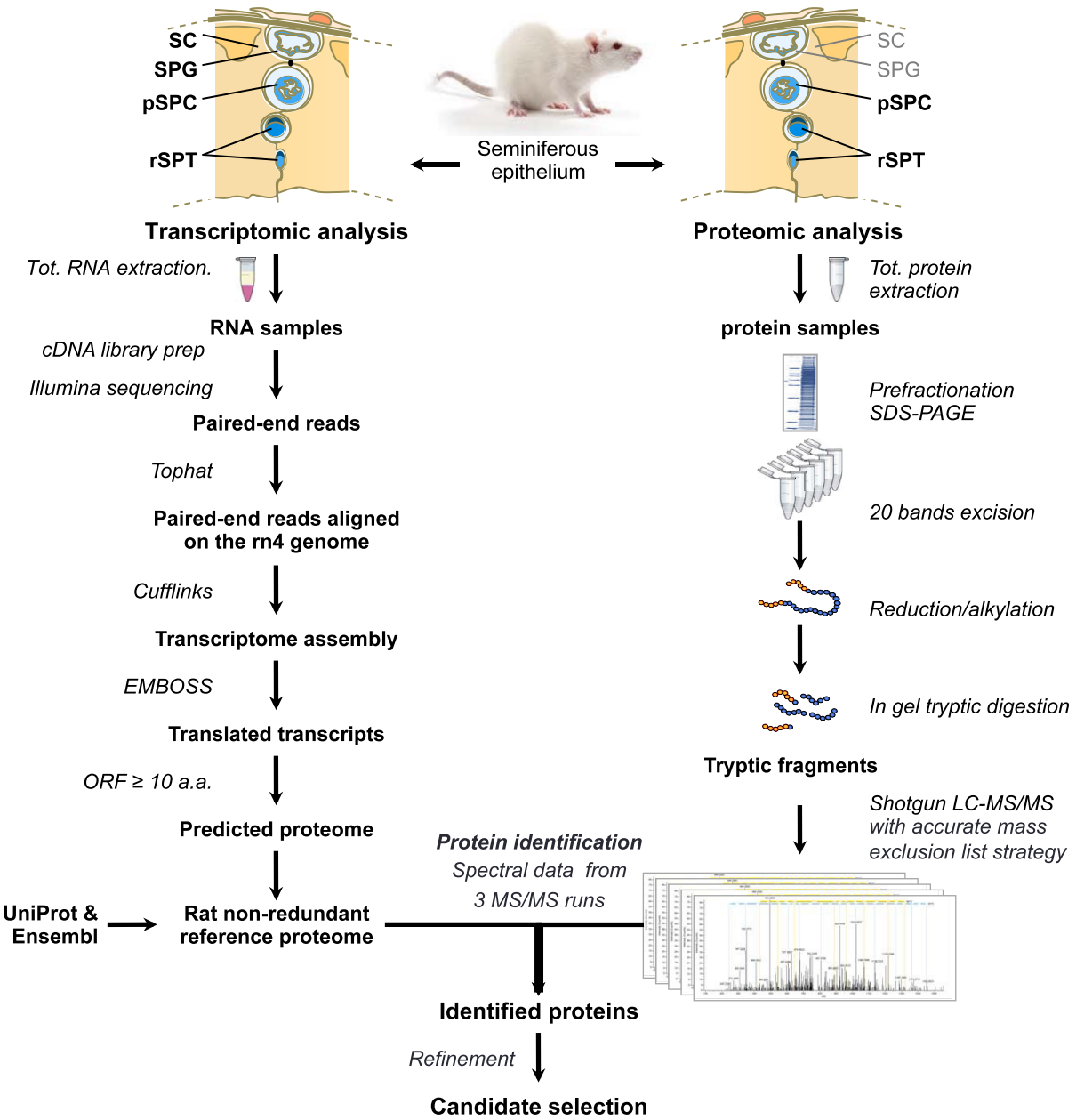

FIG. 1. Experimental design and PIT workflow. A schematic diagram of the strategy used to identify novel protein-coding loci by combining transcriptomic and proteomic data analysis.

in our RNA-seq data set and for which at least one high-confidence peptide was identified by mass spectrometry in pSPC and/or rSPT extracts.

Size characteristics and number of isoforms. MS-identified TUTs or lncRNAs (first quartile [q1] = 660 bp, median [med] = 905 bp, third quartile [q3] = 2198 bp) were significantly longer than non-MS-identified transcripts (q1 = 389 nt, med = 599 nt, q3 = 1004 nt; P value for Wilcoxon signed-rank test $< 2.10^{-8}$) but shorter than MS-identified mRNAs (q1 = 1093 nt, med = 1805 nt, q3 = 3049 nt; $P < 2.10^{-5}$) (Fig. 3A). The number of exons for MS-identified (med = 4 exons) was significantly greater than for non-MS-identified TUTs and lncRNAs (med = 2; $P < 8.10^{-9}$) but lower than for MS-identified mRNAs (med = 9, $P < 3.10^{-12}$) (Fig. 3B). Maximum ORF length was significantly longer for MS-identified transcripts (q1 = 107 aa, med = 134 aa, q3 = 335 aa) than non-MS-identified transcripts (q1 = 75 aa, med = 94 aa, q3 = 121 aa; $P < 9.10^{-13}$) but without being as long as those in MS-identified mRNAs (q1 = 229 aa, med = 374 aa, q3 = 653 aa; $P < 2.10^{-12}$) (Fig. 3C). The numbers of spliced isoforms for MS-identified TUTs and lncRNAs (q1 = 1.0, med = 2.0, q3 = 4.0) were about double those for non-MS-identified transcripts (q1 = 1, med = 1, q3 = 2; $P < 2.10^{-3}$) but only two thirds those for MS-identified mRNAs (q1 = 2.0, med = 3.0, q3 = 5.0; $P < 3.10^{-2}$) (Fig. 3D).

Sequence conservation. To assess the conservation of exons and introns in MS-identified transcripts and that in non-MS-identified transcripts, we compared the averaged base-by-base phastCons conservation score calculated for nine vertebrates, as provided by the UCSC genome browser [45].

Downloaded from www.biolreprod.org.

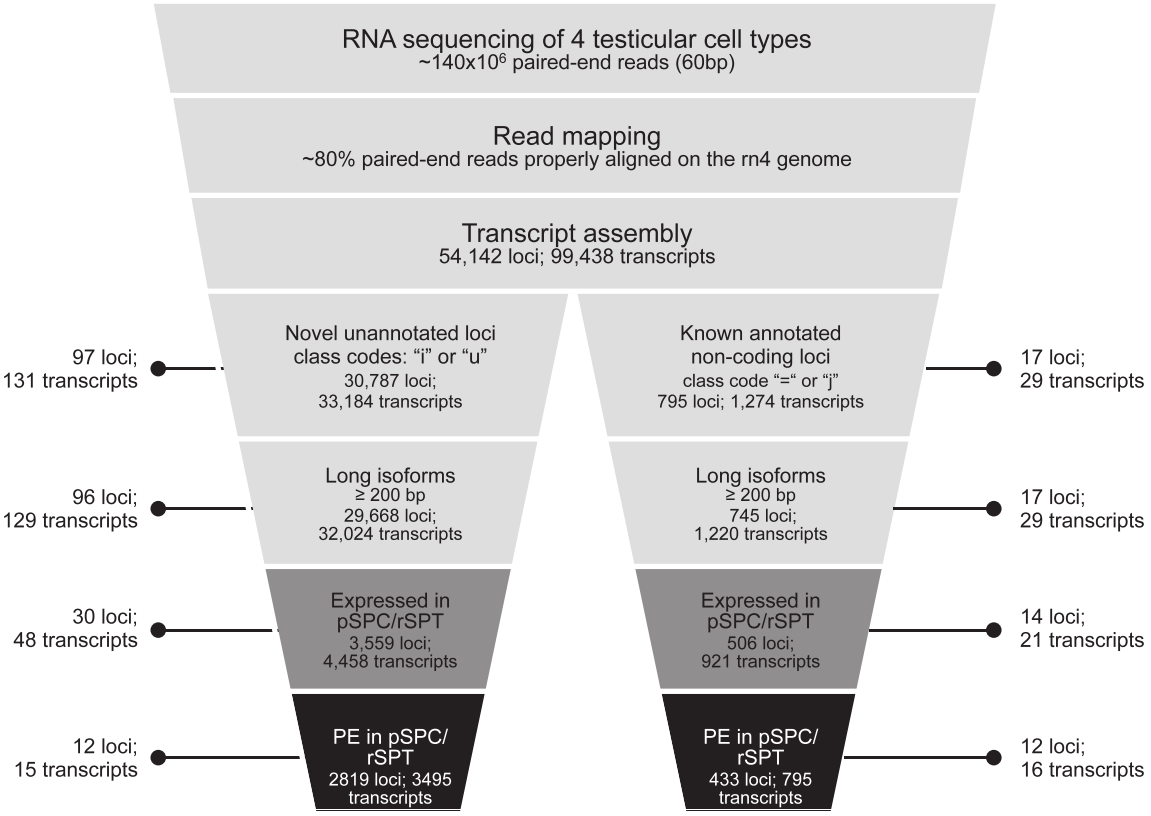

FIG. 2. PIT and refinement strategies used to select novel high-confidence protein-coding loci expressed in germ cells. From top to bottom, Illumina sequencing generated millions of 60-bp paired-end reads, which are aligned to the rat genome (release rn4) and subsequently assembled into cell-specific transcriptomes as described in the previous study [13]. We next focused on novel testicular unannotated loci (left side, class codes "i" and "u") and lncRNAs (right side, class codes "=" and "j"). We used a three-step refinement strategy to select a high-confidence set of long (≥200 bp) and detectable (in pSPC and/or rSPT) transcripts showing a peak of expression during meiotic and/or postmeiotic stages. At each step, the numbers of loci and transcripts are given. The numbers of loci and transcripts for which a predicted ORF was detected by LC-MS/MS is indicated by round-head arrows on each side of the figure.

Exon conservation was higher and intron conservation slightly lower in MS-identified (median exon conservation of 0.2 and 0.0 for intron conservation) than non-MS-identified (0.0 for exon, $P < 6.10^{-8}$; 0.0 for intron, $P < 6.10^{-3}$) TUTs and lncRNAs (Fig. 3, E and F). MS-identified mRNAs (0.7 for exon, $P < 3.10^{-20}$; 0.1 for intron, $P < 2.10^{-10}$) showed a higher exon and intron conservation than MS-identified TUTs and lncRNAs.

Abundance and cell specificity. Unexpectedly, the expression level in testicular cells was higher for MS-identified TUTs and lncRNAs (median of the highest log2FPKM of 3.6) than for non-MS-identified transcripts (median of 3.1; $P < 5.10^{-4}$) and MS-identified mRNAs (median of 3.3; $P < 5.10^{-2}$) (Fig. 3G). An expression specificity score based on the Shannon (theoretical information measure) entropy Q was calculated as an estimate of the abundance specificity for the various testicular cell types [66] as previously suggested [25, 29]. MS-identified transcripts showed intermediate cell-type specificity (median Shannon entropy-based specificity score = 1.1) significantly lower than that for non-MS-identified transcripts (0.8; $P < 2.10^{-4}$) but higher than that for MS-identified mRNAs (1.4; $P < 9.10^{-6}$) (Fig. 3H).

Distance to neighboring protein-coding genes. We investigated the relationships between MS-identified TUTs and lncRNAs and their protein-coding neighbors, and compared them with those for nonidentified transcripts. We considered the nearest known upstream and downstream protein-coding genes without distance restriction. Non-MS-identified TUTs and lncRNAs (q1 = 1225, med = 12 946, q3 = 53 684) were about six times farther from any protein-coding genes than were MS-identified TUTs and lncRNAs (q1 = 332, med = 2178, q3 = 20 701; $P < 2.10^{-2}$) (Fig. 3I). The distance to neighboring protein-coding genes was not significantly different for MS-identified TUTs and lncRNAs and for MS-identified mRNAs ($P < 0.5$).

The Genes for VAMP9 and a Testicular Enolase Domain-Containing Protein, T-ENOL: Two Novel Protein-Coding Genes Expressed in Meiotic and Postmeiotic Germ Cells

The main objective of our study was to demonstrate the existence of novel unannotated protein-coding loci by providing mass spectrometry evidence of the corresponding peptides/proteins. The validated data set is presented in Supplemental Data S4. Two TUTs were selected for further investigation to illustrate the relevance of our discovery strategy.

VAMP9 (locus XLOC_013843), a meiotic and postmeiotic VAMP7-like protein. The first selected locus (locus ID: XLOC_013843; transcript ID: TCONS_00035758) was identified from one high-confidence peptide in pSPC (protein sequence coverage ≈ 22.5%; E value $< 10^{-4}$). It maps on chromosome 14 (positions 10 975 584–10 984 162) and is composed of three exons with a cumulative exon size of 256

Downloaded from www.biolreprod.org.

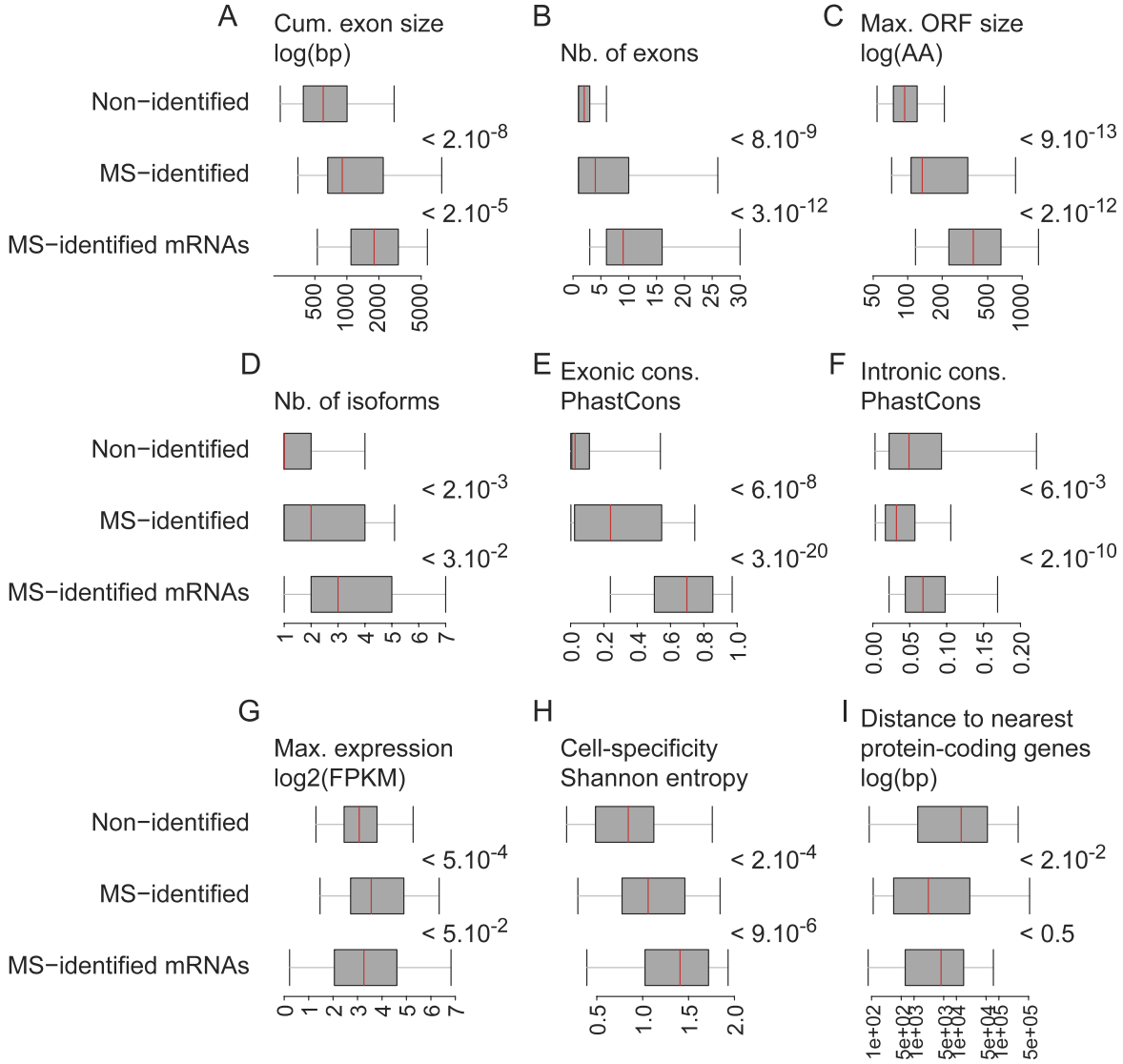

FIG. 3. Genomic and transcriptomic features of MS-identified transcripts. MS-identified TUTs, and known lncRNAs were compared to nonidentified TUTs and known lncRNAs and to MS-identified mRNAs. The box plots summarize the distributions of: cumulative exon length (**A**); number of exons (**B**); maximum ORF size in amino acids (**C**); number of isoforms (**D**); exon conservation (phastCons score; **E**); intron conservation (phastCons score; **F**); the maximum abundance in samples in log2(FPKM + 0.05) (**G**); cell-specificity measures based on Shannon entropy (**H**); and the distance to the nearest protein-coding gene (**I**). Note that the lower the value of Shannon entropy, the more the expression is restricted to one cell type. For **A** and **I**, lengths are shown in nucleotides (bp). For **A**, **C**, and **I**, x-axes are shown on a logarithmic scale.

nt and a maximum ORF size of 78 aa (Figs. 4A and 5A). This TUT was among the most conserved among vertebrates (phastCons score = 0.735; Figs. 4A and 5A). It shows a peak expression in pSPC based on the RNA-seq data set (Fig. 4A) that was confirmed by qPCR and RT-PCR using the four testicular cell populations (Fig. 4, B and C). The RT-PCR experiment also included seven healthy tissues; small amounts of the RNA were detected in somatic tissues (brain, liver and lung; Fig. 4C). Two distinct bands corresponding to two alternative isoforms were observed. Pachytene SPC up to rSPT and elongated spermatids in adult testis sections showed a cytoplasmic staining in in situ hybridization analysis (Fig. 4, D and E). Protein, genome, and EST database searches using Blast programs [54] unambiguously identified or predicted corresponding protein sequences in 13 vertebrates (Fig. 5A). This analysis indicated very strong conservation of the predicted ORF in mammals and also that the predicted secondary structures were conserved: five beta sheets and three helices. It also revealed a close paralogy relationship with the vesicle-associated membrane protein 7 (VAMP7, ~46% of sequence similarity). The similarity with VAMP7 was confirmed by the prediction of a SNARE-like domain (Pfam domain PF13774, E value = 0.02). We thus decided to name this novel protein-coding gene the vesicle-associated membrane protein 9 (VAMP9) gene. Analysis of identified ESTs indicated that the VAMP9 gene is expressed in five other mammalian species (mouse, dog, buffalo, pig, and wallaby); in four of these species, the ESTs were exclusively retrieved in the testis (Fig. 5A and Supplemental Data S5). RT-PCR evidenced substantial amounts of VAMP9 mRNA in both human and mouse testis (Supplemental Data S6).

T-ENOL (XLOC_001949), a novel mammalian meiotic and postmeiotic protein with a conserved enolase domain. The second selected candidate locus we investigated, T-ENOL

Downloaded from www.biolreprod.org.

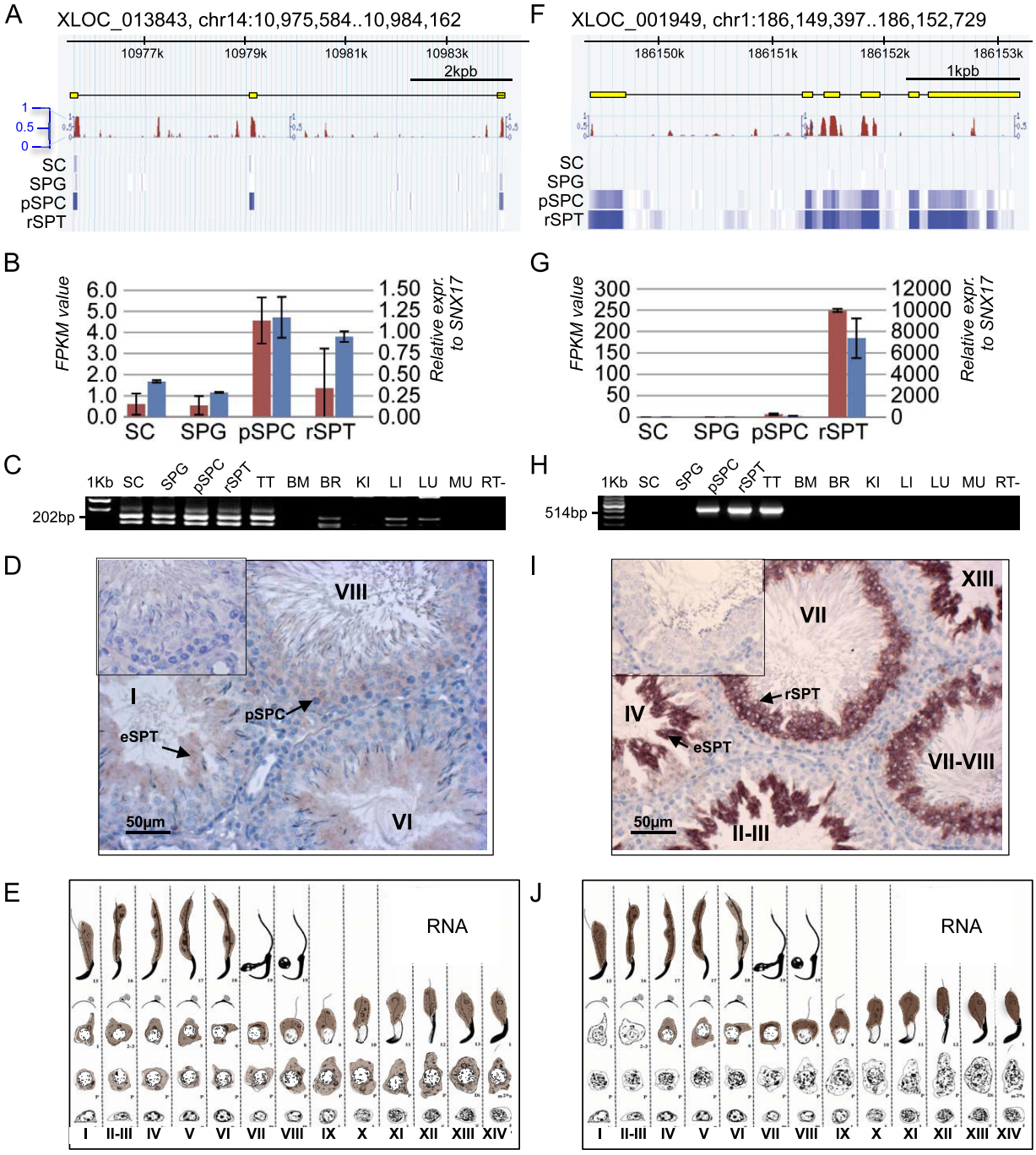

FIG. 4. Cell-specific expression patterns of XLOC_013843 (VAMP9) and XLOC_001949 (T-ENOL) transcripts. The expression patterns of XLOC_013843 (**A**–**E**) and XLOC_001949 (**F**–**J**) were investigated. **A** and **F**) The gene structure for both (yellow boxes correspond to exons), the sequence conservation between nine vertebrates as provided by the UCSC genome browser (phastCons scores, red histograms), and the transcript abundance determined by RNA-seq in four isolated testicular cell types as a color-coded blue heat map. **B** and **G**) The amounts of each transcript in each cell type according to the RNA-seq (left y-axis) and the quantitative RT-PCR (right y-axis) experiments. **C** and **H**) XLOC_013843 and XLOC_001949 transcript detection by RT-PCR in the total testis (TT), SC, SPG, pSPC and rSPT, and other tissues including bone marrow (BM), brain (BR), kidney (KI), liver (LI), lung (LU), and muscle (MU). RT–, reverse transcription negative control. **D** and **I**) Testicular in situ hybridization images with probes specific for the selected MS-identified transcripts. Insets: negative control images showing the absence of signal when sense ribonucleotide probes were used. Bars = 50 µm. **E** and **J**) A summary of XLOC_013843 and XLOC_001949 transcripts (respectively) in situ localization, superimposed on the map of spermatogenesis from Leblond and Clermont [97], as modified by Dym and Clermont [98] (reproduced with permission of John Wiley & Sons, Inc.).

(XLOC_001949/TCONS_00010279), was identified from one peptide in both pSPC and rSPT (protein sequence coverage ≈ 16%; E value $< 9.10^{-5}$). It maps to chromosome 1 (positions 186 149 397–186 152 729) and is composed of six exons with a cumulative exon size of 1045 nt and a maximum ORF size of 89 aa (Figs. 4F and 5B). Its sequence conservation among vertebrates is higher than those of other TUTs (phastCons score = 0.164), but lower than that of VAMP9 (Figs. 4F and 5B). RNA-seq data analysis indicated that this RNA was present in pSPC and rSPT (Fig. 4F), a finding confirmed by qPCR and

Downloaded from www.biolreprod.org.

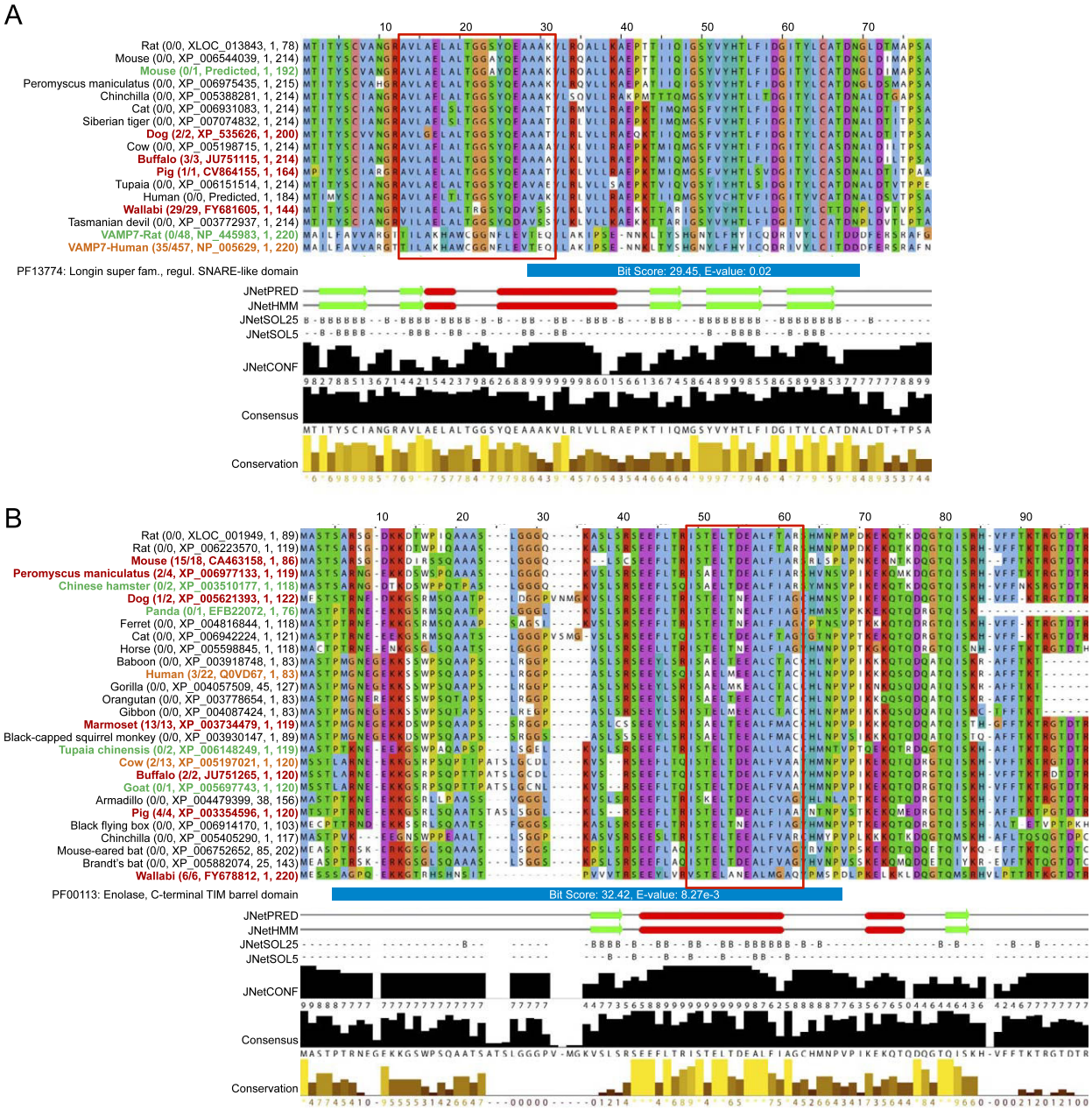

FIG. 5. Sequence conservation of the two proteins selected for detailed analysis, XLOC_013843 (VAMP9), and XLOC_001949 (T-ENOL). Predicted orthologous proteins of XLOC_013843 (**A**) and XLOC_001949 (**B**) in several mammalian species were retrieved using NCBI Blast programs with RefSeq, UniProt, and EST databases. Protein sequences were aligned using the MAFFT algorithm implemented in the JalView suite. Peptides identified by LC-MS/MS are indicated by a red rectangle on the multiple sequence alignments. Conserved residues are highlighted according to the default Clustal Color scheme as used by JalView (http://ekhidna.biocenter.helsinki.fi/pfam2/clustal_colours). On the left of each protein sequence, protein information is given as follows: organism (number of ESTs in testis /total number of EST, UniProt/ENSEMBL/RefSeq/EST identifiers, position of the first amino acid residue indicated on the alignment, total protein length). Predicted Pfam domains are depicted by blue rectangles below the alignments; the corresponding E-value and bit score are indicated within the blue rectangle. JNet structure predictions are displayed below the Pfam predictions. The annotation bars are as follows: J*NetPRE*D, the consensus prediction; JNetHMM, HMM profile based prediction; JNETSOL25 and JNETSOL5, solvent accessibility predictions (binary predictions of 25% or 5% solvent accessibility). The JNetCONF profile shows the confidence estimate for the prediction. High values mean high confidence prediction. Helices are marked as red tubes, and sheets as bright green arrows. A consensus prediction and a conservation profile with conservation scores from 0 to 10 (high values mean high conservation) are given below the JNetCONF profile. The presence of ESTs in each species is represented by a color code on the protein information: Black, no EST; green, presence of ESTs; orange, presence of ESTs sequenced in testis; and, red, a majority of ESTs sequenced in the testis.

RT-PCR (Fig. 4, G and H). The RT-PCR experiment also demonstrated that its expression was restricted to the testis (Fig. 4H). In situ hybridization analysis of adult testis sections showed that cytoplasmic staining increased from round SPT at stage IV to step 18-elongated spermatids (Fig. 4, I and J). Protein sequence analysis confirmed the good conservation of the predicted ORF and its predicted secondary structures (two beta sheets and two helices) in 26 other mammalian species, including 13 for which there is EST evidence of the corresponding transcript (Fig. 5B and Supplemental Data

Downloaded from www.biolreprod.org.

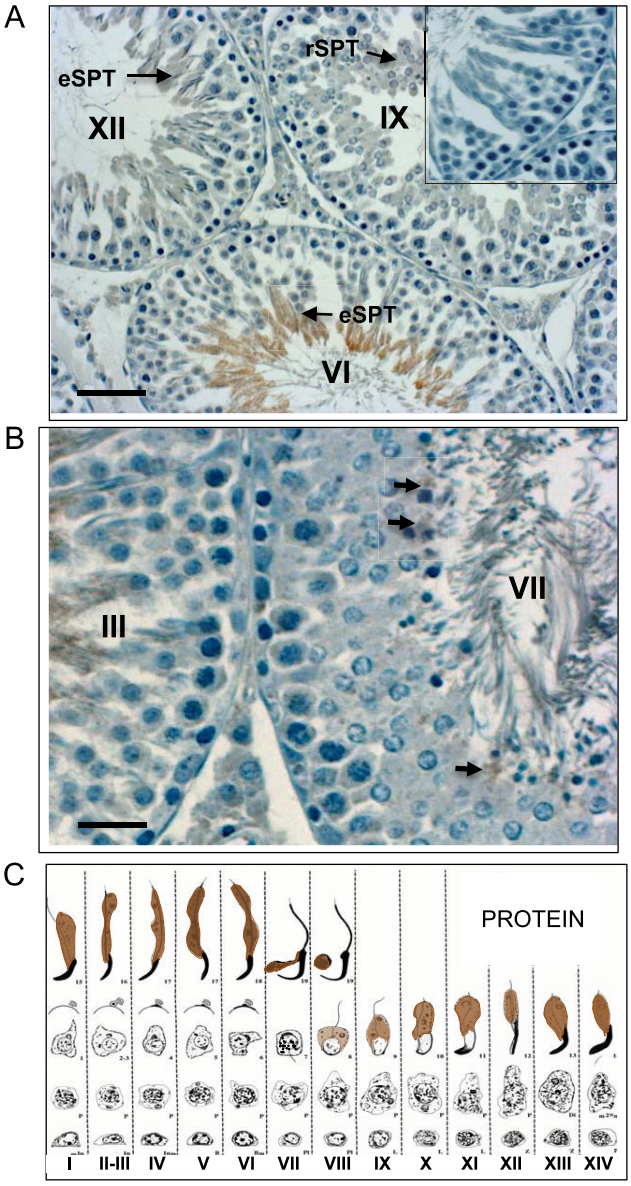

FIG. 6. Validation of the expression of XLOC_001949 (T-ENOL) by detection of a protein. **A** and **B**) Immunolocalization of the XLOC_001949 protein in adult rat testis sections as revealed with the anti $_{rec}$XLOC_001949 rabbit polyclonal antiserum. Serial rat testis sections were similarly probed with a preimmune serum as a negative control (inset in **A**). Roman numerals indicate seminiferous epithelium stages [97]. **A**) The weak immunoreactivity in rSPT at stage IX. A stronger signal was observed in elongating spermatids (eSPT), increasing from stage XII to stage VI. **B**) The localized immunoreactivity in the residual bodies (arrowheads) of spermiating spermatozoa at stage VIII. Bars = 50 μm (**A**) and 25 μm (**B**). **C**) Summarizes the distribution of the XLOC_001949 protein in situ superimposed on the map of spermatogenesis from Leblond and Clermont [97] as modified by [98] (reproduced with permission of John Wiley & Sons, Inc.).

S5). In seven mammals, a majority of EST has been sequenced in the testis. The expression of the TCONS_0012079 homolog transcript was verified by RT-PCR in the mouse and human testis (Supplemental Data S6). Domain prediction analysis indicated the presence of a highly significant enolase domain (pfam domain PF00113, E value = 8.27×10^{-3}) in the protein sequence. We further tested for the distribution of the protein during spermatogenesis by immunohistochemistry with a polyclonal antibody raised against the recombinant XLOC_001949 (T-ENOL) ORF. XLOC_001949 immunoreactivity in the adult rat testis was weak in rSPT (Fig. 6A) and stronger in elongating spermatids at all stages; localized immunoreactivity was also observed in the residual bodies of spermiating spermatozoa (Fig. 6B, arrowheads). These observations are summarized in Figure 6C.

DISCUSSION

RNA-seq has not previously been combined with shotgun proteomics to search for novel protein-coding genes in the field of male reproduction. Palmer and collaborators [67] combined RNA-seq and tandem mass spectrometry to identify novel sperm proteins in the red abalone. Red abalone is not an established model organism, so the potential for discoveries was very large. We are unaware of any studies using a PIT approach to identify novel proteins in the germ cell lineage of a model organism.

Illumina next-generation sequencing technology and highly enriched somatic and germ cell populations were recently used in our laboratory for high-resolution expression profiling of rat testicular cells [13]. Thousands of novel TUTs and long noncoding transcripts (lncRNAs) were found. Although most of the TUTs share the genomic characteristics of lncRNAs, we investigated whether some of these transcriptions correspond to novel protein-coding genes. The primary aim of our study was therefore to characterize and validate at the protein level novel protein-coding genes expressed in meiotic and postmeiotic germ cells. We exploited our RNA-seq data set by combining it with a Shotgun proteome analysis of rat pSPC and rSPT in a PIT approach [30].

Until recently, proteomic studies could detect only proteins encoded by known genes because they relied on the information present in sequence databases. However, the possibility of using sample-specific databases derived from RNA-seq data is revolutionizing large-scale proteomics [34]. PIT approaches will allow better characterization of the protein pool present in a sample, increase protein identification rates [34], and improve the genome annotations (by identifying novel genes, exons, splicing events, translated UTRs, frame shifts, and reverse strands) [37].

Among 5379 TUTs and lncRNAs (4065 loci) reconstructed from pSPC and/or rSPT by RNA-seq analysis, we report evidence for the production of the corresponding protein and thus expression of 69 transcripts (48 TUTs and 21 lncRNAs) corresponding to 44 loci. About 72% of these MS-identified transcripts showed meiotic or postmeiotic expression patterns. Various traits were compared between MS-identified and non-MS-identified TUTs and lncRNAs, as well as MS-identified mRNAs. MS-identified transcripts showed genomics features differing from those typically shared by lncRNAs: they were longer, had longer ORFs, had better exon conservation, were closer to neighboring protein-coding genes, and had less cell-type specificity. This last feature is consistent with previous observations of a significantly more specific expression pattern of noncoding loci than protein-coding genes [13, 29]. These observations are consistent with the meiotic and postmeiotic TUTs and lncRNAs, for which there was mass spectrometry evidence of protein corresponding to novel protein-coding loci.

We generated an inferred protein sequence database of about 3 million sequences derived from the reconstructed transcripts translated in the six reading frames. These sequences were merged with canonical databases. Searching a higher-eukaryote, six-frame translation database is problematic because most of the search space will consist of translated

Downloaded from www.biolreprod.org.

noncoding sequences; indeed, for example, only 1%–2% of the human genome encodes proteins [68, 69]. The problem of such custom databases is the balance between increasing the exhaustivity of the available peptide sequences and decreasing the FDR. Both sequence redundancy due to size of the database and the presence of sequence errors increase the FDR [70–72]. Nevertheless, the validation of gene products at the peptide level demonstrates the robustness of our integrative strategy and clearly establishes its relevance for the improvement of rat genome annotation. However, the completeness of our PIT approach cannot be guaranteed, as some protein-coding transcripts may have been missed. Some may not have been assembled during the RNA-seq analysis because they are rare, or may have not been identified because their expression is below the stringent BEC defined to eliminate artifacts during transcript selection refinement. Several additional inherent drawbacks remain in the use of RNA-seq data sets for proteome analyses. Both sequencing errors and errors in assembly can result in artifacts, particularly shifts in the reading frame and apparent early termination of predicted protein sequences; such errors will cause erroneous deduced protein sequence and thus tryptic peptides.

Some peptides might not be identified because their sequence is not in the database; others might be missed because of the lower sensitivity of proteomics than transcriptomics and because all the peptides in an experiment are not equally detectable [73]. In their impressive work to improve mouse genome annotation using proteomics, Brosch and coworkers [74] found that up to 91% of all protein-coding exons and 86% of all introns could theoretically be confirmed from peptides identified by proteomics experiments. The number of potentially predicted peptides does not correlate with the number of identified peptides because the latter is directly related to the expression dynamics of the proteins in a given sample [74]. Gene expression is dependent on tissue, cell type, and environmental stage, and is often transient. Consequently, a novel gene corresponding to a large number of theoretical peptides might be missed because the encoded protein expressed only weakly or not at all in the sample studied. Indeed, the absence of identification of a peptide in mass spectrometry experiments used during a PIT strategy is not proof that the corresponding gene is not expressed (or protein produced).

PIT approaches are undoubtedly becoming a major component of the integrative genomics toolbox, not only for studies on nonmodel organisms [75–78], but also to improve the genome annotation of extensively studied model organisms such as rodents and human. A high value application of PIT methodologies is the deciphering of the spliceome of a cell population. In this work, we establish an appropriate basis for a large-scale study of the numerous germ cell-specific proteoforms whose synthesis is required for the completion of meiosis and the production of mature gametes. Our PIT approach will allow us to match any identified peptide to its corresponding exon on each transcript isoform, facilitating the discovery of novel splice junctions and their analysis by mass spectrometry [36]. Several RNA-seq-based gene expression studies have reported the discovery of thousands of novel transcript isoforms [79] in reproductive tissues: germ cells and testis [14, 15, 17, 80], placenta [81], and prostate tumor [82]. Chalmel and collaborators [13] reported 12 000 novel transcript isoforms expressed during rat spermatogenesis. Here, we identified up to 4999 proteins, and the proportion of potentially new protein isoforms identified in pSPC and rSPT accounted for about 80% of our novel protein identifications (Supplemental Data S1). This indicates that our PIT strategy is useful for identifying proteins and correlating changes in isoform expression during male germ cell differentiation. Nevertheless, experimental verification may be required to determine which isoforms are present in a given sample [83].

The final objective of our study was to demonstrate that a PIT strategy could identify novel proteins important to in germ cell biology. We thus investigated in greater detail two MS-identified transcripts, corresponding to VAMP9 and the T-ENOL (testicular enolase-like) protein. All validations performed at the transcript level unambiguously confirmed the expression of both genes in meiotic or postmeiotic germ cells in the rat. Both qPCR and in situ hybridization showed that the VAMP9 gene is expressed in pSPC and in rSPT. The expression of is highly variable (250-fold difference according to RNA-seq data), confirmed by much more intense staining in rSPT than in pSPC on in situ hybridization analysis. Further transcriptional validations showed the preferential expression of T-ENOL and VAMP9 transcripts in the testes in both mouse and human.

A polyclonal antibody against the $_{rec}$XLOC_01949 (T-ENOL) protein was produced and used to study the meiotic and postmeiotic distribution of the protein; this analysis confirmed the validity of its identification by MS in protein extracts from spermatocytes and spermatids. By inference, we were able to provide a novel annotation for its human homolog locus (LOC440356), which previously was ambiguously annotated as a ''noncoding CDIPT antisense RNA 1'' in NCBI or as a ''product of a dubious gene prediction'' in UniProt (accession: Q0VD67).

The identification of T-ENOL, an enolase domain-containing protein, further implicates enolases in spermiogenesis. It has long been known that mammalian sperm contains atypical forms of enolases [84]. These are associated with the fibrous sheath in the principal piece of the sperm flagellum. Eno-S, a human sperm-specific enolase, has been found to have three different isoforms whose expression is linked to the stage of sperm maturation through epididymal transit [85, 86]. Enolases are involved in glycolysis, so, as ATP production is crucial for the motility of spermatozoa, these enzymes presumably play a key role in spermiogenesis and male gamete biology. Recently, the spermatogenic cell-specific mouse enolase 4 (ENO4) was characterized. Using a gene trap approach, the authors demonstrated that disruption of the eno4 gene led to major sperm structural defects (a coiled flagellum and a disorganized fibrous sheath) and reduced sperm motility [87].

VAMP9 may also be very important for germ cell biology and spermatogenesis. It is a SNARE-like (soluble N-ethylmaleimide-sensitive factor attachment protein receptor) domain-containing protein, apparently a member of the Longin family, as is VAMP7, the closest paralog of VAMP9 [88]. This family of proteins is essential for regulating membrane trafficking. VAMP9 may also, like other SNARE proteins, regulate the SNARE complex formation involved in secretory and endocytic pathways [89]. This complex mediates diverse biochemical functions via a range of protein-protein interactions [90]. Some members of the SNARE family have already been associated with spermatogenesis: syntaxin 2 (STX2) may be involved in the acrosome reaction [91], and VAM6P and SNAP accumulate on the acrosome during capacitation [92]. Syntaxin 17 (STX17) is abundant in steroidogenic cells [93]. Furthermore, VAMP7 is important in several cell differentiation processes, including neurite outgrowth [94] and ciliogenesis [95]; note than sperm flagella and cilia share numerous features. It has also been suggested that VAMP7 defines a novel trafficking pathway to the cell surface in both neuronal

Downloaded from www.biolreprod.org.

and nonneuronal cells [96]. VAMP9 may possibly play similar roles in meiotic and postmeiotic germ cells.

In conclusion, we report the discovery of 44 new protein-coding loci expressed in rat male germ cells by using a PIT strategy. The relevance of this type of strategy for discovering novel testicular proteins was confirmed. In particular, we experimentally validated two of the novel male germline-associated proteins identified: a vesicle-associated membrane protein named VAMP9, and an enolase domain-containing protein, T-ENOL. The data contribute to a better understanding of germ cell differentiation events and represent a valuable resource for functional investigations into the role of numerous new genes and proteins in normal and pathological spermatogenesis. Graphical displays of both transcriptomics and proteomics datasets used in this study are available in open access through the ReproGenomics Viewer (http://rgv.genouest.org). Further progress will also be made available on this comprehensive viewer to help scientists in the field to improve their understanding of the molecular events underlying spermatogenesis.

Although PIT approaches are increasingly widely used, they require extensive advanced skills in all of transcriptomics, genomics, and proteomics. This currently limits their application to various fields of cell biology. Nevertheless, we foresee that over the next few years PIT approaches will contribute to the discovery of numerous novel proteins corresponding to currently unknown events or associated with transcripts thought to be untranslated.

ACKNOWLEDGMENT

We acknowledge Olivier Sallou and Olivier Collin (Genouest Bioinformatics Platform, IRISA) and Laetitia Cloarec (Inserm U1085, IRSET) for continued development, data upload, and maintenance of the ReproGenomics Viewer database. We thank Dominique Mahe Poiron, Nathalie Dejucq-Rainsford, and Nathalie Rioux-Leclerq for providing the human samples. We thank all members of the SEQanswers forums for helpful advice; Steven Salzberg and Cole Trapnell for continuous support with the Tuxedo suite; and Emmanuelle Com and the PRIDE team for the submission of the mass spectrometry proteomics data to ProteomeXchange via the PRIDE database. Sequencing was performed by the IGBMC Microarray and Sequencing platform, member of the France Génomique program.

REFERENCES

1. Matzuk MM, Lamb DJ. Genetic dissection of mammalian fertility pathways. Nat Cell Biol 2002; 4(suppl): s41–s49.
2. Eddy EM. Male germ cell gene expression. Recent Prog Horm Res 2002; 57:103–128.
3. Griswold MD. Interactions between germ cells and Sertoli cells in the testis. Biol Reprod 1995; 52:211–216.
4. Bettegowda A, Wilkinson MF. Transcription and post-transcriptional regulation of spermatogenesis. Philos Trans R Soc Lond B Biol Sci 2010; 365:1637–1651.
5. Jégou B, Pineau C, Dupaix A. Paracrine control of testis function. In: Wang C (ed.), Male Reproductive Function Endocrine Update Series. Berlin: Kluwer Academic; 1999:41–64.
6. Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SSW, Demougin P, Gattiker A, Moore J, Patard J-J, Wolgemuth DJ, Jégou B, Primig M. The conserved transcriptome in human and rodent male gametogenesis. Proc Natl Acad Sci U S A 2007; 104:8346–8351.
7. Chalmel F, Lardenois A, Evrard B, Mathieu R, Feig C, Demougin P, Gattiker A, Schulze W, Jégou B, Kirchhoff C, Primig M. Global human tissue profiling and protein network analysis reveals distinct levels of transcriptional germline-specificity and identifies target genes for male infertility. Hum Reprod 2012; 27:3233–3248.
8. Schlecht U, Demougin P, Koch R, Hermida L, Wiederkehr C, Descombes P, Pineau C, Jégou B, Primig M. Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. Mol Biol Cell 2004; 15:1031–1043.
9. Schultz N, Hamra FK, Garbers DL. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. Proc Natl Acad Sci U S A 2003; 100:12201–12206.
10. Shima JE, McLean DJ, McCarrey JR, Griswold MD. The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. Biol Reprod 2004; 71:319–330.
11. Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen Q-R, Cenacchi N, Khan J. Database of mRNA gene expression profiles of multiple human organs. Genome Res 2005; 15:443–450.
12. Wrobel G, Primig M. Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. Reproduction 2005; 129:1–7.
13. Chalmel F, Lardenois A, Evrard B, Rolland AD, Sallou O, Dumargne M-C, Coiffec I, Collin O, Primig M, Jégou B. High-resolution profiling of novel transcribed regions during rat spermatogenesis. Biol Reprod 2014; 91:5.
14. Laiho A, Kotaja N, Gyenesei A, Sironen A. Transcriptome profiling of the murine testis during the first wave of spermatogenesis. PLoS One 2013; 8: e61558.
15. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, Dym M, de Massy B, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep 2013; 3:2179–2190.
16. Gan H, Cai T, Lin X, Wu Y, Wang X, Yang F, Han C. Integrative proteomic and transcriptomic analyses reveal multiple post-transcriptional regulatory mechanisms of mouse spermatogenesis. Mol Cell Proteomics 2013; 12:1144–1157.
17. Margolin G, Khil PP, Kim J, Bellani MA, Camerini-Otero RD. Integrated transcriptome analysis of mouse spermatogenesis. BMC Genomics 2014; 15:39.
18. Meikar O, Vagin VV, Chalmel F, Søstar K, Lardenois A, Hammell M, Jin Y, Da Ros M, Wasik KA, Toppari J, Hannon GJ, Kotaja N. An atlas of chromatoid body components. RNA 2014; 20:483–495.
19. Djureinovic D, Fagerberg L, Hallström B, Danielsson A, Lindskog C, Uhlén M, Pontén F. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. Mol Hum Reprod 2014; 20: 476–488.
20. Schmid R, Grellscheid SN, Ehrmann I, Dalgliesh C, Danilenko M, Paronetto MP, Pedrotti S, Grellscheid D, Dixon RJ, Sette C, Eperon IC, Elliott DJ. The splicing landscape is globally reprogrammed during male meiosis. Nucleic Acids Res 2013; 41:10170–10184.
21. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) project. Science 2004; 306:636–640.
22. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012; 22: 1775–1789.
23. Hung T, Chang HY. Long noncoding RNA in genome regulation: prospects and mechanisms. RNA Biol 2010; 7:582–585.
24. Bánfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE Jr, Kundaje A, Gunawardena HP, Yu Y, Xie L, Krajewski K, Strahl BD, et al. Long noncoding RNAs are rarely translated in two human cell lines. Genome Res 2012; 22:1646–1657.
25. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011; 25: 1915–1927.
26. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009; 458:223–227.
27. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 2010; 28:503–510.
28. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L, Yang X, Amit I, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature 2011; 477:295–300.
29. Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL, Regev A, Schier AF. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res 2012; 22:577–591.
30. Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods 2012; 9:1207–1211.
31. Brewis IA, Brennan P. Proteomics technologies for the global identifica-

Downloaded from www.biolreprod.org.

tion and quantification of proteins. Adv Protein Chem Struct Biol 2010; 80:1–44.
32. Lamond AI, Uhlen M, Horning S, Makarov A, Robinson CV, Serrano L, Hartl FU, Baumeister W, Werenskiold AK, Andersen JS, Vorm O, Linial M, et al. Advancing cell biology through proteomics in space and time (PROSPECTS). Mol Cell Proteomics 2012; 11:O112.017731.
33. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 2014; 42:D191–D198.
34. Wang X, Slebos RJC, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B. Protein identification using customized protein sequence databases derived from RNA-Seq data. J Proteome Res 2012; 11:1009–1017.
35. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10:57–63.
36. Sheynkman GM, Shortreed MR, Frey BL, Smith LM. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Mol Cell Proteomics 2013; 12:2341–2353.
37. Woo S, Cha SW, Merrihew G, He Y, Castellana N, Guest C, MacCoss M, Bafna V. Proteogenomic database construction driven from large scale RNA-seq data. J Proteome Res 2014; 13:21–28.
38. Pineau C, Syed V, Bardin CW, Jégou B, Cheng CY. Germ cell-conditioned medium contains multiple factors that modulate the secretion of testins, clusterin, and transferrin by Sertoli cells. J Androl 1993; 14: 87–98.
39. Com E, Evrard B, Roepstorff P, Aubry F, Pineau C. New insights into the rat spermatogonial proteome. Mol Cell Proteomics 2003; 2:248–261.
40. Skinner MK, Fritz IB. Structural characterization of proteoglycans produced by testicular peritubular cells and Sertoli cells. J Biol Chem 1985; 260:11874–11883.
41. Toebosch AM, Robertson DM, Klaij IA, de Jong FH, Grootegoed JA. Effects of FSH and testosterone on highly purified rat Sertoli cells: inhibin alpha-subunit mRNA expression and inhibin secretion are enhanced by FSH but not by testosterone. J Endocrinol 1989; 122:757–762.
42. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, et al. Ensembl 2013. Nucleic Acids Res 2013; 41:D48–D55.
43. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012; 40:D130–D135.
44. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 2006; 7(Suppl 1): S12.1–14.
45. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, et al. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res 2013; 41:D64–D69.
46. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 2012; 7:562–578.
47. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25:1105–1111.
48. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 2010; 28:511–515.
49. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000; 16:276–277.
50. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Girón CG, et al. Ensembl 2014. Nucleic Acids Res 2014; 42:D749–D755.
51. Lavigne R, Becker E, Liu Y, Evrard B, Lardenois A, Primig M, Pineau C. Direct iterative protein profiling (DIPP)—an innovative method for large-scale protein detection applied to budding yeast mitosis. Mol Cell Proteomics 2012; 11:M111.012682.
52. Chalmel F, Primig M. The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. BMC Bioinformatics 2008; 9:86.
53. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3:Article3.
54. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403–410.
55. Kulikova T, Aldebert P, Althorpe N, Baker W, Bates K, Browne P, van den Broek A, Cochrane G, Duggan K, Eberhardt R, Faruque N, Garcia-Pastor M, et al. The EMBL Nucleotide Sequence Database. Nucleic Acids Res 2004; 32:D27–D30.
56. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013; 30:772–780.
57. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009; 25:1189–1191.
58. Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic Acids Res 2004; 32:W327–W331.
59. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011; 39:D225–D229.
60. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res 2008; 36:W197–W201.
61. Calvel P, Kervarrec C, Lavigne R, Vallet-Erdtmann V, Guerrois M, Rolland AD, Chalmel F, Jégou B, Pineau C. CLPH, a novel casein kinase 2-phosphorylated disordered protein, is specifically associated with postmeiotic germ cells in rat spermatogenesis. J Proteome Res 2009; 8: 2953–2965.
62. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002; 12: 656–664.
63. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Soldà G, Simons C, Sunkin SM, Crowe ML, et al. Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res 2008; 18:1433–1445.
64. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A 2008; 105:716–721.
65. Ponjavic J, Oliver PL, Lunter G, Ponting CP. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. PLoS Genet 2009; 5:e1000617.
66. Schug J, Schuller W-P, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ. Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol 2005; 6:R33.
67. Palmer MR, McDowall MH, Stewart L, Ouaddi A, MacCoss MJ, Swanson WJ. Mass spectrometry and next-generation sequencing reveal an abundant and rapidly evolving abalone sperm protein. Mol Reprod Dev 2013; 80:460–465.
68. Claverie J-M. Fewer genes, more noncoding RNA. Science 2005; 309: 1529–1530.
69. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007; 447:799–816.
70. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J Proteomics 2010; 73:2092–2123.
71. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat Methods 2007; 4:787–797.
72. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol 2006; 7:R35.
73. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R. Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 2007; 25:125–131.
74. Brosch M, Saunders GI, Frankish A, Collins MO, Yu L, Wright J, Verstraten R, Adams DJ, Harrow J, Choudhary JS, Hubbard T. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and ''resurrected'' pseudogenes in the mouse genome. Genome Res 2011; 21:756–767.
75. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, Dodt M, Mackowiak SD, Gogol-Doering A, Oenal P, Rybak A, Ross E, et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Res 2011; 21: 1193–1200.
76. Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. Non-model organisms, a species endangered by proteogenomics. J Proteomics 2014; 105:5–18.
77. Looso M, Preussner J, Sousounis K, Bruckskotten M, Michel CS, Lignelli E, Reinhardt R, Höffner S, Krüger M, Tsonis PA, Borchardt T, Braun T. A de novo assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. Genome Biol 2013; 14:R16.
78. Wu H-X, Jia H-M, Ma X-W, Wang S-B, Yao Q-S, Xu W-T, Zhou Y-G,

Downloaded from www.biolreprod.org.

Gao Z-S, Zhan R-L. Transcriptome and proteomic analysis of mango (*Mangifera indica* Linn) fruits. J Proteomics 2014; 105:19–30.
79. Tress ML, Bodenmiller B, Aebersold R, Valencia A. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. Genome Biol 2008; 9:R162.
80. Gan Q, Chepelev I, Wei G, Tarayrah L, Cui K, Zhao K, Chen X. Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. Cell Res 2010; 20:763–783.
81. Kim J, Zhao K, Jiang P, Lu Z, Wang J, Murray JC, Xing Y. Transcriptome landscape of the human placenta. BMC Genomics 2012; 13:115.
82. Srinivasan S, Patil AH, Verma M, Bingham JL, Srivatsan R. Genome-wide profiling of RNA splicing in prostate tumor from RNA-seq data using virtual microarrays. J Clin Bioinforma 2012; 2:21.
83. Wu P, Zhang H, Lin W, Hao Y, Ren L, Zhang C, Li N, Wei H, Jiang Y, He F. Discovery of novel genes and gene isoforms by integrating transcriptomic and proteomic profiling from mouse liver. J Proteome Res 2014; 13:2409–2419.
84. Edwards YH, Grootegoed JA. A sperm-specific enolase. J Reprod Fertil 1983; 68:305–310.
85. Force A, Viallard J-L, Grizard G, Boucher D. Enolase isoforms activities in spermatozoa from men with normospermia and abnormospermia. J Androl 2002; 23:202–210.
86. Force A, Viallard J-L, Saez F, Grizard G, Boucher D. Electrophoretic characterization of the human sperm-specific enolase at different stages of maturation. J Androl 2004; 25:824–829.
87. Nakamura N, Dai Q, Williams J, Goulding EH, Willis WD, Brown PR, Eddy EM. Disruption of a spermatogenic cell-specific mouse enolase 4 (eno4) gene causes sperm structural defects and male infertility. Biol Reprod 2013; 88(4):90: 1–12.
88. Filippini F, Rossi V, Galli T, Budillon A, D'Urso M, D'Esposito M. Longins: a new evolutionary conserved VAMP family sharing a novel SNARE domain. Trends Biochem Sci 2001; 26:407–409.
89. Chaineau M, Danglot L, Galli T. Multiple roles of the vesicular-SNARE TI-VAMP in post-Golgi and endosomal trafficking. FEBS Lett 2009; 583: 3817–3826.
90. Rossi V, Banfield DK, Vacca M, Dietrich LEP, Ungermann C, D'Esposito M, Galli T, Filippini F. Longins and their longin domains: regulated SNAREs and multifunctional SNARE regulators. Trends Biochem Sci 2004; 29:682–688.
91. Hutt DM, Baltz JM, Ngsee JK. Synaptotagmin VI and VIII and syntaxin 2 are essential for the mouse sperm acrosome reaction. J Biol Chem 2005; 280:20197–20203.
92. Brahmaraju M, Shoeb M, Laloraya M, Kumar PG. Spatio-temporal organization of Vam6P and SNAP on mouse spermatozoa and their involvement in sperm-zona pellucida interactions. Biochem Biophys Res Commun 2004; 318:148–155.
93. Katafuchi K, Mori T, Toshimori K, Iida H. Localization of a syntaxin isoform, syntaxin 2, to the acrosomal region of rodent spermatozoa. Mol Reprod Dev 2000; 57:375–383.
94. Sato M, Yoshimura S, Hirai R, Goto A, Kunii M, Atik N, Sato T, Sato K, Harada R, Shimada J, Hatabu T, Yorifuji H, et al. The role of VAMP7/TI-VAMP in cell polarity and lysosomal exocytosis in vivo. Traffic 2011; 12: 1383–1393.
95. Szalinski CM, Labilloy A, Bruns JR, Weisz OA. VAMP7 modulates ciliary biogenesis in kidney cells. PLoS One 2014; 9:e86425.
96. Flowerdew SE, Burgoyne RD. A VAMP7/Vti1a SNARE complex distinguishes a non-conventional traffic route to the cell surface used by KChIP1 and Kv4 potassium channels. Biochem J 2009; 418:529–540.
97. Leblond CP, Clermont Y. Definition of the stages of the cycle of the seminiferous epithelium in the rat. Ann N Y Acad Sci 1952; 55:548–573.
98. Dym M, Clermont Y. Role of spermatogonia in the repair of the seminiferous epithelium following x-irradiation of the rat testis. Am J Anat 1970; 128:265–282.

Downloaded from www.biolreprod.org.

Chapitre 2

Découverte de nouvelles isoformes spécifiques des cellules germinales chez le rat

I. Introduction

Plusieurs études montrent qu'une approche de type PIT permet de mettre en évidence l'expression de nouvelles isoformes de protéines connues, exprimées spécifiquement dans un tissu ou dans un type cellulaire donné (Sheynkman et al., 2013; Trapnell et al., 2012). En effet, les séquences directement traduites des transcrits alternatifs présents dans l'échantillon biologique composent la base de séquences inférées des transcrits nouvellement assemblés, et les peptides théoriques qui en résultent servent à l'identification des peptides expérimentaux. Or, les peptides spécifiques d'une nouvelle isoforme, qu'ils soient jonctionnels où codés par un nouvel exon, ne sont pas présents dans les bases de séquences canoniques. Cette possibilité offerte par l'approche PIT est d'un grand intérêt pour l'étude d'un processus aussi complexe que la spermatogenèse au cours duquel s'expriment de nombreuses isoformes de gènes dans les cellules germinales (Laiho et al., 2013; Margolin et al., 2014). Pour la grande majorité des gènes, de multiples isoformes d'ARNms sont produits *via* de multiples voies d'épissage alternatif, ou bien du fait de l'utilisation de différents promoteurs ou de l'existence de sites de terminaison différents (Kwan et al., 2008; Pan et al., 2008). Comme l'utilisation de plusieurs sites d'épissage alternatif, de différents promoteurs et sites de polyadénylation, est régulée en réponse à des signaux extracellulaires ou au cours du développement, les transcrits alternatifs peuvent déterminer les fonctions d'un gène dans différentes circonstances.

Des analyses utilisant des microarrays exon-spécifiques ont détecté plus d'évènements d'épissage alternatif dans le testicule que dans de nombreux autres tissus pendant la spermatogenèse à l'exception du cerveau (Kan et al., 2005). Cependant, une telle régulation de l'épissage, qui plus est stade spécifique, n'est pas connue. La régulation de l'épissage alternatif est cependant liée à la méiose, car contrairement aux cellules en mitose dans lesquelles la transcription est éteinte, les cellules méiotiques sont transcriptionnellement très actives (Monesi, 1964). L'épissage alternatif qui a été détecté dans le testicule humain par puces à ADN exons-spécifiques montre que celui-ci peut conduire à l'introduction prématurée de codons stop, ceci se produisant individuellement à de basses fréquences (Kan et al., 2005). Ces caractéristiques donnent à penser que certains évènements d'épissage alternatif peuvent représenter du bruit de fond résultant d'une faible rigueur du contrôle de l'épissage dans les testicules (Melamud et Moult, 2009). Cependant, il existe des variations dans les niveaux d'expression d'un certain nombre de régulateurs importants d'épissage d'ARN pendant la spermatogenèse, comprenant entre autres les protéines RNP nucléaires

hétérogènes, les hnRNPs (Elliott et al., 2000), les protéines activatrices d'épissage « SR-like » (Tra2β) (Grellscheid et al., 2011), les protéines STAR : Sam 68 (Paronetto et al., 2006) et T-STAR (Venables et al., 2004). La protéine nPTB qui est aussi un régulateur d'épissage pourrait remplacer la fonction de la protéine PTBP1 (Polypyrimydine Tract Binding Protein) pour assurer une régulation correcte des événements d'épissage de certains exons cibles pendant la méiose (Schmid et al., 2013). Les évènements d'épissage spécifiques apparaissant au cours de la méiose donnent lieu à l'expression d'isoformes de protéines connues spécifiques de ces stades. Chez la souris, il a été montré que des basculements de l'expression d'isoformes d'ARNms ont lieu entre les transcriptomes aux stades pré-méiotique et méiotique, et des évènements d'épissage alternatif ont lieu dans les cellules germinales entre les stades spermatogonies et spermatocytes. Par ailleurs, les transcrits alternatifs exprimés pendant la méiose sont démontrés être spécifiques du testicule (Schmid et al., 2013). Les profils d'expression de gènes dans les cellules germinales mâles au cours de la première vague de la spermatogenèse ont été évalués par des techniques de séquençage à haut débit chez la souris, et ont ainsi permis l'étude des isoformes pendant la différenciation des cellules germinales mâles (Laiho et al., 2013; Margolin et al., 2014). Margolin et collaborateurs indiquent d'ailleurs par leur analyse du transcriptome de la spermatogenèse grâce au RNA-seq, qu'environ 1.000 nouveaux gènes sont exprimés pendant la méiose et que 5.000 isoformes sont potentiellement codantes et exprimées pendant la spermatogenèse chez la souris (Margolin et al., 2014). Laiho et collaborateurs trouvent 2.494 gènes différentiellement exprimés au cours de la spermatogenèse chez la souris, et 160.000 isoformes de transcrits parmi lesquels 29% sont de nouveaux transcrits de gènes connus (Laiho et al., 2013). Cette dernière étude a permis de mettre en évidence un grand nombre de transcrits spécifiques de chaque stade, ainsi que les processus biologiques relatifs à la spermatogenèse et les composants cellulaires associés à chaque stade (termes de la Gene Ontology).

La régulation des évènements d'épissage alternatif change pendant la méiose. Une forte corrélation entre les niveaux d'épissage alternatif entre l'homme et la souris suggère la conservation des mécanismes de régulation fonctionnelle et un rôle pour des événements d'épissage alternatif conservés. En outre, l'épissage alternatif conservé s'applique aux gènes exprimés dans les tissus neuronaux et les voies de signalisation, tandis qu'un épissage alternatif plus divergent opère davantage pour les gènes exprimés dans les testicules et les cellules de lignées cancéreuses où une augmentation du taux de l'épissage aberrant peut

résulter de conditions cellulaires anormales, de la prolifération rapide des cellules ou des mécanismes de surveillance défectueux (Kan et al., 2005). Il faudra donc garder en mémoire pour une analyse des nouvelles isoformes exprimées dans le testicule, en l'occurrence dans les cellules germinales, qu'un épissage alternatif « non fonctionnel » peut se produire.

II. Résultats et discussion

L'approche PIT décrite dans le précédant chapitre a permis d'identifier au niveau protéique un certain nombre de nouvelles protéines dans les cellules germinales méiotiques et post méiotiques chez le rat. Ces nouvelles identifications représentent 35% des identifications totales sur la banque personnalisée « Rat non redundant reference proteome » dans les spermatocytes pachytène ainsi que dans les spermatides rondes. Parmi ces identifications, 27% dans les pSPC et 28% dans les rSPT (Figure 24) correspondent à des transcrits classifiés comme de nouvelles isoformes potentielles dans l'étude de Chalmel et collaborateurs (Chalmel et al., 2014). Cette proportion correspond à environ 80% de nouvelles isoformes potentielles sur le nombre d'identifications nouvelles. A titre de comparaison, 7 à 8% des identifications dans les rSPT et les pSPC respectivement correspondent à de nouveaux évènements codants (Figure 24). Ces derniers ont fait l'objet du chapitre précédant.

A. Nouvelles isoformes potentielles spécifiques des cellules méiotiques et post méiotiques

Etudier ces identifications classées « nouvelles isoformes potentielles » nous permet d'identifier de nouvelles isoformes de protéines connues qui sont spécifiquement exprimées aux stades méiotique et post-méiotique de la spermatogenèse.

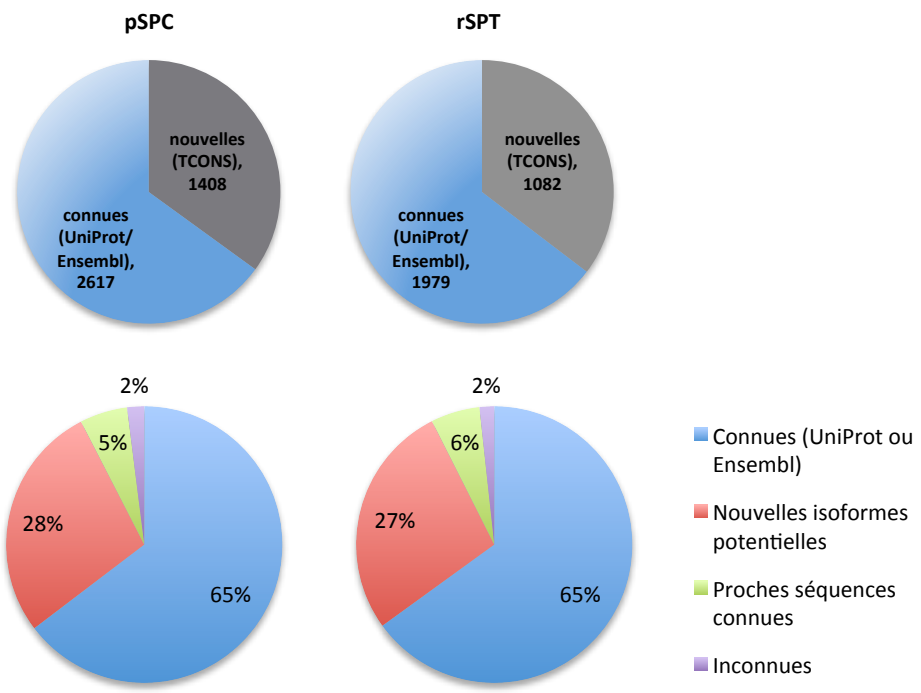

Figure 24. Proportion des différentes catégories de transcrits desquels dérivent les protéines identifiées par PIT dans les spermatocytes et les spermatides

Répartition du nombre d'identifications de protéines obtenues dans les pSPC et les rSPT par LC-MS/MS sur la banque « Rat non redundant reference proteome » comprenant les séquences traduites des transcrits reconstruits après RNA-seq (identifiants TCONS) dans les différents types cellulaires testiculaires isolés (Chalmel et al., 2014) ainsi que les séquences d'UniProt et d'Ensembl (Cf. article présenté). En haut, répartition entre les identifications de protéines connues (en bleu), et les protéines nouvelles (en gris). En bas, répartition détaillée des nouvelles identifications (TCONS) correspondant à des transcrits connus, ou à de nouvelles isoformes potentielles, ou bien à des transcrits inconnus : intergéniques ou introniques. Une autre catégorie représente les identifications de séquences proches de séquences déjà connues.

Le fait d'avoir accès à l'information de l'expression des nouvelles isoformes potentielles au niveau du transcrit grâce aux données transcriptomiques quantifiées de RNA-seq est essentiel si l'on veut s'intéresser aux isoformes susceptibles d'avoir un rôle dans la spermatogenèse (ici, aux stades méiotiques et post méiotiques). En effet, ces informations permettent de bien discriminer si ces évènements codants détectés par MS font partie du bruit d'épissage ou bien de l'épissage fonctionnel. On sait que chez l'homme 1 à 10 % d'évènements d'épissage selon le nombre d'introns et le niveau d'expression des gènes peuvent constituer des erreurs et font partie de ce bruit de fond d'épissage (Melamud et Moult, 2009). Sachant que les produits issus de ce phénomène sont non fonctionnels, il convient dans une telle recherche d'appliquer aussi des filtres de sélection selon le profil d'expression différentiel des transcrits dans les cellules germinales mâles étudiées. De plus, l'épissage fonctionnel va de paire avec un certain niveau de conservation et d'expression des transcrits. Ces deux critères doivent donc

aussi êtres pris en compte. Dans notre approche, les critères de sélection des nouvelles isoformes potentielles identifiées par MS permettant de les distinguer du bruit de fond d'épissage peuvent donc être les suivants:

- un niveau d'expression suffisant de leur transcrit dans les cellules étudiées (pSPC et rSPT);
- un profil d'expression différentiel des transcrits à ces stades de différenciation;
- la conservation de leur transcrit parmi les vertébrés.

Il est possible de réaliser ces filtrations sur l'ensemble des protéines nouvelles identifiées dans les spermatocytes pachytène et les spermatides rondes en utilisant le même système de filtration que pour les TUTs et les lncRNAs (voir Chapitre 1). En sélectionnant les nouvelles protéines dont le transcrit est classé dans la catégorie « nouvelle isoforme potentielle », dont le transcrit est plus long que 200b pour éliminer les artéfacts; exprimé dans les pSPC ou dans les rSPT et avec un profil d'expression différentiel dans les pSPC ou les rSPT; on obtient 1.350 protéines sélectionnées sans appliquer de filtre sur la conservation. Si on sélectionne les 10 protéines provenant des séquences les plus conservées (Score Phastcons), gage d'une fonctionnalité de ces protéines (Kan et al., 2005), on peut vérifier que la protéine connue correspondant à ces isoformes est toujours impliquée dans un processus lié à la spermatogenèse ou au cycle cellulaire, ou bien à l'épissage et la régulation des ARNs comme en attestent les annotations de la Gene Ontology (Tableau 6). Par exemple, la protéine codée par Usp34 est impliquée dans la voie de signalisation Wnt, qui est d'une importance capitale pour le bon déroulement de la spermatogenèse (Kerr et al., 2014), ce qui suggère que son isoforme potentielle TCONS_00036476 y est également impliquée.

Nouvelle isoforme	UniProtKB	Gène	Identifiant GO	Terme GO
TCONS_00098492	D4A3E1	Hnrnpll	GO:0006397	mRNA processing
			GO:0033120	positive regulation of RNA splicing
TCONS_00006794	Q5BJN3	Tial1	GO:0007281	germ cell development
TCONS_00012686			GO:0008284	positive regulation of cell proliferation
			GO:0017145	stem cell division
			GO:0017091	AU-rich element binding
TCONS_00035484	A4L9P7	Pds5a	GO:0007049	cell cycle
			GO:0007067	mitotic nuclear division
			GO:0008156	negative regulation of DNA replication
			GO:0051301	cell division
TCONS_00036476	F1M791	Usp34	GO:0090263	positive regulation of canonical Wnt signaling pathway
TCONS_00078575	Q99MI7	Uba3	GO:0007049	cell cycle
			GO:0045892	negative regulation of transcription, DNA-templated
TCONS_00099743	Q6AYU5	Pcbp2	GO:0003723	RNA binding
TCONS_00106235			GO:0050687	negative regulation of defense response to virus
TCONS_00106236				
TCONS_00117185	Q6P7P5	Bzw1	GO:0006355	regulation of transcription, DNA-templated

Tableau 6. Termes de la Gene Ontology associés à 10 isoformes potentielles sélectionnées

Termes de la Gene Ontology associés aux protéines connues relatives aux 10 isoformes potentielles les plus conservées dans les pSPC et les rSPT, qui sont en relation avec le développement des cellules germinales ainsi qu'avec la régulation de la transcription, l'épissage des ARNs ou la traduction des ARNms.

On peut vérifier qu'il s'agit bien de nouvelles isoformes dans la mesure où au moins un peptide permettant de les identifier est spécifique de cette isoforme, et ne s'aligne sur aucune protéine connue chez le rat. Par exemple, la portion de séquence contenant ce peptide spécifique sur la nouvelle isoforme TCONS_00006795 de Tial1 s'aligne sur la protéine Tial1 chez d'autres espèces (F7ELF9_MACMU: macaque, et Q921W2_MOUSE : souris), mais pas chez le rat. Cette isoforme est donc nouvelle chez le rat.

Etant donné leur correspondance aux critères énoncés plus haut, sans même prendre en compte leur conservation, on peut considérer que les 1.350 nouvelles isoformes identifiées dans les cellules méiotiques et post-méiotiques ont un intérêt dans le processus spermatogénétique, au moins aux étapes étudiées, car leur transcrit a une expression différentielle à l'un de ces stades. Mais, parmi ces nouvelles isoformes potentielles, nous

allons voir que si certains peuvent être validés par des peptides protéotypiques, il y a parmi eux un certain nombre de faux positifs dus à des biais dans l'identification de ces protéines.

B. Détection de nouveaux évènements d'épissage et d'UTR codants

Pour faciliter la mise en évidence de nouvelles isoformes potentielles identifiées avec des peptides spécifiques correspondant à de nouveaux exons, des UTRs ou à une nouvelle jonction d'épissage, les protéines identifiées avec un peptide non partagé par d'autres protéines sont considérées. Parmi les peptides identifiés par MS/MS qui ont permis de caractériser de potentielles nouvelles isoformes, nous identifiions des peptides jonctionnels entre deux exons qui attestent de nouveaux événements d'épissage. Donnons l'exemple de TCONS_00040250 (XLOC_015308), identifiée dans les pSPC, qui est une isoforme potentielle des protéines connues : D4A3U3, D47A4 ou D4A513 correspondant à ENSRNOG00000007046 (gène prédit à la localisation XLOC_015308). Le peptide identifié par MS/MS ayant pour séquence : SEGQTGPNTALSSLDEFLEESSK ne s'aligne sur aucune des protéines prédites répertoriées correspondant à cette localisation, mais atteste d'un nouvel événement d'épissage alternatif avec un saut d'exon dont témoigne la séquence traduite du transcrit TCONS_00040250 (Figure 25A). La nouvelle isoforme TCONS_00039708 identifiée grâce à un peptide jonctionnel résulte également d'un saut d'exon par rapport à la protéine connue Ubiquitin carboxyl-terminal hydrolaseisozyme L3 (Figure 25B).

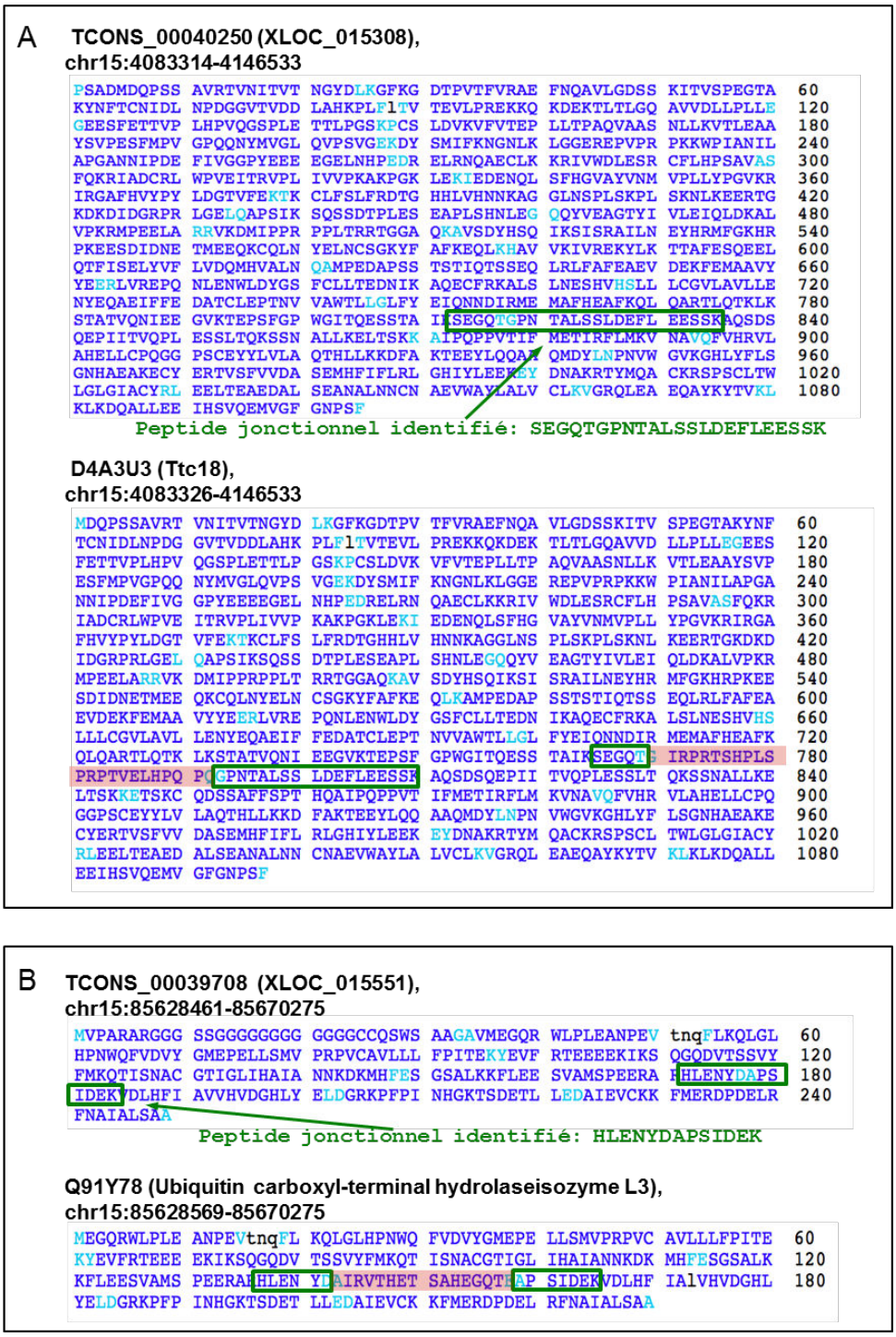

Figure 25. Mise en évidence de nouvelles isoformes avec un saut d'exon, grâce aux peptides jonctionnels identifiés par l'approche PIT

Les résultats d'un Blat UCSC des séquences indiquées sur le génome nr4 du rat sont données pour deux exemples : en **A**, TCONS_00040250 en comparaison avec la protéine prédite connue Ttc18, en **B,** TCONS_00039708, en comparaison avec la protéine connue Ubiquitin carboxyl-terminal hydrolaseisozyme L3. Les limites des exons sont matérialisées par les résidus d'acides aminés colorés en bleu clair. Les exons qui sont éliminés de la protéine connue à la nouvelle isoforme sont surlignés en rouge, et les résidus d'acides aminés correspondant aux peptides identifiés par MS/MS sont encadrés en vert.

Les peptides jonctionnels identifiés permettent aussi de mettre en évidence le remplacement ou l'insertion d'exons. Par exemple, le peptide jonctionnel identifié qui chevauche trois exons sur la nouvelle isoforme TCONS_00008859, et seulement deux exons (le second et le troisième) sur la Calpaine small subunit1, montre le remplacement du premier exon de la Calpain small subunit 1 par un autre exon plus rapproché des autres dans la nouvelle isoforme présentée (Figure 26A). Un peptide jonctionnel identifié montre l'insertion d'un exon traduit supplémentaire au sein de la nouvelle isoforme TCONS_00117809 de la protéine Vapa (Figure 26B). Une insertion d'exon est aussi montrée sur l'isoforme TCONS_00107908 de la protéine connue Rexo 2, rallongeant celle-ci en C-ter (Figure 26C).

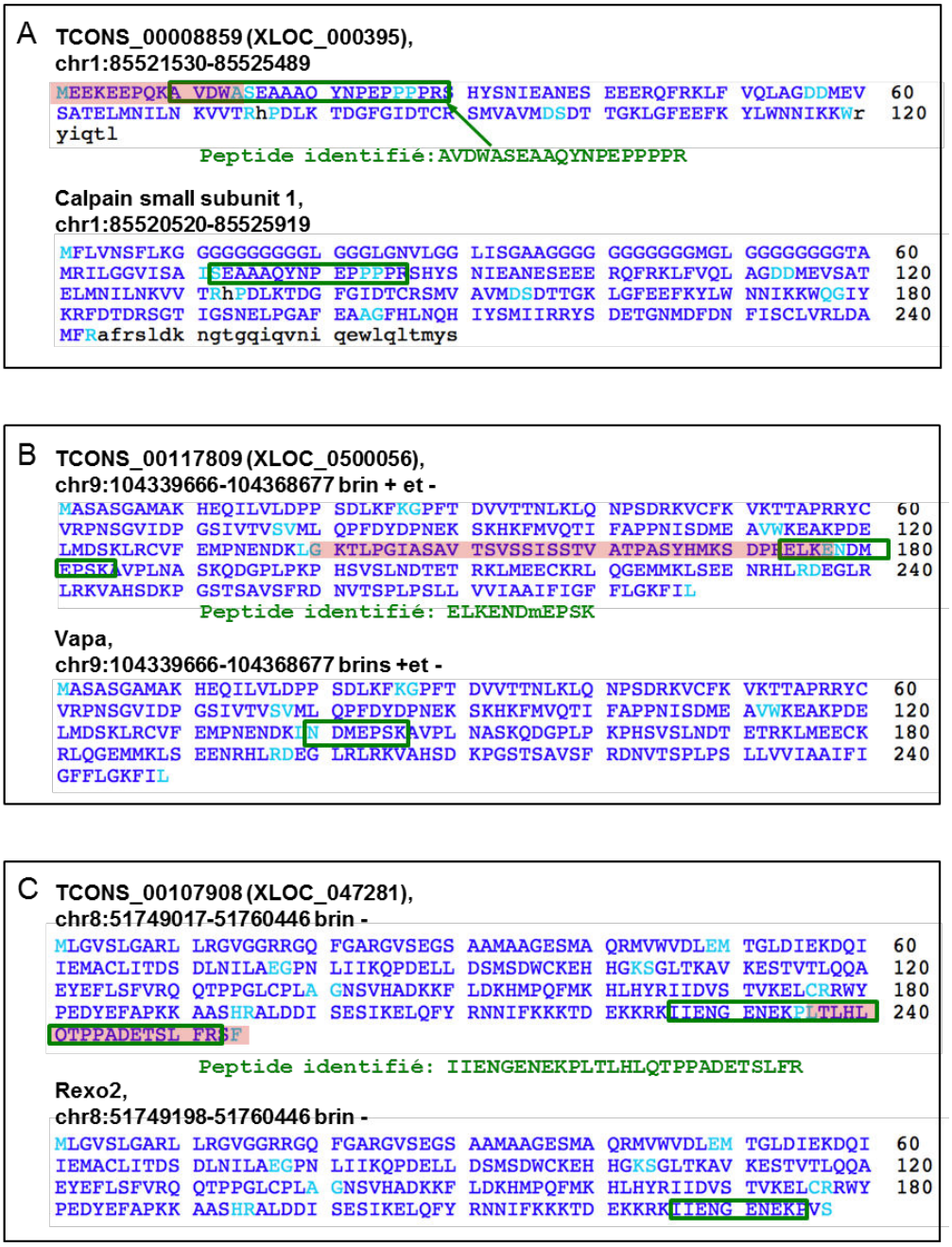

Figure 26. Mise en évidence de nouveaux évènements d'épissage alternatif grâce aux peptides jonctionnels identifiés par l'approche PIT

(Blat UCSC des séquences indiquées, sur le génome nr4 du rat). **A,** Remplacement d'exon sur le transcrit TCONS_00008859 générant une isoforme avec un N-ter différent de celui de la protéine Calpain small subunit 1. **B,** Insertion d'un exon au milieu du transcrit, ajoutant une portion à la nouvelle isoforme TCONS_00117809 par rapport à la protéine connue Vapa. **C,** Insertion d'un exon en 5' rallongeant la séquence en C-ter de la nouvelle isoforme TCONS_00107908 par rapport à la protéine connue Rexo 2. Les limites des exons sont matérialisées par les résidus d'acides aminés colorés en bleu clair. Les exons qui sont insérés sont surlignés en rouge, et les résidus d'acides aminés correspondant aux peptides identifiés par MS sont encadrés en vert.

En l'occurrence, le nouvel exon traduit de cette nouvelle isoforme TCONS_00107908 est ajouté en 5' du transcrit étant donné que cette protéine est codée par le brin inverse. Une vue du peptide IIENGENEKPLTLHLQTPPADETSLFR (Figure 26C) qui a permis d'identifier TCONS_00107908 dans les spermatides rondes, aligné sur le génome du rat, permet de vérifier que la nouvelle région codante de TCONS_00107908 se trouve bien sur le 5'UTR du gène connu Rexo2 (Figure 27).

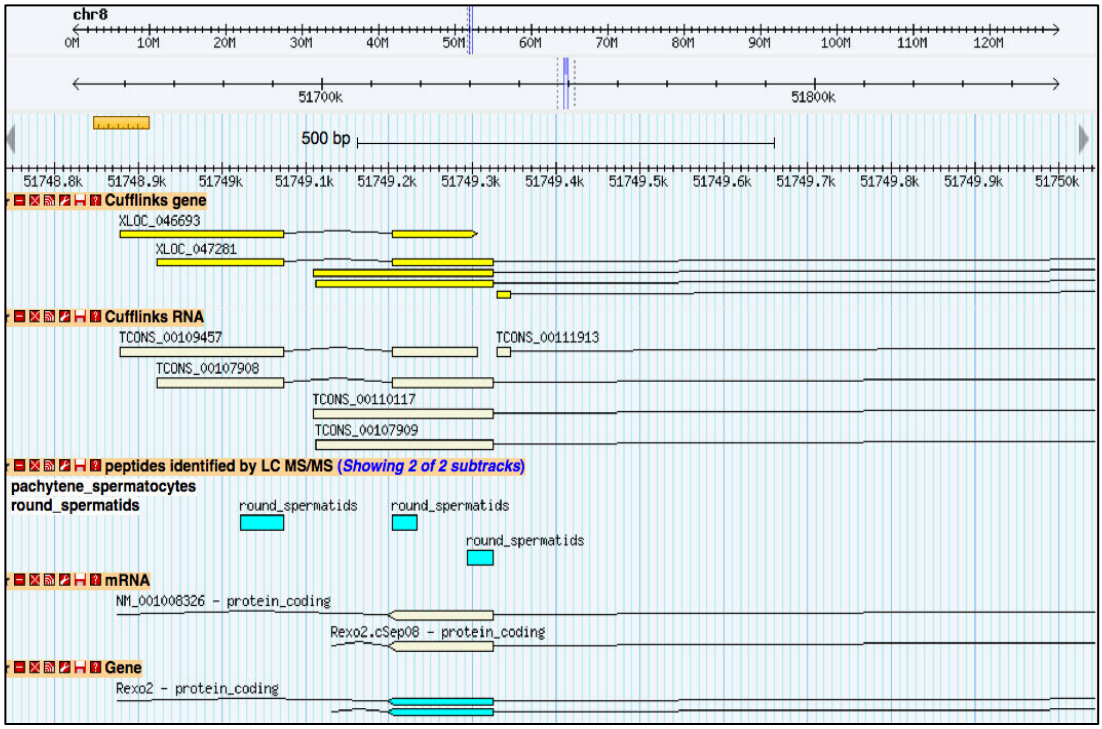

Figure 27. Localisation d'un peptide identifié correspondant à un UTR : transcrit TCONS_00107908 sur le génome rn4 du rat

Copie d'écran du navigateur RGV à la position chr8:51,748,822..51,750.000: le peptide le plus à gauche (séquence IIENGENEKPLTLHLQTPPADETSLFR), identifié dans les spermatides rondes (Boîte bleue), s'aligne sur une région non codante (5'UTR) du gène Rexo2 dont les exons sont représentés par les boîtes beiges (mRNA) et bleues (gène).

C. Problèmes rencontrés lors de la détection de nouvelles isoformes potentielles

Un certain nombre de protéines identifiées et qui sont classées parmi les isoformes potentielles sont fausses. En effet, de nombreuses séquences déduites des transcrits assemblés à partir du RNA-seq présentes dans la base de séquences peuvent être artéfactuelles, erronées ou écourtées. Même si une séquence de nouvelle isoforme potentielle est correcte, il se peut qu'elle soit identifiée de manière ambiguë par spectrométrie de masse, c'est à dire uniquement par des peptides partagés. Dans tous les cas, si les séquences des nouvelles isoformes potentielles sont plus courtes que celles des protéines connues, et si elles sont identifiées seulement avec des peptides partagés par ces deux séquences, il est impossible de savoir laquelle est réellement présente dans l'échantillon. En effet, la mieux identifiée sera la plus courte si un regroupement de protéines est appliqué à l'analyse, parce qu'elle obtient une meilleure couverture de séquence par les peptides identifiés. Plusieurs exemples pour lesquels une identification d'isoforme potentielle est obtenue de manière ambiguë peuvent se présenter (Figures 28-30).

a) Séquence plus courte : isoforme potentielle d'Ulk4

Il arrive que tous les peptides permettant d'identifier une nouvelle isoforme correspondent aussi à la (ou les) protéine(s) répertoriée(s), et que l'information d'un peptide spécifique de cette nouvelle isoforme ne soit pas accessible. Dans ce cas, même si la séquence de la nouvelle isoforme est correcte, sa présence est ambiguë. C'est l'exemple de la protéine TCONS_00112253, une nouvelle isoforme potentielle identifiée avec des peptides partagés avec Ulk4 (Figure 28). Ayant une séquence plus longue que la séquence dérivée du transcrit TCONS_00112253, la protéine Ulk4 n'est pas identifiée, au profit de la nouvelle isoforme potentielle qui peut, par ailleurs, être incorrecte.

```
TCONS_00112253_j_5_1 (nouvelle isoforme potentielle), (1198
a.a.)
3'-5' frame3
RGGSLETVIAQDQNLPEDVVREFGVDLVTGLHHLHRLGILFCDLSPGKILLEGPGTLKFS
NFCLAKVEGESLEEFFALVAAEEGGGDSGENTLRKSMKTRVRGSLIYAAPEIVTGTEFSV
TSDLWSLGCLLYEMFSGRPPFFSETMSELVEKILYEDPLPPIPKDSSFPKASSDFINLLD
GLLQKDPQKRLSWEGVLQHPFWKDALRREDSDSVSEDSTFSSRNVMECSGPHDSRELLQS
PKNGQAKGQKGAHRLSQSFRLENPTELRPKSIMGGQLNESIFLLSSRPTPRTSAMVELNP
GEGEDPSSPQKTSPLSKMTSGHLSQGALESQMRELIYTDSDLVITPIIDNPKIMKQPAIK
FDPKILHLPAYSVEKLLALKDQDWSDFLQQLCSHVDSSEKSTGALRAKLNLLCYLCVVAT
HKEVATRLLHSPLFQLLIQHLRIAPNWDIRSKVARVVGMLALHTAELQESVPVIEAITLL
TELIRENFRSGKLKQCLLPTLGQLLYLVATQEEKTQHSRECWSVPLAAYTVLMRCLREGE
ERVVNHMAAKIIENVCTTFSAQAQGFTTGEIGPVLWHLFRHSTVDALRITAISALCRITR
QSPTAFQNVIEKVGLNAVISSLASAICKVQQYMLTLFTAMLSCGIHLQRLIQEKDFVSTV
IRLLDSPSTPIRAKAFLVLLYVLIHNRDMLLLSCQARLVMYIERDSRKTSPGKELQSGNE
YLARCLDLLIQHMVQESPRILGDILNALANVSGRKHPSTVQGKQLKMCLPMMPVVLHLVM
SQVFRPQVVTEEFLFSYGTILSHIKSIDLGETNIDGAIGIVASEEFIKITLSAFEAVIQY
PVLLADYRATVVDYILPPLVSLVQSQNVEWRLFSLRLLSETTTLLVSQEPEDGDEEASCD
SDSGLLALIRDELLPQYEHILMEPDPVPAYALKLLVAMTEHNPAFTRLVAESKLVPLIFE
VILEHQESILGNTMQSVIALLNNLVAYKDSNMQLLYEQGLVGHVCNMFTETATLCLDRDN
KTNTEPAATLLASLLDILLGMLTYTSRIVRQALQAQKSGSRGDTQAAEDLLLLSKPLTDL
ISLLIPLLPSEDPEISEVSSKCLSILVQLYGGENPESLSPENLVTFADLLMAKEDPKDQK
LLLRILKRMVTSNEKLLESLRNTGSLLQALERLAPAHSSPVDTVVASLALDLLQAVGH-
```

```
>tr|D4ADQ0|D4ADQ0_RAT Protein Ulk4(1270 a.a.)

MENPVLYEEIGRGSRTVVYKGRRKGTINFVAILCTEKCKRPEITNWVSLGSKVTEAMVVV
VGGWWVVGGGWMMVGSILVGGSLETVIAQDQNLPEDVVREFGVDLVTGLHHLHRLGILFC
DLSPGKILLEGPGTLKFSNFCLAKVEGESLEEFFALVAAEEGGGDSGENTLRKSMKTRVR
GSLIYAAPEIVTGTEFSVTSDLWSLGCLLYEMFSGRPPFFSETMSELVEKILYEDPLPPI
PKDSSFPKASSDFINLLDGLLQKDPQKRLSWEGVLQHPFWKDALRREDSDSVSEDSTFSR
NVMECSGPHDSRELLQSPKNGQAKGQKGAHRLSQSFRLENPTELRPKSIMGGQLNESIFL
LSSRPTPRTSAMVELNPGEGEDPSSPQKTSPLSKMTSGHLSQGALESQMRELIYTDSDLV
ITPIIDNPKIMKQPAIKFDPKILHLPAYSVEKLLALKDQDWSDFLQQLCSHVDSSEKSTG
ALRAKLNLLCYLCVVATHKEVATRLLHSPLFQLLIQHLRIAPNWDIRSKVARVVGMLALH
TAELQESVPVIEAITLLTELIRENFRSGKLKQCLLPTLGQLLYLVATQEEKTQHSRECWS
VPLAAYTVLMRCLREGEERVVNHMAAKIIENVCTTFSAQAQGFTTGEIGPVLWHLFRHST
VDALRITAISALCRITRQSPTAFQNVIEKVGLNAVISSLASAICKVQQYMLTLFTAMLSC
GIHLQRLIQEKDFVSTVIRLLDSPSTPIRAKAFLVLLYVLIHNRDMLLLSCQARLVMYIE
RDSRKTSPGKELQSGNEYLARCLDLLIQHMVQESPRILGDILNALANVSGRKHPSTVQGK
QLKMCLPMMPVVLHLVMSQVFRPQVVTEEFLFSYGTILSHIKSIDLGETNIDGAIGIVAS
EEFIKITLSAFEAVIQYPVLLADYRATVVDYILPPLVSLVQSQNVEWRLFSLRLLSETTT
LLVSQEPEDGDEEASCDSDSGLLALIRDELLPQYEHILMEPDPVPAYALKLLVAMTEHNP
AFTRLVAESKLVPLIFEVILEHQESILGNTMQSVIALLNNLVAYKDSNMQLLYEQGLVGH
VCNMFTETATLCLDRDNKTNTEPAATLLASLLDILLGMLTYTSRIVRQALQAQKSGSRGD
TQAAEDLLLLSKPLTDLISLLIPLVRAGLCEGSEKQRRKQCYLNFLYGGENPESLSPENL
VTFADLLMAKEDPKDQKLLLRILKRMVTSNEKLLESLRNTGSLLQALERLAPAHRYEEPS
WAGMSCALVQ
```

Figure 28. Nouvelle isoforme potentielle identifiée à cause d'une séquence écourtée

Séquences d'une nouvelle isoforme potentielle TCONS_00112253_j_5_1, identifiée dans les spermatocytes, et celle de la protéine connue Ulk4, dont le gène se trouve sur la même localisation sur le génome du rat (XLOC_047617). Les peptides identifiés par MS/MS sont signalés en vert, la méthionine initiatrice, en rouge. Ulk4 n'est pas identifiée, alors que TCONS_00112253_j_5_1 est identifiée dans l'échantillon, grâce à une plus grande couverture de séquence, en effet, elles ont une longueur de 1270 a.a. et 1198 a.a. respectivement.

b) Séquence erronée : isoforme potentielle de la Cofilin1

L'exemple de l'ORF TCONS_00002856 identifié au lieu de la Cofilin1 illustre le cas d'une séquence protéique incorrecte (Figure 29). De nombreux exemples comme celui-ci, c'est à dire de séquences correctes sur une grande partie, mais écourtées et avec une petite portion incorrecte, sans doute à cause d'un changement délétère de cadre de lecture ou des erreurs de séquençage des transcrits, peuvent être trouvés dans nos listes. C'est aussi l'exemple de la protéine Ssh2 (F1M4Q5_RAT). Cette dernière n'est pas identifiée malgré les 3 peptides identifiés dans les spermatocytes qui s'alignent parfaitement sur sa séquence, au profit d'un ORF plus court: TCONS_00014761.

```
TCONS_00002856_j_1_39 (nouvelle isoforme potentielle)
5'-3'Frame2
-LHPFPRFVFRAPESAPLKSKMIYASS
KDAIKKKLTGKDPFLCMHFCVSHDGSLEYMTILGSSVVLTAGVPSSCPLF-

>sp|P45592|COF1_RAT Cofilin-1 OS=Rattus norvegicus GN=Cfl1 PE=1 SV=3
MASGVAVSDGVIKVFNDMKVRKSSTPEEVKKRKKAVLFCLSEDKKNIILEEGKEILVGDV
GQTVDDPYTTFVKMLPDKDCRYALYDATYETKESKKEDLVFIFWAPESAPLKSKMIYASS
KDAIKKKLTGIKHELQANCYEEVKDRCTLAEKLGGSAVISLEGKPL
```

Figure 29. Nouvelle isoforme potentielle identifiée avec une séquence incorrecte

Séquence de TCONS_00002856_j_1_39 identifiée dans les spermatocytes, mais qui possède une séquence non correcte. Les peptides identifiés par MS/MS sont signalés en vert, la méthionine initiatrice, en rouge. La protéine connue Cofillin1 correspondant à sa localisation XLOC_02117 n'est pas identifiée, pour les mêmes raisons que Figure 26.

c) ORFs multiples : isoformes potentielles de WDR62

Un autre exemple peut être donné par WDR62, pour laquelle de nombreux peptides sont identifiés dans les spermatocytes pachytène et les spermatides rondes, mais qui n'apparaît pas dans les listes de protéines identifiées au profit de plusieurs isoformes potentielles plus courtes (Tableau 7).

Identifiant	Locus	Cat	Gene	Taille cumulative des exons (b)	Nombre d'exons	Transcrit (RNA-seq)	Protéine (MS/MS)	Protéine connue identifiée
TCONS_00000492	XLOC_000393	iso	Wdr62	4888	31	pSPC;rSPT	pSPC;rSPT	-
TCONS_00000493	XLOC_000393	iso	Wdr62	1801	16	pSPC;rSPT	pSPC;rSPT	-
TCONS_00005093	XLOC_000393	iso	Wdr62	676	4	pSPC;rSPT	pSPC;rSPT	-
TCONS_00005094	XLOC_000393	iso	Wdr62	4372	25	pSPC;rSPT	pSPC;rSPT	-
TCONS_00005095	XLOC_000393	iso	Wdr62	3862	21	pSPC;rSPT	pSPC;rSPT	-
TCONS_00008857	XLOC_000393	iso	Wdr62	4556	32	pSPC;rSPT	pSPC;rSPT	-

Tableau 7. Nouvelles isoformes potentielles de WDR62 identifiées dans les cellules germinales

WDR62 n'est pas identifiée dans les pSPC et les rSPT au profit de ces nouvelles isoformes potentielles. Les identifiants « TCONS » pour chaque nouvelle isoforme potentielle correspondent aux séquences déduites des transcrits reconstruits par RNA-seq. Le locus, la catégorie du transcrit (« iso », pour nouvelle isoforme potentielle), le nom du gène, la taille des exons additionnés, et le nombre d'exons du transcrit sont indiqués, ainsi que les types cellulaires dans lesquels les protéines ont été détectées.

Toutes ces différentes isoformes potentielles de WDR62 ne sont certainement pas présentes, et leurs séquences dérivées des transcrits reconstruits plus courts que l'ARNm de Wdr62 donnent des ORFs plus courts identifiés plus facilement que WDR62 (Figure 30).

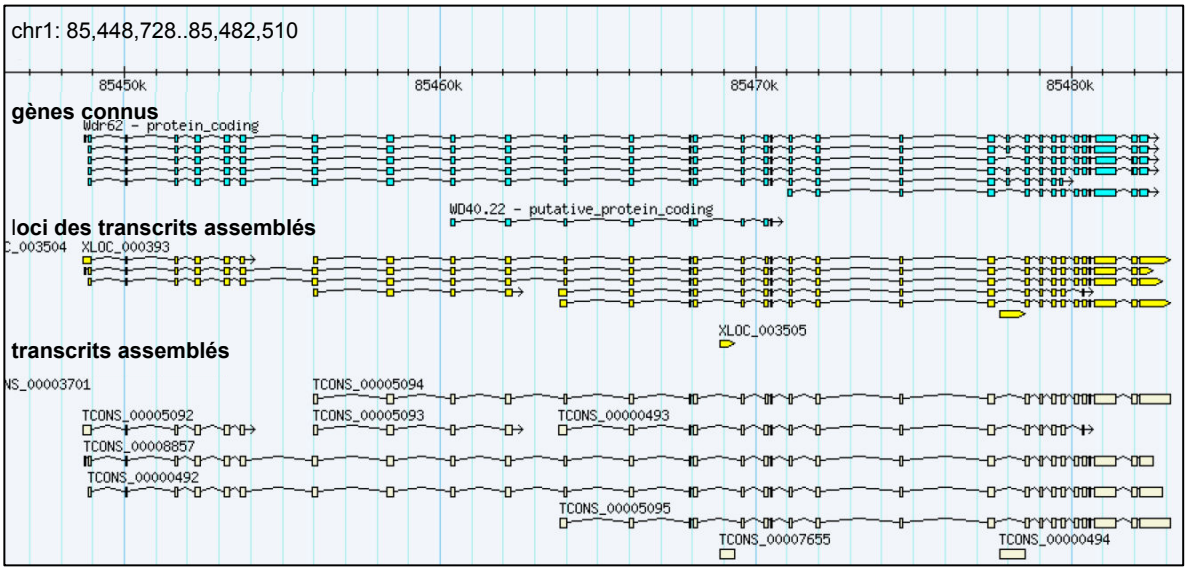

Figure 30. Localisation du gène Wdr62 et de nouveaux transcrits assemblés par RNA-seq à cette même localisation

Localisation des gènes connus en bleu (dont Wdr62), des transcrits reconstruits après RNA-seq (en beige), et leur loci (XLOCs, en jaune), à la position chr1: 85,448,728..85,482,510. On observe que de nombreux transcrits reconstruits ont des séquences plus courtes que le transcrit Wdr62.

L'approche PIT permet donc d'identifier de nombreuses isoformes potentielles parmi lesquelles un certain nombre sont de faux positifs. Des analyses manuelles ou bien l'application du filtre supplémentaire « peptide unique » lors de la sélection pour trouver des isoformes potentielles réelles, sont nécessaires si l'on veut découvrir de nouvelles isoformes à l'aide de cet ensemble de données en l'état.

III. Conclusion

Bien qu'il soit possible de découvrir de nouvelles isoformes à l'aide de peptides jonctionnels révélant de nouveaux évènements d'épissage ou à l'aide de peptides révélant des UTRs codants, la détection des identifications faites grâce à ces peptides n'est pas automatique et requiert une recherche manuelle. Pour garantir la détection de nouvelles isoformes grâce à des peptides jonctionnels ou spécifiques d'un nouvel exon, nous pourrions nous limiter à l'analyse des identifications de catégorie « iso » qui sont caractérisées par un peptide unique. En revanche, ceci conduirait à une liste très réduite des isoformes présentes dans les cellules germinales étudiées. Ici, seulement 57 nouvelles isoformes dans les pSPC et les rSPT seraient analysables avant d'appliquer des filtres de sélection du profil d'expression des transcrits. Les

résultats obtenus dans notre étude nous donnent des indices peu fiables sur leur proportion parmi les identifications nouvelles dérivées du RNA-seq. De ce fait, des étapes d'analyse supplémentaires sont nécessaires afin d'identifier de nouvelles isoformes réelles. Une étude utilisant une stratégie de type PIT qui aurait pour but la découverte de nouvelles isoformes ou de nouvelles jonctions d'épissage pourrait faire appel à une banque personnalisée différente de celle que nous avons utilisée. C'est à dire par exemple, une banque de peptides jonctionnels théoriques, ainsi que cela a été présenté récemment (Sheynkman et al., 2013). Dans cette étude, une banque de séquences nucléotidiques jonctionnelles entre plusieurs exons, étendues d'un nombre fixe de paires de bases de part et d'autre de la jonction sur le brin transcriptionnel, puis traduites dans les trois cadres de lecture, est générée. Ces séquences, transformées en peptides trypsiques théoriques constituant une base de séquences de peptides jonctionnels potentiels sont ajoutées aux bases canoniques comme UniProt et Ensembl. On obtient ainsi une banque personnalisée contre laquelle interroger les données MS/MS (Sheynkman et al., 2013). Par alignement des peptides identifiés sur le génome, ces auteurs en déduisent les nouveaux exons et évènements d'épissage.

Chapitre 3

Analyse protéomique différentielle des protéines membranaires dans les cellules méiotiques et post-méiotiques et dans les corps résiduels chez le rat

IV. Contexte et objectifs de l'étude

Les régulations paracrines et autocrines de la spermatogenèse impliquent des facteurs solubles sécrétés par les cellules de Sertoli et qui peuvent agir sur les cellules germinales (Griswold, 1998; Jégou, 1993), établissant un dialogue essentiel au fonctionnement de ces deux types de cellules (Eddy, 2002; Griswold, 1995, 1998; Jégou, 1995; Syed et Hecht, 1997). Ces facteurs sont susceptibles d'agir sur des récepteurs membranaires à la surface des cellules germinales (O'Brien et al., 1993; Tsuruta et al., 2000). Les phénomènes d'activation des cellules germinales par les cellules de Sertoli se font *via* de nombreux facteurs diffusibles dont nous avons parlé dans l'introduction générale, mais des expériences de co-culture montrent que cette stimulation est optimale lorsque les cellules germinales sont au contact des cellules de Sertoli (Saez et al., 1986), ce qui peut impliquer des protéines membranaires de ces deux types cellulaires. En effet le contact cellulaire est essentiel à la modulation de certains réseaux de signalisation conduisant à des destinées cellulaires différentes, comme l'ont montré Jorgensen et collaborateurs en étudiant la signalisation Ephrin/ Ephrin-récepteur (Jørgensen et al., 2009). Ces auteurs ont évalué les phosphorylations sur les tyrosines de protéines régulatrices impliquées dans différents processus biologiques (adhésion, polarité, signalisation phosphoinisitol, endocytose…) par protéomique différentielle dans des cellules exprimant Ephrin-récepteur stimulées par contact avec des cellules exprimant le ligand EphrinB1. À cette fin, ils ont comparé le niveau de peptides phosphorylés sur des tyrosines dans les deux types de cellules en contact. Une signalisation phospho-Tyrosine bidirectionnelle asymétrique en aval du récepteur et du ligand dans ces deux types de cellules a été démontrée. Elle affecte des régulateurs différents de la destinée cellulaire *via* un « flux d'information » cellulaire-spécifique pendant la signalisation bidirectionnelle due au contact.

D'un autre côté, il est également connu que les cellules germinales modulent la fonction des cellules de Sertoli par la sécrétion de facteurs solubles (Jégou, 1991; Onoda et al., 1991; Jégou et al., 1993; Onoda et Djakiew, 1993; Pineau et al., 1993). Des petites protéines issues de spermatides ont été associées à une modulation de la sécrétion des cellules de Sertoli (Onoda et Djakiew, 1993). Les interactions et communications Sertoli/germinales et vice-versa mettant en jeu des protéines membranaires sont loin d'être bien connues, et les études protéomiques différentielles menées jusqu'à présent n'ont pas permis d'établir de listes de protéines membranaires différentiellement exprimées dans les cellules germinales. Il nous a donc semblé intéressant d'identifier et de quantifier de manière relative un certain nombre de

protéines membranaires des spermatocytes pachytène, des spermatides rondes et des corps résiduels (dans la mesure où le protéome de ce dernier peut refléter celui des spermatides allongées), afin d'identifier des protéines susceptibles d'intervenir dans les communications cellulaires entre les cellules germinales et les cellules de Sertoli au cours de la spermatogenèse, et plus particulièrement, de la spermiogénèse. Nous proposons pour cela, l'utilisation de la technologie ICPL de quantification relative des protéines couplée à la LC-MS/MS sur une LTQ-Orbitrap.

V. Résultats et discussion

La sélection des protéines quantifiées, à savoir identifiées avec au moins deux peptides marqués (un peptide est considéré "marqué" dans l'analyse s'il est marqué par une étiquette de masse dans les trois échantillons); dans au moins deux triplex techniques; a donné 226 protéines quantifiées. Après l'application d'un seuil de ratio d'expression différentielle entre deux échantillons fixé à une valeur de 2 suivi d'un test statistique de LIMMA pour déterminer les protéines significativement différentielles, nous identifions 166 protéines différentiellement exprimées entre les spermatocytes pachytène (pSPC), les spermatides rondes (rSPT) et les corps résiduels (CRs). Ces 166 protéines ont été classées en cinq groupes selon leur profil d'expression (P1 à P5). Les protéines présentent des pics d'expression dans: P1, pSPC et rSPT; P2, rSPT; P3, pSPC et CRs; P4, rSPT et CRs; et enfin P5, CRs. Ceci nous permet d'établir une première carte d'expression des protéines membranaires des cellules germinales et des CRs.

En utilisant les annotations de la Gene Ontology ou des bases KEGG et REACTOME, j'ai sélectionné un ensemble de 8 protéines qui n'ont pas été décrites dans le testicule et sont susceptibles d'avoir un rôle dans la transduction du signal, la différenciation ou la reconnaissance cellulaire. Ces protéines candidates pourraient faire l'objet d'analyses biochimiques ultérieures afin de valider leur expression différentielle dans les cellules méiotiques et post-méiotiques sur des coupes de testicule de rat.

Par ailleurs, quatre protéines hypothétiques ou « uncharacterized » sont démontrées pour la première fois dans notre étude comme étant différentiellement exprimées dans les cellules germinales de rat. Les fonctions de ces protéines dans la spermatogenèse ne sont pas connues

mais sont susceptibles d'exister étant donné leur profil d'expression. Ces protéines pourront faire l'objet de futures investigations.

L'approche ICPL que nous avons utilisée bénéficie de la technologie Orbitrap mise au service d'une technique de quantification relative qui nous permet d'obtenir 166 protéines différentielles. Ce nombre ne représente pas une avancée spectaculaire comparé aux 123 protéines différentielles cytosoliques identifiées entre les cellules germinales, obtenues par Rolland et collègues (Rolland et al., 2007) avec la technologie DIGE. En revanche, ces deux études se complètent, l'une ayant concerné les protéines cytosoliques germinales, et notre étude, les protéines membranaires germinales. Pour diverses raisons inhérentes à l'approche ICPL, seulement 34% des protéines identifiées sont quantifiées par cette technique. En effet, certains peptides peuvent ne pas être marqués par le réactif, (même si le marquage se fait sur les lysines et les N-ter). D'autre part, les peptides sont considérés pour la quantification seulement si ils sont marqués dans les trois échantillons. Enfin, les protéines qui pourraient être quantifiées par un seul peptide non redondant ne sont pas prises en compte dans l'analyse de quantification. Ces limitations nous font perdre un certain nombre d'identifications différentielles. De plus, les protéines qui sont différentiellement exprimées mais qui sont absentes à l'un de stades étudiés malgré une forte expression à un autre stade ne peuvent pas être identifiées par une approche de quantification relative. En effet, cette technologie permet de quantifier des protéines en se basant sur la relation entre les aires acquises en MS des différents peptides marqués dans les différents échantillons, mais ne permet pas d'accéder à l'information « présent/absent » d'une protéine dans l'un ou l'autre des types cellulaires.

Si les protéines membranaires différentielles ne sont pas si nombreuses à être identifiées par cette approche, elles sont en revanche enclines à être impliquées même indirectement dans les communications cellulaires au sein de l'épithélium séminifère. Ainsi, nous décrivons dans l'article en préparation qu'un certain nombre de protéines déjà connues comme ayant un rôle crucial dans la spermatogenèse ou la fécondation, telles que EHD1 requise pour la spermiation (Rainey et al., 2010) et impliquée dans la signalisation IGF1 (Rotem-Yehudar et al., 2001), se retrouvent dans les 166 protéines différentielles. Nous montrons par exemple pour EHD1 une sur-expression relative dans les corps résiduels, et pour un certain nombre de protéines connues, nous apportons aussi un profil d'expression dans les cellules germinales et les corps résiduels. Enfin, même si l'accent est mis dans ce chapitre sur les 8 candidats listés dans l'article, les 158 autres protéines différentielles pourront aussi faire l'objet d'investigations supplémentaires.

ARTICLE 2

Quantitative proteomic Isotope-Coded Protein Label (ICPL) analysis reveals 166 membrane proteins differentially expressed in rat meiotic and post-meiotic germ cells and in residual bodies

Quantitative proteomic Isotope-Coded Protein Label (ICPL) analysis reveals 166 membrane proteins differentially expressed in rat meiotic and post-meiotic germ cells and in residual bodies

Keywords: *spermatogenesis, germ cell, proteome, membrane proteins, residual bodies*

Authors and affiliations:

Sophie Chocu [1,2], Aurélie Lardenois, Mélanie Lagarrigue, Régis Lavigne [1,2], and Charles Pineau [1,2]

[1] *Proteomics Core Facility Biogenouest, Inserm U1085, IRSET, Campus de Beaulieu, F-35042 Rennes, France*

[2] Inserm U1085-IRSET, Université de Rennes 1, F-35042 Rennes, France

Grant support : This work was supported by Biogenouest and by Infrastructures en Biologie Santé et Agronomie (IBiSA) ;Région Bretagne ; Fonds Européen de Développement Régional and Conseil Régional de Bretagne grants awarded to C.P. ; by l'Institut national de la santé et de la recherche médicale (Inserm); l'Université de Rennes 1.

Correspondence: *Charles Pineau, Proteomics Core Facility Biogenouest, Inserm U1085-IRSET, Université de Rennes 1, 263 av. du Général Leclerc, 35042 Rennes cedex, France; Tel: +33 (0)2 23 23 52 79; Fax: +33 (0)2 23 23 52 82; Email: charles.pineau@inserm.fr*

Abstract

Within the seminiferous epithelium, it is well established that Sertoli cells and germ cells communicate. Although the intricate physical interactions between Sertoli and germ cells are well described including the blood testis barrier and its regulations, their extremely elaborate dialogue remains mostly unrevealed, and many molecular actors of their communications remain unidentified. These communications involve membrane proteins on germ cells that may be specific at each step of male germ cells differentiation. We assessed membrane protein relative expression between pachytene spermatocytes, round spermatids and residual bodies to identify those proteins specifically or preferentially expressed at meiotic, post-meiotic stages and in residual bodies using a proteomic isotope coded protein labeling (ICPL) relative quantification approach. Membrane fractions obtained using differential centrifugation from pachytene spermatocytes (pSPC), round spermatids (rSPT) and residual bodies (RB) were labeled with ICPL reagents, subjected to 1D prefractionation and LC-MS/MS analysis on a LTQ-Orbitrap. After selection, we found 226 proteins quantified; and after carrying out statistical analysis and setting a relative expression ratio cutoff of 2 between two samples, we identified 166 proteins with a differential expression between spermatocytes, spermatids and residual bodies. The differential proteins were then classified into five expression patterns corresponding to a preferential relative expression in pSPC/rSPT; rSPT; pSPC:RB; rSPT/RB and RB. Using the Gene Ontology annotations, we pointed out 8 potentially interesting candidates involved in signal transduction / cell recognition or differentiation that had never been described in the testis, for further experimental validation. We also confirmed the protein existence for 23 proteins translated from GENSCAN gene predictions which have a differential expression. They constitute potential candidates for further investigation of their functions in spermatogenesis. It is also the case of four differential uncharacterized/hypothetical proteins.

Introduction

Mammalian spermatogenesis is a sophisticated process encompassing a series of events including proliferation and differentiation of spermatogonia taking place in the basal region of the seminiferous epithelium, meiotic division of primary and secondary spermatocytes, and differentiation of haploid spermatids, leading to the production of mature male gametes in the luminal region of the seminiferous epithelium (Matzuk and Lamb, 2002). Sertoli cells are intimately involved at each step of spermatogenesis through its nourishing and physical support (Jégou, 1993, 1995; Griswold, 1998). This process takes place through at coordinated and sequential gene expression program leading to gene products specifically expressed at each stage, and which are the essential actors of their smooth running (Eddy, 2002; Griswold, 1995). These gene products result in, or are the target of strict transcriptional regulation and post -transcriptional regulation at each stage (Bettegowda and Wilkinson, 2010). The testis is described to be the organ that expresses the highest number of tissue-specific genes (Son et al., 2005; Chalmel et al., 2007a, 2012; Laiho et al., 2013; Soumillon et al., 2013; Chalmel et al., 2014; Djureinovic et al., 2014). For twenty years, a large number of genes regulated in a spatiotemporal manner in the testis and in differentiating germ cells have been identified and their expression pattern were assessed by high throughput transcriptomic studies in rat, mouse and human (Schultz et al., 2003; Schlecht et al., 2004; Wrobel and Primig, 2005; Chalmel et al., 2007a, 2012). From these studies, different clusters of significantly differentially expressed transcripts with somatic, mitotic, meiotic, and post-meiotic expression profiles have emerged. These transcriptomic studies provide new insights into gene expression mechanisms that support spermatogenesis. However, the main actors of biological processes are the proteins. Thus, parallel efforts were invested into large scale study of proteins expression during spermatogenesis.

These studies lead to the establishment of isolated germ cell proteomes (Guillaume et al., 2000; Com et al., 2003a) or of the whole testis proteome during the first wave of spermatogenesis in the rat (Zheng et al., 2014) or in the adult pig (Huang et al., 2005), mouse (Zhu et al., 2006), human (Guo et al., 2008) and macaque (Wang et al., 2014). Several differential proteomic studies were also carried out on isolated germ cells with the aim to evaluate the expression profiles of a certain number of proteins at key spermatogenetic stages (Rolland et al., 2007; Gan et al., 2013). Importantly, Gan and coworkers performed a proteomic approach based on iTRAQ and LC-MS/MS that allowed them to identified 2,008 proteins in spermatogonia, pachytene spermatocytes, round spermatids and elongated spermatids in mice. They were able to bring out five mechanisms of transcriptional regulation, by comparing the proteins expression profiles (iTRAQ) with the transcripts expression profiles (by analyzing microarrays results) during different stages of spermatogenesis (Gan et al., 2013). Such differential studies are very informative about the regulation mechanisms occurring during spermatogenesis. However, these proteomic studies have not allowed to shed light on intricate interactions between Sertoli and germ cells within the seminiferous epithelium.

Interactions between germ cells and Sertoli cells are far from being known. From the functional point of view, the strategic position of the Sertoli cell allows it to receive, integrate, and transmit all the signals that act at the membrane of germ cells, required for the process of spermatogenesis from the extra-tubular compartment or from germ cells themselves. Its location also allows it to coordinate the activity of the GC within the seminiferous tubule. Thus, germ cells are capable of receiving, integrating and transmitting signals, which involve membrane proteins (Jégou, 1993).

Apart the study of Rolland and coworkers carried out on germ cells cytosolic fractions using a 2D-DIGE approach, no differential proteomic studies of male germ cells reported proposes to quantify the sub-proteomes of these cells at different stages of differentiation. Several studies were conducted on membrane proteins from male germ cells because they have a key role in the acquisition of fertilizing capacity (Belleannee et al., 2011; Cooper, 1998; Dacheux et al., 1998; Mori et al., 2012). However, to date, no differential studies were ever conducted to answer protein expression dynamics at the plasma membrane of germ cells throughout spermatogenesis. In order to tackle this subject, we compared membrane extracts from meiotic (pachytene spermatocyte), post-meiotic (round spermatids) germ cells, and residual bodies using a isotope coded protein labeling (ICPL) proteomic relative quantification approach coupled with a LC-MS/MS analysis. We identified a first set of 166 membrane proteins differentially expressed within the rat male germ lineage, and pointed out 8 potentially interesting candidates for further experimental validation.

Materials and Methods

Ethics statement

Experimental research on animal reported here was performed in conformity with the principles for the use and care of laboratory animals in compliance with French and European regulations on animal welfare. Furthermore, experimenters were delivered an authorization given by the French “Direction des Services Vétérinaires” to conduct or supervise experimentations on live animals.

Animals

Male Sprague-Dawley rats of 90 or 20 days were used for testicular cell isolation. Animals were purchased from Elevage Janvier (Le Genest-Saint-Isle, France).

Isolation of testicular cells

Pachytene spermatocytes (pSPC), early spermatids (rSPT) and residual bodies (RBs) were prepared by centrifugal elutriation with a purity greater than 90% according to a method previously described (Pineau et al., 1993) with the exception that enzymatic dissociation of cells was replaced by a mechanical dispersion. Celle pellets were frozen and maintained at -80°C until use.

Membrane protein preparation

Frozen cell pellets of rat pachytene spermatocytes and round spermatids were resuspended in a PIPES extraction buffer (100 mM PIPES, 70mM NaCl, 2 mM $MgCl_2$, pH 7.4). Frozen pellets of residual bodies were ressuspended in a Tris extraction buffer (10mM Tris,10mM $MgCl_2$: pH 7,4). A cocktail of protease inhibitors with 1mM EDTA, 0.5mM DTT, 1mM 4-(2-Aminoethyl) benzenesulfonyl fluoride hydrochloride (AEBSF), 10mM trans-Epoxysuccinyl-leucylamido(4-guanidino)butane (E64), 0.6U/mL Nuclease (Sigma-Aldrich, Saint-Quentin Fallavier, France) was added to the extraction buffer just before use. Cell suspensions were subjected to sonication with an ultrasonic processor (Bioblock Scientific, Illkirch, France) six times for 10 s with 30 s stop in between using a microtip setting power level at 40% pulse duration. Cell lysates were centrifuged at 1000g and 4°C for 10 min to remove cellular debris. The supernatants were then centrifuged at 105,000g at 4°C for 1h. The pellets were washed in 100mM Na_2C0_3, and sonicated as described above. These suspensions were centrifuged at 105,000g, 4°C for 45 minutes. The pellets were further washed in 100mM Na_2C0_3, Na Cl 1M, sonicated and again centrifuged at 105,000g, 4°C for 45 minutes. The final membrane pellets were retrieved in ICPL buffer (6 M guanidine HCl, pH8.5). The protein concentration was determined using the Bradford colorimetric assay (Bio-Rad, Marnes-la Coquette, France), and protein extracts were stored at –80°C until use.

Total protein extraction and sample preparation

In order to assess the enrichment in membrane proteins of our membrane extracts compared to total protein extracts, we prepared total proteins from a rSPT pellet. The frozen cell pellet was treated as described above, except that the first 105,000g supernatant containing the soluble proteins were pooled with the final membrane pellet. The protein concentration was determined using the Bradford colorimetric assay (Bio-Rad), and 100µg of proteins for each sample were separated by SDS-PAGE onto a 12% precast gel (NuPage Novex Bis Tris Mini Gel; Invitrogen), in a MES SDS running buffer. The gel was subsequently stained with Coomasie blue, using the EZBlue gel staining reagent (Sigma-Aldrich,Saint-Quentin Fallavier, France) for 45 minutes, destained overnight and cut into 20 slices. We processed each of the 20 bands of both samples for subsequent LC-MS:MS analysis as follows. Slices were first treated with 50mM NH_4HCO_3 in acetonitrile/water 1:1 (v/v), dehydrated with 100% acetonitrile and rehydrated in 100mM NH_4HCO_3. They were washed again with 50 mM NH_4HCO_3 in acetonitrile/water, 1:1 (v/v) and dehydrated with 100% acetonitrile. The slices were then treated with 65mM DTT for 15 min at 37 °C, and with 135mM iodoacetamide in the dark at room temperature. Finally, the samples were washed with 100mM NH_4HCO_3 in acetonitrile/water, 1:1 (v/v), and dehydrated with 100% acetonitrile before being washed with 100 mm NH_4HCO_3 in acetonitrile/water, 1:1 (v/v) and then dehydrated again with 100% acetonitrile. Gel slices were dried 20min at 37°C. In-gel digestion was performed overnight at 37°C with modified trypsin (Promega, Charbonnières-les-Bains, France) with 12,5 ng/µL trypsin in 50mM NH_4HCO_3 at 37°C overnight. Trypsic peptides were then extracted from the gel by sequential incubation in the following solutions: acetonitrile/H_2O/TFA, 70:30:0.1 (v/v/v), 100% acetonitrile and acetonitrile/H_2O/TFA, 70:30:0.1 (v/v/v), and finally extracts were concentrated by evaporation down to a final volume of 30µL.

ICPL labeling of membrane proteins and sample preparation

ICPL labelling was performed on samples from membrane fractions of pSPC, rSPT and RB, following the experimental design described in Table 1. Membrane protein extracts from pSPC, rSPT and RB were adjusted to 2.5 mg/mL by addition of 6M guanidine HCL, pH8,5. Disulfide bonds were reduced with 0.2 M tris(2-carboxyethyl)phosphine and then alkylated with 0.4 mM iodoacetamide. For each sample, 50µg of proteins were labeled using the ICPL^{TM} Quadruplex-kit (Serva Electrophoresis, Heidelberg, Germany) according to the manufacturer's instructions. Briefly, free amino groups (lysine residues and N-terminal NH_2) of proteins from different cell types or residual bodies were labelled at room temperature for 2h with different ICPL reagents: the light ^{12}C-nicotinoyloxysuccinimide (ICPL_0), the heavy ^{13}C-nicotinoy-loxysuccinimide (ICPL_6 with 6^{13}C), or the super heavy ^{13}C-^{2}D-nicotinoyloxysuccinimide (ICPL_10 with 6^{13}C and 4 deuteriums). After quenching excess reagent with 6M hydroxylamine, the three labelled samples were mixed, purified by acetone precipitation (−20 °C, overnight), and subsequently dissolved in 1× LDS NuPage Sample buffer (Invitrogen, Saint Aubin, France). For each triplex, mixed labeled proteins (50µg) were separated by SDS-PAGE onto a 12% precast gel (NuPage Novex Bis Tris Mini Gel; Invitrogen), in a MES SDS Running Buffer. The gel was subsequently stained with Coomasie blue, using

the EZBlue gel staining reagent (Sigma-Aldrich,Saint-Quentin Fallavier, France) for 45 minutes, and destained overnight. For each triplex, the entire gel lane was cut into 20 bands. Gel slices were washed with 100mM NH_4HCO_3/acetonitrile 1:1, then dehydrated with 100% acetonitrile; and washed again with 100mM NH_4HCO_3, and 100% acetonitrile. Proteins were digested and peptides were extracted as described in the last section.

LC MS/MS analysis

Data acquisition

MS measurements of peptide extracts were performed with a nanoflow high-performance liquid chromatography (HPLC) system (Ultimate 3000, Thermo Scientific Dionex, Bremen, Germany) connected to a hybrid LTQ-OrbiTrap XL (Thermo Scientific, Bremen, Germany) mass spectrometer equipped with a nanoelectrospray ion source (New Objective, Woburn, MA, USA). The HPLC system consists of a solvent degasser nanoflow pump, a thermostated column oven kept at 30 °C, and a thermostated autosampler kept at 8 °C to reduce sample evaporation. Mobile A (99.9% MilliQ water and 0.1% formic acid (v:v)) and B (99.9% acetonitrile and 0.1% formic acid (v:v)) phases for HPLC were delivered by the Ultimate 3000 nanoflow LC system (Dionex). Peptide mixture was injected with a 10 µL volume, and loaded on a trapping precolumn (5 mm × 300 µm i.d., 300 Å pore size, Pepmap C18, 5 µm) for 3 min in 2% buffer B at a flow rate of 25 µL/minute. This step was followed by reverse-phase separations at a flow rate of 0.250 µL/min using an analytical column (15 cm × 300 µm i.d., 300 Å pore size, Pepmap C18, 5 µm, Dionex). We ran a gradient ranging from 2 to 35% buffer B for the first 60 min, 35 to 60% buffer B from minutes 60–85, and 60 to 90% buffer B from minutes 85–105. Finally, the column was washed with 90% buffer B for 16 min, and with 2% buffer B for 19 min prior to loading of the next sample.

The MS instrument was operated in its data-dependent mode by automatically switching between full survey scan MS and consecutive MS/MS acquisition. Survey full scan MS spectra (mass range 400–2000) were acquired in the OrbiTrap section of the instrument with a resolution of r = 60,000 at m/z 400; ion injection times are calculated for each spectrum to allow for accumulation of 10^6 ions in the OrbiTrap. The seven most intense peptide ions in each survey scan with an intensity above 2000 counts (to avoid triggering fragmentation too early during the peptide elution profile) and a charge state ≥ 2 were sequentially isolated at a target value of 10,000 and fragmented in the linear ion trap by collision induced dissociation. Normalized collision energy was set to 35% with an activation time of 30 ms. Peaks selected for fragmentation were automatically put on a dynamic exclusion list for 120 s with a mass tolerance of ± 10 ppm to avoid selecting the same ion for fragmentation more than once. The following parameters were used: the repeat count was set to 1, the exclusion list size limit was 500, singly charged precursors were rejected, and a maximum injection time wet was set at 500 ms and 300 ms for full MS and MS/MS scan events, respectively. For an optimal duty cycle the fragment ion spectra were recorded in the LTQ mass spectrometer in parallel with the OrbiTrap full scan detection. For OrbiTrap measurements, an external calibration was used before each injection series ensuring an overall error mass accuracy below 5 ppm

for the detected peptides. MS data were saved in RAW file format (Thermo Fisher Scientific) using XCalibur 2.0.7 with tune 2.4.

Data Processing

Identification of the quantified peptides

The data analysis was performed with the Proteome Discoverer 1.2 software (Thermo Fisher Scientific) supported by Mascot (Matrixscience) database search engine for peptide and protein identification. For each ICPL triplex analysis, the 20 raw files corresponding to the 20 analyzed bands were queried trough ProteomeDiscoverer 1.2, to a custom sequence database that comprises protein entries from the UniProt (37,175 canonical and isoforms sequences, release 2012_10) and Ensembl (32,971 known and 44,993 predicted protein sequences*, release 3.4.68*) merged with a set of predicted protein sequences inferred from transcripts identified with high-throughput sequencing (Chocu et al., submitted). Proteome Discoverer provides a non-redundant identified protein list for each ICPL triplex analysis. Given that modification of lysine residues by ICPL labeling prevents their cleavage by trypsin, arginine C was selected as enzyme with one allowed miscleavage. In addition, carbamidomethylation of cysteins was set as fixed modification, and labeling of lysine residues and protein N-terminal by light, heavy, or super-heavy ICPL reagents, as well as methionine oxidation and lysine and N-terminal acetylation were considered as variable modifications. The mass tolerance for parent and fragment ions was set to 10 ppm and 0.5 Da, respectively. Identified peptides were filtered based on the Mascot score to obtain a false discovery rate of 5%. In the case of peptides shared by different proteins, proteins were automatically grouped. The proteins within a group were further ranked according to their protein score.

Relative protein quantification was obtained using the Proteome Discoverer 1.2 software, which automatically calculates the heavy to light (H/L), super-heavy to light (SH/L) and super-heavy to heavy (SH/H) ratios by comparing the relative areas of the extracted ion chromatograms (EIC), which are reconstituted by extraction of the intensities of m/z ratios corresponding to the labeled peptides observed on MS spectra. Only unique peptides (i.e., peptides not shared by different proteins) were used for protein quantification, and only peptide labeled with the three different ICPL reagents (i.e., labeled in each sample) in a given ICPL triplex were considered in the analysis.

Protein quantification in each biological sample

Protein quantification information was considered as relevant when a protein was quantified from at least two distinct peptide sequences distributed in at least two distinct experimental triplex. For a given protein, redundant peptide sequence in a single or in distinct triplex were counted as a single peptide.

The ratio data of labeled peptide areas between two samples were first log2-transformed. The median of log2-transformed ratio were calculated for the peptides that are redundant in a triplex. The quantification

of each protein in a triplex was then estimated as the median of the log2-transformed ratio of all the peptides assigned to this protein. The protein quantification data were pre-processed and analyzed using the AMEN (Annotation, Mapping, Expression and Network) suite of tools (Chalmel and Primig, 2008). Data quality was verified by plotting the log2-transformed ratios signal distribution. The log2-transformed ratio of the three triplex were then normalized using the “quantile-quantile” method.

Statistical filtration and classification

The quantified proteins displaying a high ratio (absolute value of the median of the ratio data ≥2) in at least two triplex were identified. A LIMMA statistical test (Smyth, 2004) was then used to identify the proteins that were significantly differentially expressed (F-value adjusted with the False Discovery Rate, $p \leq 0.05$). The resulting proteins were then classified into five expression patterns (P1-P5) using the PAM (Partitioning Around Medoids) algorithm.

Gene ontology enrichment analysis

The enrichments of the gene ontology (GO) terms within each of the five groups, using the GeneID as reference, were calculated using the Fisher exact probability using the Gaussian Hypergeometric test. A term was considered to be significantly over-represented as compared to the whole rat theoretical proteome when the number of genes in the group bearing this annotation is ≥3 and when the associated FDR-corrected p-value is ≤0.01.

To estimate GO terms enrichment in spermatid membrane fraction (m_rSPT) versus spermatid total extract (t_rSPT), the GO enrichments were calculated the same way, except that a term was considered to be significantly over-represented as compared to the other group when the number of genes in the group bearing this annotation is ≥15 and when the associated FDR-corrected p-value is ≤0.01.

Network analysis

The AMEN suite of tools (Chalmel and Primig, 2008) and protein interaction databases: IntAct (rhttp://www.ebi.ac.uk/intact), MINT (rhttp://160.80.34.4/mint/Welcome.do), BioGRID (http://thebiogrid.org/), and NCBI BIND and NCBI/BioGrid), have been used to assess the interactions between proteins.

RESULTS

Assessment of membrane protein enrichment

To assess the enrichment of membrane proteins in our preparations, we used a membrane extract from round spermatids (m_rSPT), and a total protein extract from round spermatids (t_rSPT) obtained as described in the materials and methods, and assessed the GO terms statistically over-represented in the membrane extract as compared to the total protein extract. We selected biological processes and molecular functions related to membrane proteins or that reflected the presence of membrane proteins. The biological processes over-represented in m_rSPT were found to be related to transport (GO: 0006810) or to localization (GO: 0051179), and the molecular functions over-represented in m_rSPT were related to transporter activity (GO: 0005215). The nucleus (GO: 0005634) and cytosol (GO: 0005829) were found to be enriched in the total protein extract (their children terms are not displayed here). On the opposite, cell surface (GO:0009986), endomembrane system (GO: 0012505) and its child term secretory granule, are over-represented in m_rSPT. As expected, membrane (GO: 0016020) and its children terms displayed without any selection in Figure 1 (e.g. plasma membrane, membrane raft) were over-represented in m_rSPT compared to the total protein extracts. Organelles (GO: 0043226) for which the children terms are also shown (Figure1) (e.g. mitochondrion), were also over-represented in m_rSPT .

Quantification of membrane proteins in meiotic and post-meiotic stages of germ cells differentiation

A pre-filtration of the quantified proteins was applied by considering only the proteins that were identified by at least two distinct labeled peptides from at least two of the three triplex (table1). This pre-filtration, yielded 226 quantified proteins.

	Triplex1	Triplex2	Triplex3	Total (Union)
Number of peptides used	**1014**	**1105**	**1104**	**1799**
Number of corresponding proteins	503	509	507	**776**
Number of selected peptides	640	715	696	997
Number of selected proteins	214	213	209	**226**

Table 1. Pre-filtration of the quantified proteins. Number of labeled peptides identified by LC-MS/MS in the three triplex that were used in the analysis. Number of selected peptides after pre-filtration of the quantified proteins i.e. proteins quantified by at least two distinct peptides and from at least two of the three triplex.

166 membrane proteins are differentially expressed in spermatocytes, spermatids and residual bodies

To assess which of the 226 selected proteins were significantly differentially expressed between pSPC, rSPT and RB, we used the strategy detailed in Figure 2A. A cutoff of 2 was applied to the median of the protein ratio per type of ratio (rSPT/pSPC, RB/pSPC and RB/rSPT) for the 226 quantified proteins. Thus, 167 proteins were considered to have a high expression variation in at least one sample comparison. Then, a LIMMA statistical test identified 166 proteins (Table S2) significantly differentially expressed (F-value adjusted with the False Discovery Rate, $p \leq 0.05$). The 166 differential proteins were classified into 5 expression patterns (P1-P5) using the PAM algorithm (Figure 2B). 52 proteins display an expression that gradually decreases from pSPC to RBs (P1); 23 proteins are highly over-expressed in rSPT (P2); 20 proteins present a low decrease of expression between the pSPC and the rSPT and are highly over-expressed in RBs (P3); 33 proteins present a gradual increase of expression from the pSPC to the RBs and are highly over-expressed in RBs (P4); and finally the expression of 38 proteins dramatically increases in RBs as compared to pSPC and rSPT (P5). The individual areas of each labeled peptide were retrieved from the (H/L), (SH/L) and (SH/H) ratios, in order to illustrate in an expression heatmap the relative expression in each biological sample of the differential proteins distributed in these 5 groups (Figure 2B). We thus obtained a membrane protein expression profiling in rat pSPC, rSPT germ cells and residual bodies.

Functional analysis of membrane proteins differentially expressed.

A functional analysis of the 5 groups of differential membrane proteins was performed. We assess the biological processes in which they are involved, and the cellular component to which they are associated using an Gene Ontology (GO) terms over-representation analysis (Figure 3).

Groups P1 and P3

The first group P1 (pSPC/rSPT) is only enriched in the biological process terms: energy derivation by oxydation of organic compounds related to cell respiration (n=9; $p<4x10^{-6}$) which is also enriched in the group P3 (pSPC /RB); mitochondrion (n=10; $p<3x10^{-9}$); and lipid particle (n=6; $p<5x10^{-5}$), also over-represented in P3: (n=12; $p<7x10^{-7}$) and (n=5; $p<4x10^{-5}$ respectively. The term protein complex is over represented in P1 (n=16; $p<3x10^{-6}$); and in P3 (n=13, $p<7x10^{-5}$). We observe that all the GO terms enriched in P1 are commonly enriched in P3.

Groups P2 and P4

In the group P2 corresponding to proteins preferentially expressed in rSPT, the reproduction process (n=12; $p<4x10^{-5}$) is enriched, as in groups P4 (n=13; $p<3x10^{-4}$). Microtubule cytoskeleton organization in P2 (n=5; $p<4x10^{-3}$) is over-represented like it is in P4 (n=6, $p<2x10^{-3}$). RNA biosynthetic process in P2

(n=9; p<10x10-^{3}); is over-represented, as well as mRNA metabolic process: P2 (n=9; p<3x10^{-7}), P4:(n=7, p<6x10^{-4}); ribosome biogenesis: P2 (n=6; p<2x10^{-5}), P4(n= 8; p<3x10^{-7}); and ribonucleoprotein complex: P2 (n10=; p<2x10^{-7}), P4 (n=10; p<3x10^{-6}). P4 has the greater number of terms related to transport over-represented compared to other groups: protein targeting (n=7; p<5x10^{-4}), also enriched in P2 (n=8; p<4x10^{-6}).

Groups P3 and P5

In the P3 group of proteins preferentially expressed in pSPC and RB, several terms related to cell cycle such as: negative regulation of ubiquitin protein ligase activity involved in mitotic cell cycle (n=3; p=0,002); DNA endoreduplication (n=3; p<2x10^{-3}) and mitotic DNA damage checkpoint (n=3, p<8x10^{-3}), are solely over-represented in this group. Terms related to metabolism are also over-represented in P3, such as: energy derivation by oxydation of organic compounds (n=5; p<2x10^{-3}). In P5 where proteins are preferentially expressed in residual bodies, the reproduction term is over-represented (n=16; p<6x10^{-5}).
Several terms such as Lipid particle, P3 (n=5; p<4x10^{-5}), P5 (n=3;p<4x10^{-4}); mitotic cell cycle: P3 (n=6, p<5x.10^{-3}), P5 (n=8; p<3x10^{-3}); and determination of adult lifespan P3 (n=6; p<5 x10^{-4}), P5 (n=6; p<3x10^{-3}) are common in P3 end P5, where proteins are over-expressed in residual bodies.

Group P5

Interestingly, spermatogenesis (n=7; p<2x10-^{3}), spermatid development (n=4; p=0,004) and binding of sperm to zona pellucida (n=4; p<6x10^{-5}) are specifically enriched in residual bodies (P5). Processes related to transport: receptor mediated endocytosis (n=6; p<3x10^{-3}), protein import (n=5; p<3x10^{-3}), protein targeting (n=7; p=0,002), SRP dependent cotranslational protein targeting to membrane (n=7, p<4x10^{-9}), RNA transport (n=4, p<4x10^{-3}), and macropinocytose (n=3; p<2x10^{-3}) are over-represented. Macromolecular complex assembly is also represented (n=9; p<5x10^{-3}), as well as cell body (n=6; p<4x10^{-3}) and cytoskeletal part (n=11; p<3x10^{-4}). Unexpected organelles in RBs such as acrosomal vesicle (n= 3; p<6x10^{-3}); cytoplasmic membrane bound vesicle (n=8; p<4x10^{-3}), and melanosome (n=4; p<4x10^{-4}) are over-represented. Protein complexes are also present: protein complex (n=15; p<4x10^{-3}), and interestingly, the chaperonin-containing T-complex (n= 5; p<3x10^{-10}), and the zona pellucida receptor complex (n=4; p<5x10^{-7}) are significantly together over-represented in the RBs, but not in other groups.

Selection of eight membrane candidate proteins in pSPC, rSPT, and RBs

Among the proteins differentially expressed in this study, eight membrane proteins in germ cells or RBs were selected for further experimental validation. These candidates represented differentially expressed proteins in pSPC, rSPT or RB, that were either:

1) Potentially involved in signal transduction suggesting a role in paracrine communication

2) Or involved in cell differentiation and not previously described in germ cells context
3) Or linked with biological processes involving phagocytosis or cell recognition, such as immunity.

Among the 166 differential proteins identified, those matching one or more of these criteria were selected using the GO annotations or the pathway annotations: KEGG pathway database; and REACTOME database with no restriction in term of expression pattern.

	UniProt	Protein description	Selection by biological process annotations or pathways (KEGG/REACTOME)	Gene Name	Pattern	PMID
1	G3V6N2	Transmembrane emp24 protein transport domain containing 4	Signal transduction (0007165);positive regulation of I-kappaB kinase/NF-kappaB cascade (0043123)	Tmed4	P1 (pSPC,rSPT).	23076522
2	Q5I0E7	Transmembrane emp24 domain-containing protein 9	Dorsal/ventral pattern formation (0009953)	Tmed9	P1 (pSPC,rSPT).	22114321
3	D3ZSA9	Nodal modulator	Negative regulation of nodal signaling / determination of lateral mesoderm left/right asymmetry (1900176)	Nomo1	P1 (pSPC/rSPT)	22832245
4	D4A899	Vps13a	Nervous system development (0007399)	Vps13a	P1 (pSPC/rSPT)	22366033
5	D3ZD31	Mannose receptor, C type 1;macrophage mannose receptor 1	Receptor-mediated endocytosis (0006898); signal transduction (0007165); Mannose binding (0005537); Immune System (REACT:6900); Phagosome (KEGG:04145);	Mrc1	P2 (rSPT)	24672807 24838383
6	D4A478	Nucleoside-triphosphatase, cancer-related;nucleoside-triphosphatase C1orf57 homolog	Imaginal disc-derived wing morphogenesis (0007476)	Ntpcr	P2 (rSPT)	-
7	Q4FZS5	Hypothetical LOC287798	Integral to membrane (0016021)	MGC95210	P4 (rSPT/RB)	
8	P62944	Ap2b1;LOC100912146	Antigen processing and presentation of exogenous peptide antigen via MHC class II (0019886)	Ap2b1	P5 (RB)	-

Table 3. Candidate proteins differentially expressed between pSPC, rSPT and RBs, potentially involved in signal transduction, development, phagocytosis or cell recognition.

Eight candidate proteins are identified: two TMED (Transmembrane emp24 domain-containing protein) proteins: TMED4 and TMED9 which are involved in immunity or signal transduction ; the Nodal modulator NOMO1 ; the macrophage receptor Mrc1 ; another protein involved in innate immune response Ap2b1; the proteins C1orf57 and Vps13a involved in differentiation, and the uncharacterized protein “hypothetical LOC287798”.

DISCUSSION

The aim of our study was to identify membrane proteins differentially expressed in pSPC, rSPT and in RB. We first showed that our experimental protocol allowed to extract membrane proteins, among which endomembrane system proteins could be present, including proteins from secretory granules; and from organelles. In this study, we confirm the differential expression between pSPC, rSPT and RBs for 43 proteins known to be expressed in the testis or known to be involved in reproduction. Among them, 27 proteins are known to have a crucial role in spermatogenesis or fertilization.

Expression pattern of known proteins

Proteins involved in post-transcriptional fate control of mRNAs, and thus in spermatogenesis are found such as PiwiL1 (Uniprot D3ZTP9), and Tdrkh proteins (Uniprot G3V8T7), which are over-expressed in P1, i.e. with an expression decreasing from pSPC to rSPT stages, before transcriptional machinery is completely shut down when the spermatid begins the nuclear condensation and elongation. The mRNA synthesis and storage are indeed necessary so that the elongating spermatid can translate mRNAs when the corresponding proteins are needed (Idler and Yan, 2012; Sassone-Corsi, 2002). Another example of protein described as crucial for spermatogenesis is the EH domain-containing protein 1 (EHD1, Uniprot Q641Z6), shown as highly over-expressed in RBs (P5). This protein has been predicted to participate in clathrin coated pit endocytosis of IGF1 receptor following ligand binding (Kirchhausen et al., 1997). Note that Clathrin, heavy chain 1 (Cltc, Uniprot F1M779) and clathrin adaptor AP-2 complex subunit beta (Uniprot P62944), are also identified in our study in the pattern rSPT/RBs (P4), and are known to be involved in the endocytosis of ligand-bound receptors (Chen et al., 1998). The protein EHD1, is associated with endocytic vesicles, and is strongly expressed in the testis (Mintz et al., 1999). Interestingly, Ehd1-/- mice show spermatogenesis defects, including fusion of spermatids heads and tails, and spermatids failing to form cytoplasmic lobes and residual bodies at later stages. Spermatids at stage VIII are phagocyted by Sertoli cells at apical ectoplasmic specialization in these Ehd1 K.O. mice. EHD1-dependent endocytic recycling and trafficking may be required for spermiation in mice (Rainey et al., 2010). In this study, we show the over-expression of EHD1 in RBs compared to meiotic and post-meiotic germ cells.

Many examples of proteins involved in sperm migration and fertilization are found. Angiotensin converting enzyme (ACE, Uniprot: P47820) is expressed in rSPTor RBs (group P4). ACE provokes the shedding of GPI anchored proteins involved in sperm migration (GPI-anchored protein complex LY6K/TEX101), and of ADAM3 (Fujihara et al., 2014). The disintegrin and metalloproteinase ADAM6 (G3V9N4) that we identified as over-expressed in rSPT and RBs (group P4), belongs to the first group of ADAMs expressed in the male reproductive tissues, which expression is lost in Tex101 KO mice (Li et al., 2013). ADAM6 forms a complex with ADAM2 and ADAM3 that is required for fertilization in mouse.(Han et al., 2009). The protein SSP411 (Uniprot FILSR7), is a potentially secreted protein associated with spermatogenesis, and is a potential testis-specific thioredoxin with a suggested role in sperm maturation, fertilization, and/or embryo development (Shi et al., 2004). We show SSP411 over-expression in RBs (group P5). We confirm the expression pattern for other proteins such as Dickkopf-like 1 protein (DKKL1, Uniprot D4A444), for which we show a preferential expression in pSPC and rSPT (group P1), as previously observed in human (Yan et al., 2012). Although DKKL1 is also described as an acrosomal protein facilitating fertilization (Kohn et al., 2005, 2010), it does not seem to be accumulated in the RBs. In summary, our study reinforce previous observations concerning 27 proteins known to be involved in spermatogenesis, using a quantitative proteomic analysis on germ cells, and residual bodies membrane proteins, performed for the first time.

Eight candidate proteins potentially involved in communications between Sertoli and germ cells

Eight membrane proteins emerged from this work as potential interesting candidate for further experimental validations. Among them, the TMED4 and TMED9 are identified, and these family of proteins seem to have an important function in germ cells. The gene*Tmed9* is a predicted target of miR-296 (a paternally imprinted miRNA), in rat and primates (Robson et al., 2012). NOMO, also a target of paternally imprinted miRNA (mir675) in human trophoblast (Gao et al., 2012). The differential expression of TMED4 or "transmembrane emp24 protein transport domain containing 4" (Uniprot: G3V6N2), and TMED9 or "Transmembrane emp24 domain-containing protein 9" (Uniprot Q5I0E7) is shown in the group P1 (pSPC and rSPT). The proteins of the TMED family have a GOLD domain of category 1 (emp24 family of proteins), conserved in eukaryotes which is related to secretion or protein sorting and thought to have a role in protein/protein interactions, helping the assembly of protein complexes on membranes (Anantharaman and Aravind, 2002). Importantly, a member of this family, TMED7, is shown to prevent TLR4 (Toll Like receptor 4) signalling upon LPS stimulation. The TMED proteins are conserved and seem to have a conserved role in the regulation of innate immunity (Doyle et al., 2012). This putative regulatory role of TMEDs in the TLRs signalling following apoptotic germ cells phagocytosis by Sertoli cells might be important to explain why germ cells phagocytosis does not trigger IL-1 alpha expression, whereas RBs phagocytosis does trigger IL-1expression and the subsequent events (IL-1 alpha expression). The TMED candidates gives some clues concerning the synchrony of the rat seminiferous epithelium ruled by the RBs phagocytosis, and subsequent production of IL-1/IL-1 alpha by Sertoli cells, leading to IL-6 production (Syed et al., 1995); both cytokines acting on the BTB dynamics (Lie et al., 2011; Zhang et al., 2014). It is known that IL-1 alpha production by Sertoli cells is a major regulator of spermatogenesis and synchrony of the spermatogenetic wave. It has also been demonstrated that IL-1alpha production by Sertoli cells is dependent upon the phagocytosis of RBs by these cells, while apoptotic pSPC and rSPT phagocytosis by Sertoli cells have no effect on IL-1alpha production (Gérard et al., 1992).
Indeed TMED proteins (preferentially expressed in pSPC and rSPT) may be able to inhibit TLR signalling *via* TLR degradation in Sertoli cells after internalization of apoptotic bodies from germ cells, and thus prevent NF-KB signalling, thereby blocking the transcription of interleukins IL-1 alpha and other inflammatory cytokine genes by Sertoli cells. On the opposite, the TMED4 and 9 proteins expression are highly decreased in RBs, suggesting that the TLR signaling upon RB phagocytosis by Sertoli cells could normally trigger interleukin production, thus triggering a new spermatogenic wave in the seminiferous epithelium.

Differential germ cell membrane proteins potentially interacting with Sertoli proteins

A network analysis between proteins from Sertoli cells extracts obtained from another experiment (not

shown), between Sertoli cell proteins and P1, P2, P3, P4, P5 differential proteins was assessed. The complexity of the network does not allow it to be displayed, but the network shows the protein/protein interaction of GORASP2 (in Sertoli cells), and TMED9 present in the P1 group (pSPC/rSPT). GORASP2 also interacts with TMED10, TMED9 and TMED7, which are present in Sertoli cell extracts. Importantly, GORASP2 is one of the Golgi reassembly stacking protein (GRASP) paralogues, required for autophagy driven IL1-beta secretion in mammalian cells (Dupont et al., 2011). TMED9 also interacts with the protein Lmna1 present in Sertoli cells.

The differential expression of the TMED 4 and TMED9 in pattern P1 allows to propose the hypothesis that IL-1 alpha production by Sertoli cells triggered by phagocytosis of RB could not be due to a specific recognition signals at the RB surface, but rather to the absence of some TMED proteins that could play a role in an inhibition of the TLRs, responsible of IL-1 alpha production or inhibition of GORASP required for IL1-beta secretion. This hypothesis could explain why latex beads and LPS also promote IL-1alpha (and iL-6) production by Sertoli cells (Cudicini et al., 1997; Gérard et al., 1992), as do RBs, through the TLR signaling. A further experimental confirmation of the differential expression of the TMED4 and TMED9 proteins by western blot in isolated germ cells and in residual bodies is necessary.

Germ cell differential predicted membrane proteins

Interestingly, predicted membrane proteins differentially expressed in meiotic and post meiotic germ cells are also identified in this study. Thus, 23 proteins translated from GENSCAN gene predictions and identified in the five expression patterns constitute potential candidates for further investigation of their functions in spermatogenesis.

Germ cell differential uncharacterized membrane proteins

Four uncharacterized/hypothetical proteins: Q4FZS5, D3ZLL8, D4A0P4, Q6AY52, have also differential expression patterns in pSPC, rSPT and RBs, and could deserve investigation for their functions in spermatogenesis.

Many more candidate proteins could have been found given the complexity of the cell types that we study, but the drawbacks of this technology hinder the exhaustivity of our analysis. First, not all the tryptic peptides may carry an ICPL label. Second, only the quantified peptides that were labeled in the three samples, (i.e. for which the three ratios SH/H, H/L, SH/L) are identified in the ICPL study. It means that with this approach we miss proteins that are expressed in some samples and absent in others.

Conclusion

Using an ICPL approach, we found 166 proteins with a differential expression in the pSPC, rSPT and RBs, potentially involved in germ cell/ Sertoli cell communication and differentiation events occurring during spermiogenesis. Our results are consistent with previous observations on some membrane proteins known to be involved in spermatogenesis and fertilization as we confirmed the differential expression for 43 proteins already known to be expressed in the testis or involved in reproduction, and among them, 27 involved in spermatogenesis or fertilization. For those that were only evidenced at a transcript level, or predicted, this study confirms their differential expression by mass spectrometry at a protein level. We also report the differential expression of 23 proteins coming from GENSCAN predictions, available on the rat genome. Yet uncharacterized differential proteins at the meiotic and post meiotic stages of spermatogenesis which were not selected as candidates could also deserve further investigation.

The study of Rolland et al., 2007 (Rolland et al., 2007) had a similar goal for cytosolic proteins from spermatogonia, spermatocytes and spermatids, and allowed the identification of proteins for which the expression and the activity had been further assessed. The membrane proteins that we identify as differentially expressed at the meiotic and post meiotic stages come to complete the previous work from our laboratory.

References

Anantharaman, V., and Aravind, L. (2002). The GOLD domain, a novel protein module involved in Golgi function and secretion. Genome Biol. *3*, research0023.

Belleannee, C., Belghazi, M., Labas, V., Teixeira-Gomes, A.-P., Gatti, J.L., Dacheux, J.-L., and Dacheux, F. (2011). Purification and identification of sperm surface proteins and changes during epididymal maturation. Proteomics *11*, 1952–1964.

Bettegowda, A., and Wilkinson, M.F. (2010). Transcription and post-transcriptional regulation of spermatogenesis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *365*, 1637–1651.

Chalmel, F., and Primig, M. (2008). The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. BMC Bioinformatics *9*, 86.

Chalmel, F., Rolland, A.D., Niederhauser-Wiederkehr, C., Chung, S.S.W., Demougin, P., Gattiker, A., Moore, J., Patard, J.-J., Wolgemuth, D.J., Jégou, B., et al. (2007a). The conserved transcriptome in human and rodent male gametogenesis. Proc. Natl. Acad. Sci. U. S. A. *104*, 8346–8351.

Chalmel, F., Lardenois, A., and Primig, M. (2007b). Toward Understanding the Core Meiotic Transcriptome in Mammals and Its Implications for Somatic Cancer. Ann. N. Y. Acad. Sci. *1120*, 1–15.

Chalmel, F., Lardenois, A., Evrard, B., Mathieu, R., Feig, C., Demougin, P., Gattiker, A., Schulze, W., Jégou, B., Kirchhoff, C., et al. (2012). Global human tissue profiling and protein network analysis reveals distinct levels of transcriptional germline-specificity and identifies target genes for male infertility. Hum. Reprod. *27*, 3233–3248.

Chalmel, F., Lardenois, A., Evrard, B., Rolland, A.D., Sallou, O., Dumargne, M.-C., Coiffec, I., Collin, O., Primig, M., and Jégou, B. (2014). High-Resolution Profiling of Novel Transcribed Regions During Rat Spermatogenesis. Biol. Reprod. biolreprod.114.118166.

Chen, H., Fre, S., Slepnev, V.I., Capua, M.R., Takei, K., Butler, M.H., Di Fiore, P.P., and De Camilli, P. (1998). Epsin is an EH-domain-binding protein implicated in clathrin-mediated endocytosis. Nature *394*, 793–797.

Com, E., Evrard, B., Roepstorff, P., Aubry, F., and Pineau, C. (2003a). New insights into the rat spermatogonial proteome: identification of 156 additional proteins. Mol. Cell. Proteomics MCP *2*, 248–261.

Com, E., Evrard, B., Roepstorff, P., Aubry, F., and Pineau, C. (2003b). New Insights into the Rat Spermatogonial Proteome. Mol. Cell. Proteomics *2*, 248–261.

Cooper, T.G. (1998). Interactions between epididymal secretions and spermatozoa. J. Reprod. Fertil. Suppl. *53*, 119–136.

Cudicini, C., Lejeune, H., Gomez, E., Bosmans, E., Ballet, F., Saez, J., and Jégou, B. (1997). Human Leydig cells and Sertoli cells are producers of interleukins-1 and -6. J. Clin. Endocrinol. Metab. *82*, 1426–1433.

Dacheux, J.L., Druart, X., Fouchecourt, S., Syntin, P., Gatti, J.L., Okamura, N., and Dacheux, F. (1998). Role of epididymal secretory proteins in sperm maturation with particular reference to the boar. J. Reprod. Fertil. Suppl. *53*, 99–107.

Djureinovic, D., Fagerberg, L., Hallström, B., Danielsson, A., Lindskog, C., Uhlén, M., and Pontén, F. (2014). The human testis-specific proteome defined by transcriptomics and antibody-based profiling. Mol. Hum. Reprod.

Doyle, S.L., Husebye, H., Connolly, D.J., Espevik, T., O'Neill, L.A.J., and McGettrick, A.F. (2012). The GOLD domain-containing protein TMED7 inhibits TLR4 signalling from the endosome upon LPS stimulation. Nat. Commun. *3*, 707.

Dupont, N., Jiang, S., Pilli, M., Ornatowski, W., Bhattacharya, D., and Deretic, V. (2011). Autophagy-based unconventional secretory pathway for extracellular delivery of IL-1β. EMBO J. *30*, 4701–4711.

Eddy, E.M. (2002). Male germ cell gene expression. Recent Prog. Horm. Res. *57*, 103–128.

Fujihara, Y., Okabe, M., and Ikawa, M. (2014). GPI-Anchored Protein Complex, LY6K/TEX101, Is Required for Sperm Migration into the Oviduct and Male Fertility in Mice. Biol. Reprod. biolreprod.113.112888.

Gan, H., Cai, T., Lin, X., Wu, Y., Wang, X., Yang, F., and Han, C. (2013). Integrative proteomic and transcriptomic analyses reveal multiple post-transcriptional regulatory mechanisms of mouse spermatogenesis. Mol. Cell. Proteomics.

Gao, W.-L., Liu, M., Yang, Y., Yang, H., Liao, Q., Bai, Y., Li, Y.-X., Li, D., Peng, C., and Wang, Y.-L. (2012). The imprinted H19 gene regulates human placental trophoblast cell proliferation via encoding miR-675 that targets Nodal Modulator 1 (NOMO1). RNA Biol. *9*, 1002–1010.

Gérard, N., Syed, V., and Jégou, B. (1992). Lipopolysaccharide, latex beads and residual bodies are potent activators of Sertoli cell interleukin-1α production. Biochem. Biophys. Res. Commun. *185*, 154–161.

Griswold, M.D. (1995). Interactions between germ cells and Sertoli cells in the testis. Biol. Reprod. *52*, 211–216.

Griswold, M.D. (1998). The central role of Sertoli cells in spermatogenesis. Semin. Cell Dev. Biol. *9*, 411–416.

Guillaume, E., Dupaix, A., Moertz, E., Courtens, J.-L., Jégou, B., and Pineau, C. (2000). Proteome analysis of spermatogonia: identification of a first set of 53 proteins. Proteome *1*, 1–20.

Guo, X., Zhang, P., Huo, R., Zhou, Z., and Sha, J. (2008). Analysis of the human testis proteome by mass spectrometry and bioinformatics. Proteomics Clin. Appl. *2*, 1651–1657.

Han, C., Choi, E., Park, I., Lee, B., Jin, S., Kim, D.H., Nishimura, H., and Cho, C. (2009). Comprehensive Analysis of Reproductive ADAMs: Relationship of ADAM4 and ADAM6 with an ADAM Complex Required for Fertilization in Mice. Biol. Reprod. *80*, 1001–1008.

Huang, S.-Y., Lin, J.-H., Chen, Y.-H., Chuang, C., Lin, E.-C., Huang, M.-C., Sunny Sun, H.-F., and Lee, W.-C. (2005). A reference map and identification of porcine testis proteins using 2-DE and MS. Proteomics *5*, 4205–4212.

Idler, R.K., and Yan, W. (2012). Control of Messenger RNA Fate by RNA-Binding Proteins: An Emphasis on Mammalian Spermatogenesis. J. Androl. *33*.

Jégou, B. (1993). The Sertoli-germ cell communication network in mammals. Int. Rev. Cytol. *147*, 25–96.

Jégou, B. (1995). La cellule de Sertoli: actualisation du concept de cellule nourricière. Médecine/sciences *11*, 519.

Kirchhausen, T., Bonifacino, J.S., and Riezman, H. (1997). Linking cargo to vesicle formation: receptor tail interactions with coat proteins. Curr. Opin. Cell Biol. *9*, 488–495.

Kohn, M.J., Kaneko, K.J., and DePamphilis, M.L. (2005). DkkL1 (Soggy), a Dickkopf family member, localizes to the acrosome during mammalian spermatogenesis. Mol. Reprod. Dev. *71*, 516–522.

Kohn, M.J., Sztein, J., Yagi, R., DePamphilis, M.L., and Kaneko, K.J. (2010). The acrosomal protein Dickkopf-like 1 (DKKL1) facilitates sperm penetration of the zona pellucida. Fertil. Steril. *93*, 1533–1537.

Laiho, A., Kotaja, N., Gyenesei, A., and Sironen, A. (2013). Transcriptome profiling of the murine testis during the first wave of spermatogenesis. PloS One *8*, e61558.

Li, W., Guo, X.-J., Teng, F., Hou, X.-J., Lv, Z., Zhou, S.-Y., Bi, Y., Wan, H.-F., Feng, C.-J., Yuan, Y., et al. (2013). Tex101 is essential for male fertility by affecting sperm migration into the oviduct in mice. J. Mol. Cell Biol. *5*, 345–347.

Lie, P.P.Y., Cheng, C.Y., and Mruk, D.D. (2011). Interleukin-1alpha is a regulator of the blood-testis barrier. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *25*, 1244–1253.

Matzuk, M.M., and Lamb, D.J. (2002). Genetic dissection of mammalian fertility pathways. Nat. Cell Biol. *4 Suppl*, s41–49.

Mintz, L., Galperin, E., Pasmanik-Chor, M., Tulzinsky, S., Bromberg, Y., Kozak, C.A., Joyner, A., Fein, A., and Horowitz, M. (1999). EHD1--an EH-domain-containing protein with a specific expression pattern. Genomics *59*, 66–76.

Mori, E., Fukuda, H., Imajoh-Ohmi, S., Mori, T., and Takasaki, S. (2012). Purification of N-acetyllactosamine-binding activity from the porcine sperm membrane: possible involvement of an ADAM complex in the carbohydrate-binding activity of sperm. J. Reprod. Dev. *58*, 117–125.

Pineau, C., Syed, V., Bardin, C.W., Jégou, B., and Cheng, C.Y. (1993). Germ cell-conditioned medium contains multiple factors that modulate the secretion of testins, clusterin, and transferrin by Sertoli cells. J. Androl. *14*, 87–98.

Rainey, M.A., George, M., Ying, G., Akakura, R., Burgess, D.J., Siefker, E., Bargar, T., Doglio, L., Crawford, S.E., Todd, G.L., et al. (2010). The endocytic recycling regulator EHD1 is essential for spermatogenesis and male fertility in mice. BMC Dev. Biol. *10*, 37.

Robson, J.E., Eaton, S.A., Underhill, P., Williams, D., and Peters, J. (2012). MicroRNAs 296 and 298 are imprinted and part of the GNAS/Gnas cluster and miR-296 targets IKBKE and Tmed9. RNA *18*, 135–144.

Rolland, A.D., Evrard, B., Guitton, N., Lavigne, R., Calvel, P., Couvet, M., Jégou, B., and Pineau, C. (2007). Two-dimensional fluorescence difference gel electrophoresis analysis of spermatogenesis in the rat. J. Proteome Res. *6*, 683–697.

Sassone-Corsi, P. (2002). Unique chromatin remodeling and transcriptional regulation in spermatogenesis. Science *296*, 2176–2178.

Schlecht, U., Demougin, P., Koch, R., Hermida, L., Wiederkehr, C., Descombes, P., Pineau, C., Jégou, B., and Primig, M. (2004). Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. Mol. Biol. Cell *15*, 1031–1043.

Schultz, N., Hamra, F.K., and Garbers, D.L. (2003). A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. Proc. Natl. Acad. Sci. U. S. A. *100*, 12201–12206.

Shi, H.-J., Wu, A.Z., Santos, M., Feng, Z.-M., Huang, L., Chen, Y.-M., Zhu, K., and Chen, C.-L.C. (2004). Cloning and characterization of rat spermatid protein SSP411: a thioredoxin-like protein. J. Androl. *25*, 479–493.

Smyth, G.K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol. *3*, Article3.

Son, C.G., Bilke, S., Davis, S., Greer, B.T., Wei, J.S., Whiteford, C.C., Chen, Q.-R., Cenacchi, N., and Khan, J. (2005). Database of mRNA gene expression profiles of multiple human organs. Genome Res. *15*, 443–450.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Rep. *3*, 2179–2190.

Syed, V., Stéphan, J.P., Gérard, N., Legrand, A., Parvinen, M., Bardin, C.W., and Jégou, B. (1995). Residual bodies activate Sertoli cell interleukin-1 alpha (IL-1 alpha) release, which triggers IL-6 production by an autocrine mechanism, through the lipoxygenase pathway. Endocrinology *136*, 3070–3078.

Wang, J., Xia, Y., Wang, G., Zhou, T., Guo, Y., Zhang, C., An, X., Sun, Y., Guo, X., Zhou, Z., et al. (2014). In-depth proteomic analysis of whole testis tissue from the adult rhesus macaque. Proteomics.

Wrobel, G., and Primig, M. (2005). Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. Reproduction *129*, 1–7.

Yan, Q., Wu, X., Chen, C., Diao, R., Lai, Y., Huang, J., Chen, J., Yu, Z., Gui, Y., Tang, A., et al. (2012). Developmental expression and function of DKKL1/Dkkl1 in humans and mice. Reprod. Biol. Endocrinol. RBE *10*, 51.

Zhang, H., Yin, Y., Wang, G., Liu, Z., Liu, L., and Sun, F. (2014). Interleukin-6 disrupts blood-testis barrier through inhibiting protein degradation or activating phosphorylated ERK in Sertoli cells. Sci. Rep. *4*, 4260.

Zheng, B., Zhou, Q., Guo, Y., Shao, B., Zhou, T., Wang, L., Zhou, Z., Sha, J., Guo, X., and Huang, X. (2014). Establishment of a proteomic profile associated with gonocyte and spermatogonial stem cell maturation and differentiation in neonatal mice. Proteomics *14*, 274–285.

Zhu, Y.-F., Cui, Y.-G., Guo, X.-J., Wang, L., Bi, Y., Hu, Y.-Q., Zhao, X., Liu, Q., Huo, R., Lin, M., et al. (2006). Proteomic analysis of effect of hyperthermia on spermatogenesis in adult male mice. J. Proteome Res. *5*, 2217–2225.

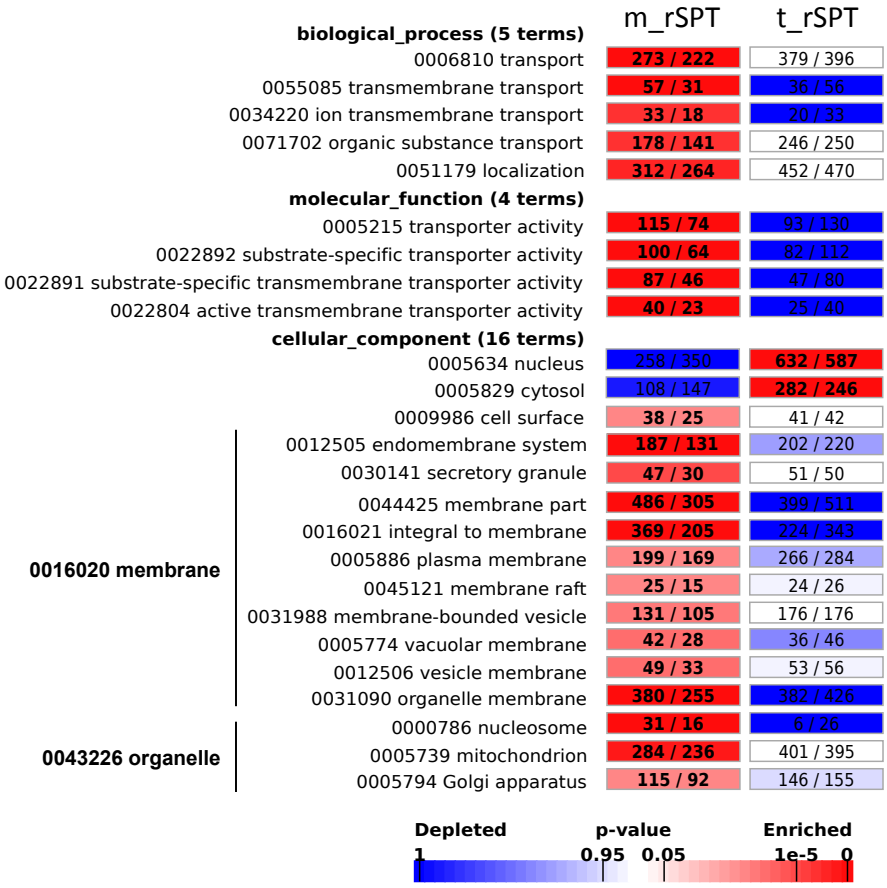

Figure 1: Assessment of rSPT membrane enrichment. Significantly enriched GeneOntology (GO) terms, among the identified proteins (GeneID) from the membrane fraction (m_rSPT) as compared to the total rSPT protein extracts (t_rSPT). The total numbers of proteins (NCBI entrez gene identifiers) are given within rectangles as observed (on the left), and as expected by chance (on the right). A color scale illustrating p-values is displayed, for enriched terms in red and depleted terms in blue. Numbers in bold indicate a statistically significant over-representation for a given GO term.

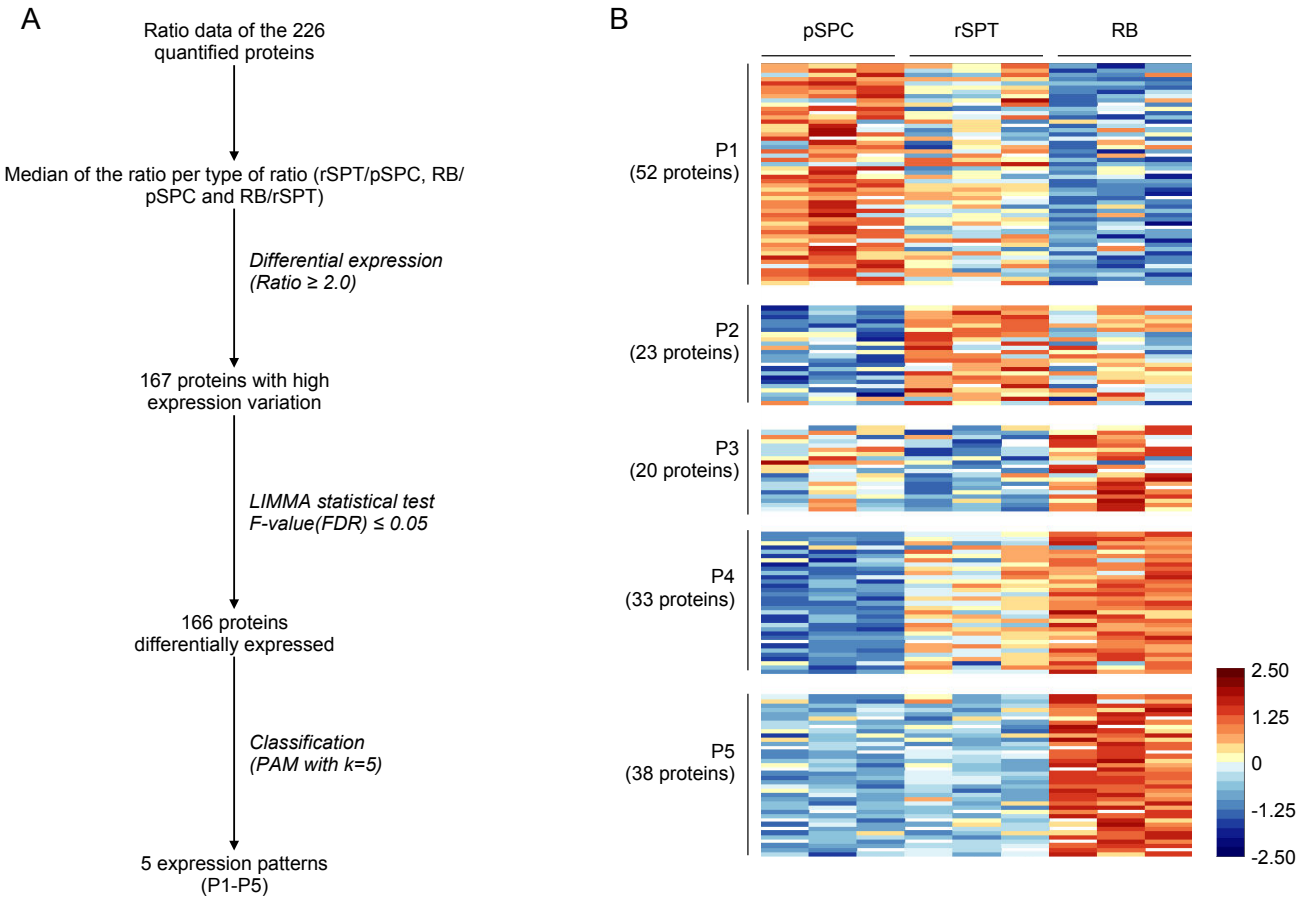

Figure 2: Selection and classification of the differential proteins from pSPC, rSPT and RB membrane fractions

A, Flowchart summarizing the filtration steps to select significantly differentially expressed proteins identified by LC MS/MS and quantified by ICPL, in rat pachytene spermatocytes (pSPC), round spermatids (rSPT), and residual bodies (RB); and their classification. The number of selected proteins is given at each filtration step. B, False-color heatmap that summarizes the five expression patterns of the 166 differential proteins, defined according to their relative expression level in pachytene spermatocyte (pSPC), round spermatids (rSPT) and residual bodies (RBs). Cell types and residual bodies are represented by columns subdivided in three columns corresponding a technical triplex. Each line corresponds to a protein. For each expression pattern, P1 to P5 the number of differential proteins is indicated. The standardized log2-transformed areas are displayed according to a color scale ranging from -2.50 (blue) to 2.50 (red) .

	biological_process	P1	P2	P3	P4	P5
Reproduction and Development	0000003 reproduction	9 / 5	12 / 3	8 / 3	13 / 4	16 / 4
	0008340 determination of adult lifespan	2 / 1	2 / 0	6 / 0	4 / 1	6 / 1
	0009792 embryo development ending in birth or egg hatching	8 / 3	7 / 2	7 / 2	10 / 2	9 / 3
	0007283 spermatogenesis	4 / 1	4 / 1	1 / 1	3 / 1	7 / 1
	0007286 spermatid development	2 / 0	0 / 0	1 / 0	1 / 0	4 / 0
	0007339 binding of sperm to zona pellucida	0 / 0	1 / 0	2 / 0	1 / 0	4 / 0
Transport	0006605 protein targeting	3 / 1	8 / 1	1 / 1	7 / 1	7 / 1
	0006614 SRP-dependent cotranslational protein targeting to membrane	2 / 0	8 / 0	0 / 0	7 / 0	0 / 0
	0050658 RNA transport	1 / 0	0 / 0	0 / 0	4 / 0	1 / 0
	0044351 macropinocytosis	2 / 0	0 / 0	0 / 0	3 / 0	1 / 0
	0017038 protein import	0 / 1	2 / 0	0 / 0	0 / 0	5 / 0
	0006898 receptor-mediated endocytosis	3 / 1	2 / 0	3 / 0	1 / 1	6 / 1
Cell cycle	0051436 negative regulation of ubiquitin-protein ligase activity ...	0 / 0	0 / 0	3 / 0	0 / 0	1 / 0
	0042023 DNA endoreduplication	0 / 0	0 / 0	3 / 0	0 / 0	0 / 0
	0044773 mitotic DNA damage checkpoint	0 / 0	0 / 0	3 / 0	0 / 0	1 / 0
	0000278 mitotic cell cycle	3 / 2	5 / 1	6 / 1	7 / 1	8 / 1
	0022402 cell cycle process	5 / 3	6 / 1	6 / 1	7 / 2	10 / 2
Metabolism	0015980 energy derivation by oxidation of organic compounds	9 / 1	0 / 0	5 / 0	1 / 0	1 / 1
	0010467 gene expression	4 / 6	11 / 3	4 / 3	10 / 4	3 / 5
	0006629 lipid metabolic process	1 / 2	1 / 1	7 / 1	1 / 2	6 / 2
	0005975 carbohydrate metabolic process	2 / 2	1 / 1	6 / 1	4 / 1	11 / 2
	0071704 organic substance metabolic process	16 / 16	15 / 8	11 / 8	16 / 11	23 / 13
Other	0032774 RNA biosynthetic process	2 / 4	9 / 2	1 / 2	7 / 3	0 / 3
	0006996 organelle organization	8 / 6	10 / 3	7 / 3	9 / 4	11 / 5
	0016071 mRNA metabolic process	0 / 1	9 / 1	2 / 1	7 / 1	2 / 1
	0042254 ribosome biogenesis	0 / 1	6 / 0	1 / 0	8 / 0	1 / 1
	0000226 microtubule cytoskeleton organization	3 / 1	5 / 1	4 / 1	6 / 1	6 / 1
	0065003 macromolecular complex assembly	3 / 3	6 / 1	7 / 1	4 / 2	9 / 2
	0012501 programmed cell death	3 / 3	3 / 1	8 / 1	8 / 2	8 / 2
	0051788 response to misfolded protein	1 / 0	0 / 0	3 / 0	0 / 0	1 / 0
	0010038 response to metal ion	1 / 1	1 / 1	3 / 1	2 / 1	6 / 1
	0006457 protein folding	0 / 0	0 / 0	2 / 0	0 / 0	11 / 0
	cellular_component					
Protein complex	0043234 protein complex	16 / 7	4 / 4	13 / 4	6 / 5	15 / 6
	0005832 chaperonin-containing T-complex	0 / 0	0 / 0	2 / 0	0 / 0	5 / 0
	0002199 zona pellucida receptor complex	0 / 0	0 / 0	2 / 0	1 / 0	4 / 0
Organelle and cell part	0005739 mitochondrion	19 / 4	9 / 2	12 / 2	8 / 2	6 / 3
	0005811 lipid particle	6 / 0	0 / 0	5 / 0	3 / 0	5 / 0
	0005634 nucleus	10 / 11	13 / 6	8 / 5	10 / 7	13 / 9
	0030529 ribonucleoprotein complex	4 / 2	10 / 1	0 / 1	10 / 1	3 / 1
	0005777 peroxisome	0 / 0	1 / 0	3 / 0	0 / 0	1 / 0
	0005794 Golgi apparatus	5 / 3	3 / 1	1 / 1	8 / 2	1 / 2
	0005773 vacuole	3 / 1	2 / 1	2 / 1	9 / 1	2 / 1
	0016023 cytoplasmic membrane-bounded vesicle	4 / 2	0 / 1	1 / 1	6 / 1	8 / 2
	0001669 acrosomal vesicle	0 / 0	0 / 0	0 / 0	2 / 0	3 / 0
	0042470 melanosome	0 / 0	0 / 0	1 / 0	1 / 0	4 / 0
	0044430 cytoskeletal part	5 / 3	0 / 2	7 / 1	4 / 2	11 / 2
	0044297 cell body	1 / 1	0 / 1	2 / 1	1 / 1	6 / 1

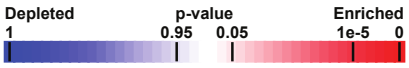

Figure 3: Functional analysis of differentially expressed proteins between pSPC, rSPT and RB membrane fractions. Significantly over-represented GeneOntology (GO) terms, associated with the 166

differential proteins from the membrane fractions of pSPC, rSPT and RBs distributed in five patterns, P1 to P5 (Figure 2B); as compared with those of the whole rat proteome are illustrated. The total numbers of proteins are given within rectangles as observed (on the left), and as expected by chance (on the right). A color scale illustrating p-values is displayed, for over-represented terms in red and depleted terms in blue. Numbers in bold indicate an over-representation for a given GO term.

Supplemental tables

Table S1. Inverted labeling of pachytene spermarocyte, round spermatids and residual bodies, using the light (ICPL_0), heavy (ICPL_6) and super heavy (ICPL_10) ICPL reagents.

	pSPC	**rSPT**	**RBs**
Triplex 1	ICPL_0	ICPL_6	ICPL_10
Triplex 2	ICPL_6	ICPL_10	ICPL_0
Triplex 3	ICPL_10	ICPL_0	ICPL_6

Table S2. List of the 166 differential proteins, defined according to their relative expression level in pachytene spermatocyte (pSPC), round spermatids (rSPT) and residual bodies (RBs). The UniProt identifier (ID) is given, the gene name and the description as well as the expression patterns P1 to P5 are indicated. The log2-transformed values are given for each type of ratio: SPTvsSPC, RBvsSPT, RBvsSPC are displayed.

ID	Gene Name	Description	Group	SPTvsSPC	RBvsSPT	RBvsSPC
D3ZF12	Spcs3	signal peptidase complex subunit 3 homolog (S. cerevisiae);signal peptidase complex subunit 3	P1	-1.55	-0.77	-1.35
H0VNW7	PCBP1	Uncharacterized protein	P1	-0.46	-2.14	-2.14
D3ZSA9	Nomo1	nodal modulator 1	P1	-0.62	-1.93	-1.61
D3ZTP9	Piwil1	piwi-like RNA-mediated gene silencing 1;piwi like homolog 1;piwi-like 1;piwi-like protein 1	P1	-0.49	-1.96	-1.65
D3ZUX5	Chchd3	coiled-coil-helix-coiled-coil-helix domain containing 3;coiled-coil-helix-coiled-coil-helix domain-containing protein 3, mitochondrial	P1	-0.88	-1.15	-1.21
Q62991	Scfd1	Sec1 family domain-containing protein 1	P1	-1.00	-1.02	-1.17
D4A0P4	LOC691496	similar to histone 1, H2ai;uncharacterized protein LOC691496	P1	1.50	-1.32	-0.28
D4A0T0	Ndufb10	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 10;NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 10	P1	-2.61	-0.49	-1.32
D4A3P0	Ybx2	Y box binding protein 2;Y box protein 2;Y-box-binding protein 2	P1	-0.61	-1.69	-1.66
D4A444	Dkkl1	dickkopf-like 1;dickkopf-like protein 1	P1	0.29	-2.37	-2.09
D4A7N1	Chchd6	coiled-coil-helix-coiled-coil-helix domain containing 6;coiled-coil-helix-coiled-coil-helix domain-containing protein 6, mitochondrial	P1	-0.58	-1.36	-1.40
D4AAT2	RGD1306782	similar to RIKEN cDNA 1700029P11	P1	-1.97	0.04	-0.88
Q6P6U3	Lmna	Lmna protein	P1	-4.59	-0.77	-3.53
ENSRNOP00000026949	-	-	P1	0.66	-3.53	-3.02
ENSRNOP00000056733	-	-	P1	0.59	-3.19	-3.03
F1LP82	-	-	P1	-1.19	0.12	-0.56
F1LXA0	Ndufa12	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 12	P1	-1.08	-1.43	-1.50
Q3KRE0	Atad3	ATPase family AAA domain-containing protein 3	P1	-1.35	-0.28	-0.81
G3V6N2	Tmed4	transmembrane emp24 protein transport domain containing 4;transmembrane emp24 domain-containing protein 4	P1	-1.64	-0.86	-1.45
G3V7E6	Pcm1	pericentriolar material 1	P1	1.07	-2.86	-1.93
G3V7U4	Lmnb1	lamin B1;lamin-B1	P1	-2.46	-1.53	-2.08
G3V7Y3	Atp5d	ATP synthase, H+ transporting, mitochondrial F1 complex, delta subunit;ATP synthase subunit delta, mitochondrial;F-ATPase delta subunit	P1	-0.09	-1.76	-1.55
G3V8T7	Tdrkh	tudor and KH domain containing	P1	-1.97	-1.93	-2.37

GENSCAN00000005489	-	-	P1	-0.31	-1.27	-1.19
GENSCAN00000010132	-	-	P1	-2.09	0.11	-0.81
GENSCAN00000015481	-	-	P1	-1.13	-0.74	-1.07
GENSCAN00000018257	-	-	P1	-0.10	-1.19	-1.06
GENSCAN00000021810	-	-	P1	-1.22	-1.97	-2.02
GENSCAN00000024730	-	-	P1	-0.24	-1.06	-0.89
GENSCAN00000024814	-	-	P1	0.14	-2.22	-1.91
GENSCAN00000031571	-	-	P1	-1.92	-1.52	-1.95
GENSCAN00000032509	-	-	P1	-1.24	-1.06	-1.46
GENSCAN00000043299	-	-	P1	-1.43	-0.49	-0.93
GENSCAN00000044966	-	-	P1	-2.36	-1.59	-2.36
P10888	Cox4i1	cytochrome c oxidase subunit IV isoform 1;COX IV-1;cytochrome c oxidase polypeptide IV;cytochrome c oxidase subunit 4 isoform 1, mitochondrial;cytochrome c oxidase, subunit 4a;cytochrome c oxidase, subunit IV;cytochrome c oxidase, subunit IVa	P1	-1.45	-0.12	-0.79
P11951	Cox6c	cytochrome c oxidase, subunit VIc;cytochrome c oxidase polypeptide VIc-2;cytochrome c oxidase subunit 6C-2;cytochrome oxidase subunit VIc	P1	-1.09	-1.11	-1.43
P32551	Uqcrc2	ubiquinol cytochrome c reductase core protein 2;complex III subunit 2;core protein II;cytochrome b-c1 complex subunit 2, mitochondrial;ubiquinol-cytochrome C reductase complex core protein 2, mitochondrial precursor (Complex III subunit II);ubiquinol-cytochrome c reductase core protein II;ubiquinol-cytochrome-c reductase complex core protein 2	P1	-3.52	-0.61	-1.93
P62494	Rab11a	RAB11a, member RAS oncogene family;24KG;rab-11;ras-related protein Rab-11A	P1	-2.60	0.72	-0.81
P63031	Mpc1	mitochondrial pyruvate carrier 1;apoptosis-regulating basic protein;brain protein 44-like protein	P1	-2.25	-0.40	-1.21
P67779	Phb	prohibitin	P1	-1.13	-0.65	-0.95
Q06437	Pdha2	pyruvate dehydrogenase (lipoamide) alpha 2;PDHE1-A type II;pyruvate dehydrogenase E1 alpha 2;pyruvate dehydrogenase E1 component subunit alpha, testis-specific form, mitochondrial	P1	-1.46	0.51	-0.47

Q08013	Ssr3	signal sequence receptor, gamma;SSR-gamma;TRAP-complex gamma subunit;TRAP-gamma;signal sequence receptor subunit gamma;translocon-associated protein subunit gamma	P1	-1.83	-0.28	-0.71
Q5BK63	Ndufa9	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 9;CI-39kD;NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial;NADH-ubiquinone oxidoreductase 39 kDa subunit;complex I-39kD;sperm flagella protein 3	P1	-0.96	-1.04	-1.25
Q5I0E7	Tmed9	transmembrane emp24 protein transport domain containing 9;p24 family protein alpha-2;p24alpha2;transmembrane emp24 domain-containing protein 9	P1	-1.32	0.75	-0.51
Q5RJR8	Lrrc59	leucine rich repeat containing 59;leucine-rich repeat-containing protein 59;protein p34	P1	0.26	-1.11	-0.69
Q5XIG4	Ociad1	OCIA domain containing 1;OCIA domain-containing protein 1;ovarian carcinoma immunoreactive antigen	P1	1.13	-1.24	-0.28
Q5XIM3	RSA-14-44	RSA-14-44 protein;ras homolog gene family, member A	P1	0.04	-1.28	-1.04
Q66HF3	Etfdh	electron-transferring-flavoprotein dehydrogenase;ETF dehydrogenase;ETF-QO;ETF-ubiquinone oxidoreductase;electron transfer flavoprotein-ubiquinone oxidoreductase, mitochondrial;electron transferring flavoprotein dehydrogenase	P1	-1.76	-1.26	-1.69
Q7TMZ5	Arl6ip1	ADP-ribosylation factor-like 6 interacting protein 1;ADP-ribosylation factor-like protein 6-interacting protein 1;ADP-ribosylation-like factor 6-interacting protein	P1	-1.53	-0.68	-1.15
Q7TP91	Surf4	Ab1-205	P1	-1.75	-0.72	-1.34
Q7TQ84	Fyttd1	forty-two-three domain containing 1;UAP56-interacting factor;forty-two-three domain-containing protein 1;protein 40-2-3	P1	-3.18	-4.59	-4.59
Q9WVB1	Rab6a	Ras-related protein Rab-6A	P1	-1.46	-0.71	-1.11
B2RZD4	Rpl34	ribosomal protein L34;60S ribosomal protein L34;ribosomal protein L34-like 2	P2	2.65	-0.59	1.03
D3ZD31	Mrc1	mannose receptor, C type 1; macrophage mannose receptor 1	P2	2.07	-1.50	-0.00
D3ZHD8	-	-	P2	2.56	-0.34	1.14
D3ZZW6	Hils1	histone linker H1 domain, spermatid-specific 1;histone H1-like protein in spermatids 1;spermatid-specific linker histone H1-like protein	P2	2.58	-0.93	0.83
D4A3K5	Hist1h1a	histone cluster 1, H1a;histone 1, H1a;histone H1.1;histone H1a	P2	4.29	-1.09	1.84
D4A478	Ntpcr	nucleoside-triphosphatase, cancer-related;nucleoside-triphosphatase C1orf57 homolog	P2	1.65	0.07	0.83
GENSCAN00000000770 1	-	-	P2	1.24	-1.24	-0.20

GENSCAN00000001389 9	-	-	P2	2.98	-0.34	1.40
GENSCAN00000002005 5	-	-	P2	1.88	-1.97	-0.74
H7C5Y5	Rpl6	ribosomal protein L6;60S ribosomal protein L6;neoplasm-related protein C140	P2	1.16	-1.22	-0.36
P06349	Hist1h1t	histone cluster 1, H1t;H1 histone family member T (testis-specific);H1 histone family, member 3;H1 histone family, member T (testis-specific);histone 1 a family 3;histone 1, H1t;histone 1, family 3;histone 1t;histone H1t;testis-specific histone 1 probably same as Hh1tts;testis-specific histone 1, probably same as Hh1tts	P2	3.93	-3.18	-0.42
P13471	Rps14	ribosomal protein S14;40S ribosomal protein S14	P2	3.29	-1.03	1.07
P18445	Rpl27a	ribosomal protein L27a;60S ribosomal protein L27a	P2	1.74	-0.81	0.36
P29314	Rps9;LOC 10090946 6	ribosomal protein S9;40S ribosomal protein S9;40S ribosomal protein S9-like	P2	1.21	-0.56	0.11
P36970-2	Gpx4	glutathione peroxidase 4;GSHPx-4;phospholipid hydroperoxide glutathione peroxidase, mitochondrial;phospholipid hydroperoxide glutathione peroxidase, nuclear	P2	4.29	-1.84	1.18
P61354	Rpl27	ribosomal protein L27;60S ribosomal protein L27	P2	1.79	-0.36	0.63
P62243	Rps8	ribosomal protein S8;40S ribosomal protein S8	P2	2.04	-0.95	0.37
P62755	Rps6;LOC 10091137 2	ribosomal protein S6;40S ribosomal protein S6;40S ribosomal protein S6-like	P2	1.17	0.00	0.35
P62832	Rpl23	ribosomal protein L23;60S ribosomal protein L23	P2	1.55	-0.18	0.62
P62850-2	Rps24	ribosomal protein S24;40S ribosomal protein S24	P2	3.29	-0.91	1.09
P83732	Rpl24	ribosomal protein L24;60S ribosomal protein L24;L30	P2	2.13	-0.45	0.73
Q68FP7	Gk2	glycerol kinase 2;glucokinase activity, related sequence 2	P2	1.20	-0.17	0.32
Q6IMY8	Hnrnpu	heterogeneous nuclear ribonucleoprotein U;system N1 Na+ and H+-coupled glutamine transporter;transporter protein; system N1 Na+ and H+-coupled glutamine transporter	P2	2.97	-2.90	-1.06
B2RYS8	Ndufb8	NADH dehydrogenase (ubiquinone) 1 beta subcomplex 8;NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 8, mitochondrial	P3	-1.11	-0.02	-0.65
Q5XI04	Stom	Protein Stom	P3	-0.15	1.21	0.53
D4AC23	Cct7	chaperonin containing Tcp1, subunit 7 (eta);T-complex protein 1 subunit eta;chaperonin subunit 7 (eta)	P3	0.31	1.81	1.02

F1LP05	Atp5a1	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle;ATP synthase subunit alpha, mitochondrial;ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit;mitochondrial H+-ATP synthase alpha subunit	P3	-0.51	1.19	0.22
F1LQ62	-	-	P3	-0.44	1.71	0.65
GENSCAN00000000698 9	-	-	P3	-0.47	1.28	0.42
GENSCAN00000002748 8	-	-	P3	-0.14	1.79	0.86
O88994	Marc2	mitochondrial amidoxime reducing component 2;MOCO sulphurase C-terminal domain containing 2;MOSC domain-containing protein 2, mitochondrial;moco sulfurase C-terminal domain-containing protein 2;molybdenum cofactor sulfurase C-terminal domain-containing protein 2	P3	-0.84	1.49	0.32
P20070-3	Cyb5r3	cytochrome b5 reductase 3;B5R;Diaphorase (NADH) (cytochrome b-5 reductase);NADH-cytochrome b5 reductase 3;diaphorase 1;diaphorase-1	P3	-1.61	1.58	0.27
P32089	Slc25a1	solute carrier family 25 (mitochondrial carrier, citrate transporter), member 1;citrate transport protein;citrate transporter) member 1;mitochondrial tricarboxylate carrier;tricarboxylate transport protein, mitochondrial	P3	-1.11	1.19	0.06
P62193	Psmc1	proteasome (prosome, macropain) 26S subunit, ATPase, 1;26S protease regulatory subunit 4;26S proteasome AAA-ATPase subunit RPT2;P26s4;peptidase (prosome, macropain) 26S subunit, ATPase 1;protease (prosome, macropain) 26S subunit, ATPase 1;proteasome 26S subunit ATPase 1	P3	-0.34	1.47	0.51
P62260	Ywhae	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide;14-3-3 epsilon;14-3-3 protein epsilon;MSF L;mitochondrial import stimulation factor (MSF) L subunit;mitochondrial import stimulation factor L subunit;tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activatioprotein epsilon polypeptide	P3	-0.23	1.31	0.54
Q05962	Slc25a4	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 4;ADP,ATP carrier protein 1;ADP/ATP translocase 1;ANT 1;adenine nucleotide translocator 1;mitochondrial adenine nucleotide translocator;solute carrier family 25 (mitochondrial adenine nucleotide translocator) member 4;solute carrier family 25 member 4	P3	-1.39	2.39	0.76
Q09073	Slc25a5	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5;ADP,ATP carrier protein 2;ADP/ATP translocase 2;ANT 2;Adenine nucleotid translocator 2 fibroblast isoform (ATP-ADP carrier protein);Adenine nucleotid translocator 2, fibroblast isoform (ATP-ADP carrier protein);adenine nucleotide translocator 2 fibroblast isoform (ATP-ADP carrier protein);adenine nucleotide translocator 2, fibroblast isoform (ATP-ADP carrier protein);solute carrier family 25 member 5	P3	-0.81	2.01	0.58
Q498D8	Rbx1	ring-box 1, E3 ubiquitin protein ligase;RING-box protein 1	P3	-0.07	1.47	0.69
Q4KM24	Pex11b	peroxisomal biogenesis factor 11 beta;peroxisomal biogenesis factor 11b;peroxisomal membrane protein 11B	P3	0.29	1.07	0.60

Q5U2S7	Psmd3	proteasome (prosome, macropain) 26S subunit, non-ATPase, 3;26S proteasome non-ATPase regulatory subunit 3;proteasome 26S non-ATPase subunit 3	P3	-0.35	2.29	1.06
Q5XIM9	Cct2	chaperonin containing TCP1, subunit 2 (beta);CCT-beta;T-complex protein 1 subunit beta;TCP-1-beta	P3	-0.81	1.41	0.24
Q64428	Hadha	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), alpha subunit;TP-alpha;hydroxyacyl-Coenzyme A dehydrogenase/3-ketoacyl-Coenzyme A thiolase/enoyl-Coenzyme A hydratase (trifunctional protein), alpha subunit;trifunctional enzyme subunit alpha, mitochondrial	P3	-1.02	0.64	-0.25
Q6AXX6	RGD1309676	similar to RIKEN cDNA 5730469M10;UPF0765 protein C10orf58 homolog;peroxiredoxin-like 2 activated in M-CSF stimulated monocytes;redox-regulatory protein FAM213A;redox-regulatory protein PAMM;sperm head protein 1	P3	0.34	1.34	0.81
B0BN81	Rps5	ribosomal protein S5;40S ribosomal protein S5	P4	1.19	0.61	0.70
B0BNC7	Armc12	armadillo repeat containing 12;armadillo repeat-containing protein 12	P4	2.45	1.77	2.29
D3Z8F7	Gykl1	glycerol kinase-like 1	P4	1.54	0.84	0.96
D3ZAC5	Dnajb3	DnaJ (Hsp40) homolog, subfamily B, member 3;dnaJ homolog subfamily B member 3	P4	2.19	2.00	2.58
D3ZFA8	LOC100364909	Protein LOC100362366	P4	2.54	0.77	1.84
D3ZLL8	LOC687680	Protein LOC100909878	P4	2.01	1.33	2.04
D3ZPN7	LOC100360604	ribosomal protein L21-like	P4	1.67	0.31	0.97
D3ZVY6	LOC100359986	Protein LOC100359986	P4	5.33	0.32	3.55
D3ZX01	Rps4y2	ribosomal protein S4, Y-linked 2	P4	1.33	0.79	0.96
D4A1P2	Rpl10l	ribosomal protein L10-like;60S ribosomal protein L10-like	P4	2.08	0.59	1.49
P47820	Ace	Angiotensin-converting enzyme	P4	1.95	2.80	3.41
F1M779	Cltc	clathrin, heavy chain (Hc);clathrin heavy chain 1;clathrin, heavy polypeptide (Hc)	P4	1.36	1.27	1.54
G3V6I9	Rpl26	ribosomal protein L26;60S ribosomal protein L26	P4	3.53	0.12	1.96
G3V9N4	Adam6	a disintegrin and metallopeptidase domain 6;a disintegrin and metalloproteinase domain 6;tMDC IV	P4	2.42	1.45	2.36
GENSCAN00000002761 0	-	-	P4	1.30	1.70	1.65
P09895	Rpl5	ribosomal protein L5;60S ribosomal protein L5	P4	3.69	0.10	1.91
P17077	Rpl9; LOC1003604 LOC100364457	ribosomal protein L9;60S ribosomal protein L9;ribosomal protein L9-like	P4	2.80	1.81	2.80

P17078	Rpl35	ribosomal protein L35;60S ribosomal protein L35	P4	2.57	0.23	1.38
P35571	Gpd2	glycerol-3-phosphate dehydrogenase 2, mitochondrial;GPD-M;GPDH-M;Glycerol-3-phosophate dehydrogenase 2 (mitochondrial);glycerol-3-phosphate dehydrogenase, mitochondrial;mtGPDH gene, promoter region and alternative transcripts	P4	2.72	0.83	2.13
P39069	Ak1	adenylate kinase 1;ATP-AMP transphosphorylase 1;ATP:AMP phosphotransferase;adenylate kinase isoenzyme 1;adenylate monophosphate kinase;myokinase	P4	1.10	1.65	1.70
P55063	Hspa1l	heat shock protein 1-like;HSP70.3;Heat shock protein 70-1l;heat shock 70 kDa protein 1-like;heat shock 70 kDa protein 1L;heat shock 70 kDa protein 3	P4	2.36	3.06	3.69
P60868	Rps20; LOC100359951; LOC100362149; LOC100362684	ribosomal protein S20;40S ribosomal protein S20;ribosomal protein S20-like	P4	2.25	1.38	2.31
P62278	Rps13; LOC684988	ribosomal protein S13;40S ribosomal protein S13;similar to ribosomal protein S13;ribosomal protein S13-like	P4	2.54	1.03	2.10
P62845	Rps15	ribosomal protein S15;40S ribosomal protein S15;RIG protein;insulinoma	P4	2.61	0.41	1.67
P62909	Rps3	ribosomal protein S3;40S ribosomal protein S3	P4	1.39	0.92	1.17
P62914	-	-	P4	2.16	0.91	1.71
Q4FZS5	MGC95210	hypothetical LOC287798;uncharacterized protein LOC287798	P4	2.71	1.06	2.21
Q62665	-	-	P4	1.68	0.47	0.81
Q62803	Spam1	sperm adhesion molecule 1 (PH-20 hyaluronidase, zona pellucida binding);hyal-PH20;hyaluronidase PH-20;hyaluronoglucosaminidase PH-20;sperm surface antigen 2B1;sperm surface protein PH-20	P4	1.71	1.36	1.89
Q68FR6	Eef1g	eukaryotic translation elongation factor 1 gamma;EF-1-gamma;eEF-1B gamma;elongation factor 1-gamma	P4	2.76	2.17	2.71
Q6AXV6	Spert	spermatid associated;spermatid-associated protein	P4	3.12	2.08	3.07
Q920Q3	Spata19	spermatogenesis associated 19;spermatogenesis-associated protein 19, mitochondrial;spermatogenic cell-specific gene 1 protein;spermatogenic specific-gene1	P4	2.61	1.24	2.41
Q9Z1B2	Gstm5; LOC100912430	glutathione S-transferase, mu 5;GST class-mu 5;glutathione S-transferase Mu 5;glutathione S-transferase Mu 5-like	P4	2.10	2.04	2.68
D3ZDK7	Pgp	phosphoglycolate phosphatase	P5	1.84	2.70	2.76
P21769	Odf1	Outer dense fiber protein 1	P5	1.79	3.30	3.46
D4A5R7	-	-	P5	0.64	2.68	2.08
D4A781	Ipo5	importin 5;RAN binding protein 5;importin-5;karyopherin (importin) beta 3	P5	0.28	3.05	2.21

Q7TNZ0	LOC365778	HCBP6	P5	1.64	3.00	2.80
F1LSR7	Spata20	spermatogenesis associated 20;sperm protein SSP411;sperm-specific protein 411;spermatogenesis-associated protein 20	P5	2.24	5.33	5.33
F1M2D2	Cct6b	chaperonin containing Tcp1, subunit 6B (zeta 2);T-complex protein 1 subunit zeta-2;chaperonin subunit 6b (zeta)	P5	0.35	2.18	1.41
P97536	Cand1	Cullin-associated NEDD8-dissociated protein 1	P5	0.38	1.89	1.35
G8JLS1	-	-	P5	0.24	2.45	1.64
GENSCAN00000007350	-	-	P5	2.11	2.86	3.16
GENSCAN00000009290	-	-	P5	0.29	4.29	3.29
GENSCAN00000012385	-	-	P5	-0.10	2.98	1.91
GENSCAN00000031767	-	-	P5	1.45	2.87	2.72
GENSCAN00000036233	-	-	P5	-0.89	3.39	1.92
GENSCAN00000038686	-	-	P5	-0.21	1.24	0.56
P05708	Hk1	hexokinase 1;HK I;brain form hexokinase;hexokinase type I;hexokinase-1	P5	1.59	3.23	3.46
P07687	Ephx1	epoxide hydrolase 1, microsomal (xenobiotic);epoxide hydratase;epoxide hydrolase 1;epoxide hydrolase 1 (microsomal xenobiotic hydrolase);liver microsomal xenobiotic epoxide hydrolase;microsomal epoxide hydrolase	P5	0.21	2.75	1.77
P11598	Pdia3	protein disulfide isomerase family A, member 3;58 kDa glucose-regulated protein;58 kDa microsomal protein;ER protein 57;ER protein 60;ER-60 protease;ERp60;HIP-70;Q-2;disulfide isomerase ER-60;endoplasmic reticulum resident protein 57;endoplasmic reticulum resident protein 60;glucose regulated protein, 58 kDa;oxidoreductase ERp57;p58;protein disulfide isomerase associated 3;protein disulfide-isomerase A3	P5	0.84	3.13	2.54
P14659	Hspa2	heat shock protein 2;HST;heat shock 70kDa protein 2;heat shock protein 70.2;heat shock protein alpha 2;heat shock-related 70 kDa protein 2;testis-specific heat shock protein-related gene hst70	P5	0.87	1.39	1.31
P18420	Psma1	proteasome (prosome, macropain) subunit, alpha type 1;macropain subunit C2;multicatalytic endopeptidase complex subunit C2;proteasome alpha 1 subunit;proteasome component C2;proteasome nu chain;proteasome subunit alpha type-1	P5	0.72	1.61	1.23
P23711	Hmox2	heme oxygenase (decycling) 2;heme oxygenase 2;heme oxygenase-2 non-reducing isoform	P5	1.23	2.21	2.16
P28480	Tcp1	t-complex 1;CCT-alpha;T-complex protein 1 subunit alpha;TCP-1-alpha	P5	-0.43	2.39	1.33

P55054	Fabp9	fatty acid binding protein 9, testis;15 kDa perforatorial protein;PERF 15;T-FABP;fatty acid-binding protein 9;testis lipid binding protein;testis lipid-binding protein;testis-type fatty acid-binding protein	P5	0.91	2.08	1.74
P62944	Ap2b1; LOC100912146		P5	1.44	2.13	2.17
P63102	Ywhaz	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide;14-3-3 protein zeta/delta;KCIP-1;mitochondrial import stimulation factor S1 subunit;protein kinase C inhibitor protein 1	P5	0.47	2.60	2.01
P82995	Hsp90aa1	heat shock protein 90, alpha (cytosolic), class A member 1;HSP 86;heat shock 86 kDa;heat shock protein 1, alpha;heat shock protein 86;heat shock protein HSP 90-alpha	P5	-0.13	3.29	2.00
P83868	Ptges3	prostaglandin E synthase 3 (cytosolic);cPGES;cytosolic prostaglandin E2 synthase;hsp90 co-chaperone;progesterone receptor complex p23;prostaglandin E synthase 3;prostaglandin-E synthase 3;telomerase-binding protein p23	P5	0.62	2.21	1.61
Q0VGK4	Gdpd1	glycerophosphodiester phosphodiesterase domain containing 1;glycerophosphodiester phosphodiesterase domain-containing protein 1	P5	1.81	2.76	3.19
Q3MHS9	Cct6a	chaperonin containing Tcp1, subunit 6A (zeta 1);T-complex protein 1 subunit zeta;chaperonin subunit 6a (zeta)	P5	-0.25	2.16	0.93
Q4V8H5	Dnpep	aspartyl aminopeptidase	P5	0.36	4.82	4.29
Q5XI00	Tcp11	t-complex protein 11;T-complex protein 11 homolog;t-complex 11	P5	2.99	3.93	4.82
Q60587	Hadhb		P5	0.11	2.16	1.25
Q641Z6	Ehd1	EH-domain containing 1;EH domain-containing protein 1	P5	0.81	1.52	1.30
Q66H38	Fam71b	family with sequence similarity 71, member B;Golgi-associated Rab2B interactor-like 3;protein FAM71B	P5	0.77	1.31	1.05
Q68FQ0	Cct5	chaperonin containing Tcp1, subunit 5 (epsilon);CCT-epsilon;T-complex protein 1 subunit epsilon;TCP-1-epsilon;chaperonin subunit 5 (epsilon)	P5	-0.22	2.33	1.20
Q6AY56	Tuba8	tubulin, alpha 8;alpha-tubulin 8;tubulin alpha-8 chain	P5	1.44	2.50	2.61
Q71SY3	Tsn	translin	P5	-0.09	3.69	2.50
Q7TPB1	Cct4	chaperonin containing Tcp1, subunit 4 (delta);CCT-delta;T-complex protein 1 subunit delta;TCP-1-delta;chaperonin subunit 4 (delta)	P5	0.22	1.84	0.87
B0K031	Rpl7	ribosomal protein L7;60S ribosomal protein L7		-0.01	-0.57	-0.45

B1WBY5	Dnajc11	DnaJ (Hsp40) homolog, subfamily C, member 11;dnaJ homolog subfamily C member 11		-0.25	-0.14	-0.49
B1WBY7	Erlin1	ER lipid raft associated 1;SPFH domain family, member 1;erlin-1		-0.28	0.40	-0.12
B2RZD1	Sec61b	Sec61 beta subunit;protein transport protein Sec61 subunit beta		0.12	-0.35	-0.42
D3ZE15	LOC100911483	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 13-like		-0.93	-0.09	-0.51
D3ZGW9	Gml	glycosylphosphatidylinositol anchored molecule like;GPI anchored molecule like protein;glycosyl-phosphatidylinositol-anchored molecule-like protein;glycosylphosphatidylinositol anchored molecule like protein		0.63	0.30	0.18
D3ZKD8	Fam209a	family with sequence similarity 209, member A;uncharacterized protein LOC296411		0.93	0.52	0.59
Q3B7V5	Rab2b	Protein Rab2b		-0.85	0.46	-0.29
P61314	Rpl15	60S ribosomal protein L15		-0.32	0.02	-0.36
D4A6B2	-	-		-0.89	-0.32	-0.70
D4A899	Vps13a	Protein Vps13a		0.85	0.48	0.50
D4AA84	-	-		0.79	-0.51	0.00
ENSRNOP00000026197	-	-		0.27	0.04	-0.05
F1LRA1	Lman1	lectin, mannose-binding, 1;ER-Golgi intermediate compartment 53 kDa protein;endoplasmic reticulum-golgi intermediate compartment protein 53;protein ERGIC-53		-1.42	-0.99	-1.64
F1LSW7	Rpl14	ribosomal protein L14;60S ribosomal protein L14		0.28	-0.26	-0.23
F1M2H8	-	-		0.80	-0.18	0.07
P51146	Rab4b	Ras-related protein Rab-4B		-0.30	-0.13	-0.38
G3V6H5	-	-		-0.69	-0.32	-0.55
G3V8N9	Tex101	testis expressed 101;lipid raft-associated glycoprotein TEC-21;testis expressed gene 101;testis-expressed protein 101;testis-expressed sequence 101 protein		0.52	0.06	-0.06
G3V9S0	Cyb5r1	cytochrome b5 reductase 1;NAD(P)H:quinone oxidoreductase type 3, polypeptide A2;NADH-cytochrome b5 reductase 1;b5R.1		-0.24	0.31	-0.26
G3V9Y7	-	-		0.04	0.39	0.04
GENSCAN00000004573	-	-		0.08	0.28	-0.05
GENSCAN00000008925	-	-		0.58	-0.63	-0.26
GENSCAN00000019271	-	-		-0.67	-0.03	-0.53

GENSCAN00000026104	-	-		-0.86	0.29	-0.33
GENSCAN00000032928	-	-		-0.03	-0.51	-0.54
GENSCAN00000041608	-	-		0.51	0.48	0.32
P06761	Hspa5			-0.18	0.37	-0.14
P07153	Rpn1	ribophorin I;RPN-I;dolichyl-diphosphooligosaccharide--protein glycosyltransferase 67 kDa subunit;dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 1;ribophorin-1;ribophorin1		-0.68	0.44	-0.23
P12001	Rpl18	ribosomal protein L18;60S ribosomal protein L18		0.23	-0.40	-0.32
P29419	Atp5i			-0.79	-0.69	-0.95
P35435	-	-		0.46	-0.46	-0.24
P50878	Rpl4	ribosomal protein L4;60S ribosomal protein L1;60S ribosomal protein L4;L1		-0.01	0.32	-0.04
P62198	Psmc5			-0.20	0.88	0.18
P62250	Rps16	ribosomal protein S16;40S ribosomal protein S16		0.92	0.93	0.72
P62271	Rps18; LOC100360679	ribosomal protein S18;40S ribosomal protein S18;ribosomal protein S18-like		0.62	-0.20	0.10
P62718	Rpl18a	ribosomal protein L18A;60S ribosomal protein L18a		0.71	0.45	0.37
P62919	Rpl8;LOC100360117; LOC100910370	ribosomal protein L8;60S ribosomal protein L8;ribosomal protein L8-like;60S ribosomal protein L8-like		0.81	-0.27	0.40
P62982	Rps27a; LOC100912032	ribosomal protein S27a;40S ribosomal protein S27a;ubiquitin carboxyl extension protein 80;ubiquitin-40S ribosomal protein S27a;ubiquitin-40S ribosomal protein S27a-like		0.56	0.54	0.52
P63322	Rala	v-ral simian leukemia viral oncogene homolog A (ras related);-ral simian leukemia viral oncogene homolog A (ras related);ras-related protein Ral-A		-0.27	0.81	0.09
P84100	Rpl19	ribosomal protein L19;60S ribosomal protein L19		0.20	0.67	0.23
P97521	-	-		-0.48	0.25	-0.24
Q3ZU82	Golga5	golgin A5;Golgin subfamily A member 5;golgi autoantigen, golgin subfamily a, 5;golgin-84		0.40	0.87	-0.27
Q5EB77	Rab18	RAB18, member RAS oncogene family;ras-related protein Rab-18		-0.73	0.17	-0.38
Q5HZY0	Ubxn4	UBX domain protein 4;UBX domain containing 2;UBX domain-containing protein 2;UBX domain-containing protein 4;erasin		-0.59	-0.37	-0.67
Q5RJY4-2	Dhrs7b	dehydrogenase/reductase (SDR family) member 7B;SDR family dehydrogenase/reductase member 7B;dehydrogenase/reductase SDR family member 7B		-0.45	0.48	-0.31

Q5RKG3	H1fnt	H1 histone family, member N, testis-specific;haploid germ cell-specific nuclear protein 1;histone H1t2;testis-specific H1 histone		0.49	0.26	0.05
Q5U2X4	Dpep3	dipeptidase 3;putative membrane-bound dipeptidase 3		0.59	0.27	0.25
Q5XIH7	Phb2	prohibitin 2;B-cell receptor associated protein 37;B-cell receptor-associated protein 37;B-cell receptor-associated protein BAP37;prohibitin-2		-0.57	-0.71	-0.79
Q5XIU4	Bcap29	B-cell receptor-associated protein 29;B-cell receptor-associated protein BAP29		0.42	-0.19	-0.12
Q5XIU9	Pgrmc2	progesterone receptor membrane component 2;membrane-associated progesterone receptor component 2;progesterone membrane binding protein		-0.40	-0.44	-0.60
Q63584	Tmed10	transmembrane emp24-like trafficking protein 10 (yeast); 21 kDa transmembrane-trafficking protein; integral membrane protein Tmp21-I (p23); p24 family protein delta-1;p24delta1; transmembrane emp24 domain-containing protein 10; transmembrane protein Tmp21; transmembrane trafficking protein 21		-0.55	0.13	-0.33
Q641Y0	Ddost			-0.33	0.18	-0.41
Q642E2	Rpl28	ribosomal protein L28;60S ribosomal protein L28		0.33	0.26	0.13
Q66HA6	Arl8b	ADP-ribosylation factor-like 8B;ADP-ribosylation factor-like protein 8B		-0.38	-0.36	-0.53
Q68FW4	Stx18	syntaxin 18;syntaxin-18		-0.59	-0.56	-0.73
Q6AXV4	Samm50	SAMM50 sorting and assembly machinery component;sorting and assembly machinery component 50 homolog;sorting and assembly machinery component 50 homolog A		-0.17	-0.19	-0.22
Q6AY30	Sccpdh;LOC100910414	saccharopine dehydrogenase (putative);probable saccharopine dehydrogenase;saccharopine dehydrogenase-like oxidoreductase;saccharopine dehydrogenase-like oxidoreductase-like		0.51	0.11	0.14
Q6AY52	LOC100362783	Uncharacterized protein C7orf61 homolog;uncharacterized protein LOC100362783		0.42	0.63	0.35
Q9Z270	Vapa	VAMP (vesicle-associated membrane protein)-associated protein A;33 kDa Vamp-associated protein;VAMP-A;VAMP-associated protein A;VAP-33;VAP-A;vesicle-associated membrane protein, associated protein a;vesicle-associated membrane protein-associated protein A		-0.80	0.87	-0.03

Chapitre 4

Identification de protéines impliquées dans le devenir des corps résiduels par protéomique Shotgun chez le rat

I. Contexte et objectifs

Les corps résiduels (CRs) sont la portion cytoplasmique laissée par les spermatides matures derrière elles lorsqu'elles se détachent des cellules de Sertoli au moment de la spermiation, au stade VII-VIII de l'épithélium séminifère (Figure 31). Comme nous l'avons vu en introduction générale, la formation des CRs et leur phagocytose par les cellules de Sertoli semblent avoir une importance capitale dans la synchronisation de l'épithélium séminifère. Cette dernière pourrait être régulée par la phagocytose des CRs. Le laboratoire a montré que la phagocytose des CRs induit la production d' IL-1 et par conséquent d'IL-6 par les cellules de Sertoli (Syed et al., 1995). L'hypothèse émise par notre laboratoire sur la base de ces travaux était que les deux cytokines, IL-1 et IL-6, agissaient sur l'entrée en méiose des spermatocytes pré-leptotène (Dugast et Jégou, 1994) et sur la dynamique de la BHT (Lie et al., 2011; Zhang et al., 2014), pour au final stimuler la prolifération des spermatogonies (Gérard et al., 1992). De même, Liu et collaborateurs ont montré que la phagocytose des CRs induisait la production de l'activateur du plasminogène (Liu, 2007; Liu et al., 1995).

Le laboratoire a montré qu'après leur phagocytose par les cellules de Sertoli, les CRs fusionnent avec des lysosomes, et leur membrane est progressivement dégradée dans le cytoplasme des cellules de Sertoli; les spermatides allongées subissant la dégradation sont aussi observables dans le cytoplasme de la cellule de Sertoli (Pineau et al., 1991). Cependant, le processus qui permet à la cellule de Sertoli de discriminer les cellules germinales apoptotiques des CRs pour permettre leur dégradation ciblée reste mal connu, malgré quelques études (Pineau et al., 1991; Kawasaki et al., 2002; Yefimova et al., 2008; Elliott et al., 2010). De même, le devenir des CRs au sein des cellules de Sertoli après leur phagocytose par ces dernières est encore mal documenté. On sait qu'après reconnaissance par les récepteurs SR-BI (class B scavenger receptor type I) des phosphatidylsérines exposées à la surface des CRs, (Blanco-Rodríguez et Martínez-García, 1999; Kawasaki et al., 2002; Nakanishi et Shiratsuchi, 2004), ces derniers sont internalisés par la cellule de Sertoli, et il s'y forme un phagosome, avec lequel les lysosomes sertoliens fusionnent pour former des phagolysosomes (Morales et al., 1985). Ces derniers peuvent être transportés le long des microtubules jusqu'à la base de l'épithélium séminifère (Qian et al., 2014). Il existe aussi une hypothèse selon laquelle les CRs subiraient une autophagie concomitante avec leur phagocytose, ce qui donnerait une explication supplémentaire pour leur dégradation au sein des cellules de Sertoli (Chemes, 1986; Russell et al., 1989). Ceci expliquerait aussi la

proportion plus importante de lysosomes dans les cellules de Sertoli au moment de la spermiation, comparée à cette proportion aux autres stades du cycle de l'épithélium séminifère (Chemes, 1986). Les CRs, *via* leur phagocytose et leur dégradation par les cellules de Sertoli, représentent d'ailleurs une source énergétique importante pour ces dernières *via* la β-oxydation des lipides (Xiong et al., 2009). Certains mécanismes concernant la phagocytose par les cellules de Sertoli sont décrits, mais dans ces études, il n'est pas question d'une différenciation entre les cellules germinales apoptotiques et les CRs. En effet, les récepteurs TAM: Tyrosine kinase (Tyro3), AXL et MER (Lemke et Rothlin, 2008; Xiong *et al*., 2008), promeuvent la reconnaissance et la phagocytose du CR ou des cellules apoptotiques par la cellule de Sertoli, et y sont d'ailleurs nécessaires. En revanche, le ligand Gas6 de ces récepteurs n'est pas spécifique de la membrane des CRs ou des cellules apoptotiques. Ces récepteurs TAM ont un rôle dans la modulation de l'immunité innée, et donc de l'inflammation (Lemke et Rothlin, 2008). La phagocytose par les cellules de Sertoli implique donc des mécanismes proches de ceux de l'immunité innée, comme évoqué dans le chapitre 3 au sujet des TLRs (Toll-like receptors).

Bien que les CRs soient décrits dans un certain nombre d'études, notamment sur leur fonction de régulation de l'épithélium séminifère, et que leur contenu soit partiellement connu (Pineau et al., 1991); les mécanismes de leur formation au sein des spermatides en allongement ne sont pas totalement élucidés. Une étude décrit un mécanisme de formation des corps résiduels, selon lequel la portion cytoplasmique caudale de la spermatide allongée qui se destine à devenir un corps résiduel, se met à présenter des caractéristiques de corps apoptotique, en dépit du noyau sain de la spermatide allongée (Blanco-Rodríguez et Martínez-García, 1999). Cette portion cytoplasmique « apoptotique » exprime des molécules typiques de l'apoptose telles que les protéines p53, p21 et c-jun (Blanco-Rodríguez et Martínez-García, 1999).

En caractérisant de manière aussi exhaustive que possible le protéome des CRs, nous espérons pouvoir mettre en évidence, par une analyse de données orientée à l'aide des annotations de la Gene Ontology, d'autres protéines susceptibles d'être impliquées dans la formation ou le devenir des corps résiduels. Ces protéines pourraient jouer un rôle dans la synchronisation du cycle spermatogénétique. Notre étude est originale, puisque aucune étude protéomique des corps résiduels n'a été menée à ce jour.

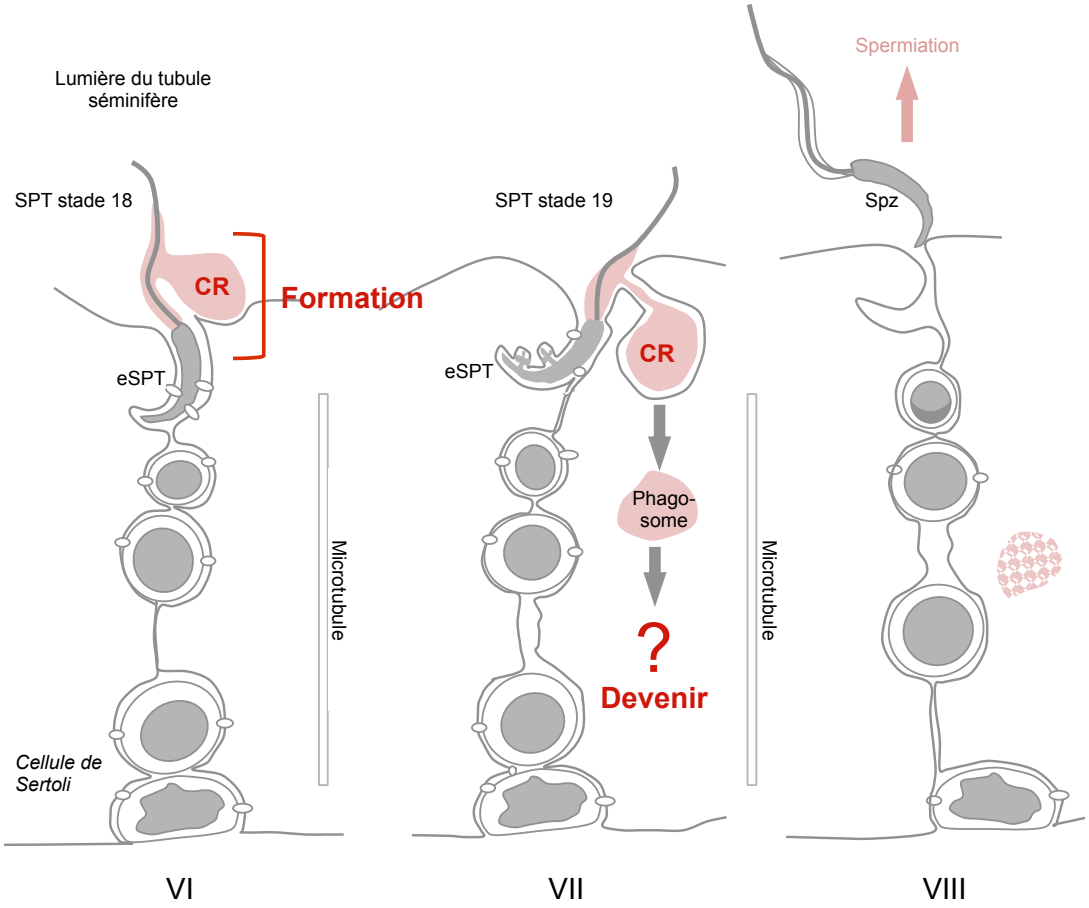

Figure 31. Objectif de l'étude consistant à identifier des protéines potentiellement impliquées dans la formation et le devenir des corps résiduels chez le rat

Les corps résiduels sont formés à partir du cytoplasme des spermatides en allongement. Les mécanismes de leur formation sont encore mal connus. Les corps résiduels sont phagocytés par la cellule de Sertoli au moment de la spermiation, au stade VII-VIII de l'épithélium séminifère. Ils sont internalisés et dirigés vers le pôle basal des cellules de Sertoli le long de microtubules, mais leur devenir au sein de la cellule de Sertoli reste mal connu. L'objectif de ce projet est d'identifier des protéines impliquées au sens large dans la formation et / ou le devenir des corps résiduels.

II. Méthodes

Les corps résiduels de rats mâles Sprague-Dawley de 90 jours ont été purifiés par la méthode d'élutriation centrifuge décrite dans l'article Pineau et collaborateurs, sans la dissociation trypsique (Pineau et al., 1993), avec une pureté de plus de 80%. Trois expériences indépendantes ont été réalisées, chacune à partir d'un pool de huit rats. Pour chaque expérience, l'extraction des protéines totales des corps résiduels, la séparation des protéines sur gel 1D et les digestats protéiques ont été réalisés comme décrit dans l'article présenté

dans le Chapitre 1. Pour chaque expérience indépendante, les digestats protéiques issus des protéines séparées sur gel (20 digestats correspondant à 20 bandes découpées du gel) ont été analysés chacun en triple injection, afin d'obtenir un maximum d'identifications de protéines. L'analyse en triple injection avec liste d'exclusion consiste à établir une liste d'exclusion des m/z entre chaque injection pour permettre au spectromètre de masse d'ignorer les peptides déjà identifiés lors de l'injection précédente (Lavigne et al., 2012). L'analyse LC-MS/MS en triple injection a également été réalisée comme décrit dans l'article du Chapitre 1. Pour chaque expérience, les données de spectres des trois injections sont interrogées *via* les moteurs de recherche SEQUEST et Mascot contre la base de donnée UniProt *rattus norvegicus* (release 2014_02), avec le logiciel Proteome Discoverer™ 1.2 (Thermo Scientific). Les résultats de recherche sont extraits des fichiers .msf à l'aide de Proteome Discoverer. Les peptides pris en compte dans l'analyse sont les peptides « high confident » (FDR 1%), et de rang 1.

III. Résultats et discussion

L'analyse Shotgun des extraits de protéines totales de corps résiduels par nano-LC MS/MS en triple injection et avec une liste d'exclusion dynamique permet d'accéder au protéome en profondeur. Elle a permis d'identifier dans les corps résiduels environ 20% de protéines en plus qu'avec une simple acquisition LC-MS/MS. Les 3 expériences indépendantes ont permis d'identifier au total 3.740 protéines non redondantes, dont 1.422 (soit un peu plus d'un tiers des identifications totales) se trouvent dans l'intersection de ces trois expériences. Une filtration des protéines identifiées sur la base des termes associés de la Gene Ontology (GO) permet d'orienter le choix de protéines candidates pour une étude ciblée en fonction de notre questionnement concernant : 1) la formation, et 2) le devenir des CRs. Sur l'ensemble des 3.740 protéines identifiées dans les corps résiduels, 2.679 ont un Gene ID qui permet de retrouver ces annotations GO. Nous pouvons choisir par exemple, les termes protéasome (135 protéines), phagocytic vesicle/phagocytosis (63 protéines), phagosome (5 protéines) et autophagy (55 protéines), pour extraire des groupes restreints de protéines potentiellement intéressantes dans l'étude du devenir des corps résiduels. Le terme: innate immunity (45 protéines), et les termes associés à inflammatory response (138 protéines), semblent importants à explorer en ce qui concerne le devenir de CRs, étant donné l'importance qu'ont

les récepteurs de l'immunité dans la phagocytose des CRs, et l'importance d'une signalisation de type inflammatoire (synthèse d'IL-1 et IL-6) associée à cette phagocytose.

Pour étudier la formation des CRs, nous pourrions sélectionner les protéines annotées avec le terme vesicle (2.680 protéines) ou exocytosis (138 protéines), afin de trouver des protéines potentiellement impliquées dans celle-ci. Mais ces listes sont encore trop importantes, et ces termes sont encore trop larges pour permettre une analyse ciblée. Par ailleurs, nous avons vu l'importance des phénomènes apoptotiques dans la formation du CR et son devenir. Il est donc aussi intéressant de regarder de plus près les protéines impliquées dans les phénomènes apoptotiques ou nécrotiques, donc annotées par des termes tels que : cell death. Dans notre liste, 1.184 protéines sont annotées par ce terme dont 2 sont annotées avec le terme autophagic cell death, (la protéine Lamp1 et la Cathepsin L1); une protéine est annotée avec le terme positive regulation of necrotic cell death (Cyclophilin D); 4 protéines sont annotées avec le terme programmed necrotic cell death (phosphoglycerate mutase family member 5 ; dynamin 1-like ; peptidylprolyl isomerase F; Sirtuin 2) ; parmi d'autres. Deux protéines peu connues relatives à la mort cellulaire et qui semblent intéressantes sont identifiées dans les corps résiduels: Pdcd10 (Programmed cell death protein1), et ALG-2-interacting protein 1 (Programmed cell death 6-interacting protein). La protéine Pdcd10 joue un rôle dans la régulation de l'exocytose induite par un ligand (Zhang et al., 2013), et elle est requise dans la migration des neurones en formation (Louvi et al., 2014). La protéine, ALG-2-interacting protein 1 est un régulateur de l'apoptose *via* l'activation de la caspase 9 induite par le calcium, et par l'intermédiaire de son partenaire: la protéine Alix (Strappazzon et al., 2010). Son partenaire ALG-2 est aussi impliqué dans la déformation de la membrane (Sadoul, 2006), le bourgeonnement et la régulation des récepteurs de surface cellulaire (Chen et al., 2005; Shi et al., 2010; Wang et al., 2014b), ce qui est très intéressant dans le cadre de l'étude de la formation du corps résiduel. Cette protéine ALG-2 n'est pas décrite dans le testicule. Voici donc un candidat à considérer pour l'étude de la formation du corps résiduel.

Pour chercher quelles protéines pourraient interagir entre les cellules de Sertoli et les CRs, j'ai réalisé le protéome des cellules de Sertoli, par les techniques d'isolement et de culture maîtrisées au laboratoire sur des rats de 20 jours (Skinner et Fritz, 1985; Toebosch et al., 1989). Les données de protéomique pour les cellules de Sertoli ont été acquises comme pour les extraits protéiques des autres types cellulaires en LC-MS/MS. Cette étude n'est pas présentée, car la limite en temps ne m'a pas permis d'exploiter les données. En revanche, j'ai

pu rechercher par une analyse informatique les protéines des cellules de Sertoli susceptibles d'interagir avec les protéines des CR. Par là même, j'ai recherché un signal protéique potentiel à la membrane des CRs qui serait en jeu dans leur formation ou leur phagocytose par les cellules de Sertoli. Cette recherche a été effectuée *via* l'interrogation des banques d'interactomique Mint, Intact, NCBI BioGRID, et NCBI-BIND pour les listes des protéines identifiées dans ces deux types cellulaires, à l'aide du logiciel AMEN (Chalmel et Primig, 2008). Cette première analyse n'a pas permis de mettre en évidence des couples d'interactants entre les cellules de Sertoli et les corps résiduels. Aucun couple de partenaires intéressants ne transparaît de cette analyse, à part ceux qui sont connus et décrits dans le testicule tels que Akt1 /PdpK1 (Dong et al., 2002; Siu et al., 2005).

La question de la formation des CRs, et la question: « pourquoi la phagocytose des CRs déclenche-t-elle la synthèse d'IL-1/ IL-6 et la synchronisation de l 'épithélium séminifère, contrairement à la phagocytose des cellules germinales apoptotiques par les cellules de Sertoli ? » demeurent. Pourtant, il semble, étant donné les résultats de l'étude différentielle décrite au précédent chapitre, qu'il faille plutôt, pour répondre à cette dernière question, chercher du côté de l'absence d'un signal dans les CRs *versus* sa présence dans les cellules germinales plutôt que l'inverse. Il est à noter que chez l'homme, la phagocytose des CRs par la cellule de Sertoli n'a pas lieu (Breucker et al., 1985), et que ceux-ci jouent un rôle important dans le devenir du spermatozoïde. En effet, la portion cytoplasmique associée avec les spermatozoïdes matures est le siège d'activités enzymatiques importantes, comme celle de l'énolase par exemple, qui semblent être impliquées dans le potentiel fécondant des spermatozoïdes (Force et al., 2002, 2004).

DISCUSSION ET PERSPECTIVES

Les travaux réalisés au cours de cette thèse ont été très informatifs tant sur le plan biologique pour la connaissance de la spermatogenèse, que sur le plan méthodologique en apportant une preuve de concept validant une approche de génomique intégrative appliquée à la spermatogenèse. Je discuterai des apports de mes études protéomiques pour la compréhension de la spermatogenèse, et de la nécessité d'intégrer les données Omiques pour rendre ces études protéomiques exploratoires fructueuses. Puis, je discuterai de certaines limites des différentes approches utilisées et des moyens possibles pour les améliorer. Enfin, je proposerai des perspectives qui peuvent s'ouvrir suite à ces travaux de recherche, au sein de notre équipe et à l'international.

L'objectif de cette thèse était de mettre en évidence des protéines susceptibles d'avoir un rôle dans la spermatogenèse et plus particulièrement dans la spermiogénèse chez le rat. Trois études de protéomique exploratoire ont été menées, dont l'une a pu aboutir à la mise en évidence de nouveaux évènements codants par des analyses biochimiques de l'expression du transcrit et de la protéine. Notre seconde étude a permis de mettre en évidence des protéines membranaires d'intérêt avec une expression différentielle au cours de la maturation des cellules germinales. Une troisième étude, à l'état préliminaire, a consisté à établir le protéome du corps résiduel et constitue un potentiel de découverte important pour la compréhension de sa biologie. La protéomique et la spectrométrie de masse sur lesquelles s'appuient ces études devraient encore impacter notre compréhension de la spermatogenèse. En effet, mes travaux ont permis l'identification par LC-MS/MS de plus de 3.000 protéines non redondantes dans les cellules de Sertoli et 5.860 protéines non redondantes sur l'ensemble des cellules germinales et des corps résiduels. Ils ont aussi permis de découvrir 69 nouveaux transcrits correspondant à 44 nouveaux gènes exprimés de manière spécifique dans les cellules méiotiques et post-méiotiques, grâce à l'utilisation d'une banque personnalisée de séquences comprenant des séquences protéiques déduites de transcrits nouvellement séquencés. Pour deux protéines d'intérêt, la nouvelle énolase T-ENOL et la nouvelle protéine « VAMP-7 like » que nous avons appelée VAMP9 (XLOC_013843), l'expression du transcrit a pu être confirmée dans les cellules méiotiques et post-méiotiques chez le rat, et dans le testicule chez la souris et chez l'homme. Ces protéines, produits de loci non annotés ou annotés de manière incertaine chez le rat s'expriment spécifiquement dans le testicule, et cette expression est spécifique des cellules méiotiques et post-méiotiques, ce qui a été confirmé par IHC pour la nouvelle énolase T-ENOL. Etant donné la conservation importante chez les mammifères pour

cette famille de protéines, leur prédiction de domaines « énolase » et « Longin-SNARE-like domain », ainsi que leur niveau et leur profil d'expression dans les cellules méiotiques et post-méiotiques, nous supposons qu'elles ont un rôle dans la spermiogénèse. Par ailleurs, d'autres protéines de leurs familles respectives : Longines (Hutt et al., 2005; Katafuchi et al., 2000; Steegmaier et al., 2000) et énolases (Edwards et Grootegoed, 1983; Force et al., 2004), ont des fonctions avérées dans la spermatogenèse. Il n'est pas encore possible de montrer le rôle crucial des fonctions exercées par VAMP9 et par l'énolase T-ENOL dans la spermatogenèse sans études fonctionnelles incluant l'invalidation de leur gène. Il est important de noter toutefois qu'une perte de fonction de VAMP9 qui soit délétère pour la spermatogenèse sera difficile à prouver dans la mesure où celle-ci possède une protéine paralogue, VAMP7, susceptible de compenser cette perte de fonction. Même dans le cas où une protéine n'a pas de paralogue, comme c'est cas de CLPH mise en évidence par une étude protéomique différentielle antérieure (Rolland et al., 2007; Calvel et al., 2009), le K.O. ne donne pas forcément de phénotype dans les conditions étudiées. Pour CLPH en effet, l'établissement d'une lignée de souris K.O. n'a pas permis de mettre en évidence un phénotype d'infertilité ni même une altération de la qualité spermatique. Le rôle de cette protéine reste obscur mais il est possible que sa fonction dans la spermatogenèse normale ne se manifeste que dans des conditions de stress particulières qui restent à trouver pour permettre l'observation d'un phénotype testiculaire d'intérêt. L'invalidation des gènes dans d'autres espèces comme *C.elegans* ne serait pas possible dans le cas de VAMP9 ou de la nouvelle énolase T-ENOL, car celles-ci ne sont pas conservées en dehors des mammifères. Pour l'étude du rôle de VAMP9, la génération d'un double KO chez la souris pour les gènes Vamp7 et Vamp9 serait nécessaire.

La découverte de la protéine VAMP9 qui pourrait intervenir dans le trafic membranaire au cours de la spermiogénèse comme sa paralogue VAMP7 (Sato et al., 2011a), de même que la découverte d'autres protéines du trafic mises en évidence dans notre analyse différentielle par ICPL, ouvrent de nouvelles pistes de recherche. En effet, le trafic membranaire tient un rôle important de régulation de la signalisation cellulaire. Il régit des évènements inattendus dans la signalisation cellulaire tels que les interactions entre certaines protéines comme Rab (Wang et al., 2010) ou les protéines TMED avec des récepteurs membranaires. Les protéines de la famille TMED conservées chez les eucaryotes (Jerome-Majewska et al., 2010) et qui sont impliquées dans le trafic intracellulaire (Carney et Bowen, 2004), sont nouvellement considérées, tout comme les membres la famille Rab, comme d'importants régulateurs dans le

contrôle de la signalisation immunitaire innée (Doyle et al., 2012). Notre étude protéomique différentielle sur les protéines membranaires montre que plusieurs membres de la famille TMED (TMED4, TMED9), et de la famille Rab (Rab2A, Rab11a, Rab6a) sont sur-exprimés dans les cellules germinales méiotiques et post-méiotiques, mais pas dans les corps résiduels. La sur-expression relative des protéines Rab et TMED dans les cellules germinales et leur sous-expression relative dans les CRs suggèrent un rôle d'inhibition de la signalisation inhérente à la phagocytose (signalisation associée aux Toll like récepteurs), par ces protéines. Cette inhibition pourrait faire la différence entre la signalisation en aval de la phagocytose des CRs, et celle en aval de la phagocytose des cellules germinales au sein des cellules de Sertoli. Notons que certaines protéines TMED sont par ailleurs nécessaires au développement du placenta et de l'embryon chez la souris, telles que TMED2 (Jerome-Majewska et al., 2010). Or, le placenta est un organe immuno privilégié (Kauma et al., 1999), donc les TMEDs joueraient potentiellement un rôle dans l'immunosuppression. De plus, les protéines TMED jouent leur rôle d'immuno-modulation en interagissant avec les récepteurs TAM (Doyle et al., 2012), récepteurs récemment reconnus pour leur rôle clé dans la modulation de l'immunité innée (Lemke et Rothlin, 2008), notamment dans les cellules de Sertoli (Lemke et Rothlin, 2008; Lu et al., 1999; Xiong et al., 2008), et donc dans la modulation de l'inflammation.

Mes travaux de thèse permettent, en outre, de proposer des candidats intéressants pour la compréhension des communications entre les cellules au sein de l'éphithélium séminifère, qui mériteraient des validations biochimiques. Parmi les 166 protéines différentielles membranaires qui ont été mises en évidence dans notre étude, 27 protéines sont impliquées dans la signalisation cellulaire, dont 12 n'ont jamais été mises en évidence auparavant dans le testicule. En s'intéressant aux protéines dont le rôle n'est pas directement relié à la spermatogenèse ni à la reproduction, parmi ces 166 protéines, je propose une liste de 8 candidats pour une future investigation. Aussi, seulement 2 protéines parmi les 69 nouvelles protéines (44 loci) découvertes par l'approche PIT exprimées dans les cellules méiotiques ou post-méiotiques, ont été étudiées, et seulement 3 gènes parmi ces 44 ont été clonés. Ceci ouvre encore des possibilités de caractériser de nouveaux gènes exprimés par les cellules germinales au cours de leur développement.

Dans le cadre d'une collaboration, ces travaux ont contribué à enrichir les données actuelles chez le rat. Celles-ci sont rendues disponibles sur un navigateur dédié à la visualisation des données sur la reproduction, pour le rat, l'homme et la souris : le ReproGenomics Viewer

(http://rgv.genouest.org/). Les peptides identifiés par MS/MS dans les différents types cellulaires de la lignée germinale, dans les corps résiduels et les cellules de Sertoli de rat s'affichent à leur localisation sur le génome rn4. L'accès à l'information de l'expression d'un gène au niveau du transcrit dans les cellules testiculaires isolées, et maintenant de l'expression de sa protéine, pourra faciliter la formulation d'hypothèses, orienter les analyses biochimiques de protéines candidates et donc en réduire les coûts. Ce navigateur permet de vérifier en première intention si une hypothèse est valable. Par exemple, il permet de vérifier si une interaction entre deux protéines, telle que l'interaction potentielle entre HSP90-alpha et N-WASP, est spécifique de cellules germinales en différenciation à un stade donné. L'interaction potentielle entre HSP90-alpha et N-WASP avait été mise en évidence *in silico* à partir de nos listes de protéines identifiées par LC-MS/MS dans les spermatides et les cellules de Sertoli. Une vue de HSP90-alpha à sa localisation sur RGV permet de dire que celle-ci est exprimée dans tous les types cellulaires testiculaires, donc que son interaction avec N-WASP n'est pas spécifique des spermatides et des cellules de Sertoli. Ce navigateur permet aussi de voir si des peptides identifiés par PIT correspondent à de nouveaux évènements codants : exons non annotés, UTRs, nouvelle jonction d'épissage, région intronique ou intergénique.

Au cours de cette thèse, j'ai été en mesure de caractériser des protéomes relativement exhaustifs des cellules de Sertoli, des spermatogonies, des spermatocytes pachytène, des spermatides rondes ainsi que des corps résiduels chez le rat. Les questions biologiques posées pendant cette thèse m'ont amenée à laisser de côté certains jeux de données pouvant être analysés ultérieurement au laboratoire. En effet, je me suis focalisée sur la recherche de protéines spécifiques des cellules germinales par une approche de type PIT. Notre objectif concernait la validation de nouveaux évènements codants. Avant de réaliser cette étude sur l'ensemble des types cellulaires, il convenait dans un premier temps d'en prouver l'efficacité en comparant seulement deux types cellulaires: les spermatocytes et les spermatides. L'approche protéomique shotgun sur un spectromètre LTQ-Orbitrap a fait augmenter d'un facteur 10 le nombre de protéines identifiées par type cellulaire par rapport aux années 2000. Ainsi, la 2D MS a permis au laboratoire d'obtenir un nombre d'identifications de l'ordre d'une centaine : 153 protéines dans les spermatogonies (Com et al., 2003), et 123 protéines différentielles avec un ratio >2,5 avaient été identifiées par DIGE dans les spermatogonies, spermatocytes pachytène et spermatides rondes (Rolland et al., 2007). Une étude plus fine des protéines identifiées dans ces listes n'est donc plus possible manuellement. Aujourd'hui, des étapes de présélection sont nécessaires pour orienter la recherche de candidats dans des

grands ensembles de données. Pour notre étude sur la recherche de nouveaux évènements codants, les critères de sélection des protéines candidates sont d'ordre génomique, alors que pour l'étude quantitative différentielle, la sélection est inhérente à la nécessité d'analyser des protéines significativement différentielles, et donc réalisée par l'élimination les protéines non quantifiables. Dans les deux cas, les choix étant très stricts, nous avons laissé de côté de nombreux candidats potentiels qui pourraient être retrouvés en relâchant certains filtres.

Les travaux menés au cours de cette thèse ont été réalisés en faisant appel à des technologies de spectrométrie de masse qui, de la façon dont nous les avons utilisées, ne permettent pas d'avoir accès à certaines informations biologiques. De toute évidence, la structure quaternaire, l'état de clivage ou d'oligomérisation d'un facteur qui induira des destinées cellulaires très différentes par exemple (Chaigne-Delalande et al., 2008; Tauzin et al., 2011), ainsi que certaines modifications post-traductionnelles cruciales pour la fonction d'une protéine ou pour sa sécrétion, seraient extrêmement informatifs mais n'ont pas été recherchés dans nos analyses. En conséquence, l'étude descriptive et fonctionnelle des protéines d'intérêt est nécessaire en aval de l'analyse Shotgun pour caractériser les nouvelles protéines ou celles dont la fonction est inconnue. Le point fort de la protéomique Shotgun telle que nous l'avons mise en œuvre est de permettre d'établir le protéome quasi complet des cellules étudiées (Lavigne et al., 2012), à ceci près que les protéines discrètes peuvent ne pas être détectées malgré la sensibilité et la résolution des spectromètres de masse. Le problème est encore plus important avec une analyse différentielle de type ICPL. En effet, seulement 34% des protéines que nous identifions ont été quantifiées, ce qui laisse passer de nombreux candidats potentiels. Pourtant, le pourcentage de peptides marqués peut être important, de l'ordre de 90% (Nogueira et al., 2012). Si cette technique de quantification relative vise à quantifier des protéines potentiellement impliquées dans les communications cellulaires, et donc moins abondantes que des protéines de transport par exemple, le problème de la quantification des protéines se pose. Dans notre cas, pour pallier partiellement ce problème, le protéome a été fractionné (nous avons analysé les protéines membranaires), pour rendre plus accessibles les protéines impliquées dans la signalisation cellulaire.

Dans l'étude visant à découvrir de nouveaux évènements codants par une approche PIT, nous montrons que l'approche utilisée a des limitations inhérentes à l'emploi d'une grande base de donnée de séquences traduites des transcrits assemblés. En particulier, le fait que cette base contienne des séquences erronées ou écourtées peut biaiser l'identification de nouvelles isoformes potentielles si l'information d'un peptide spécifique de cette isoforme n'est pas en

notre possession. En effet, des peptides identifiés sur des exons communs à plusieurs isoformes ne donnent pas d'information sur une nouvelle jonction d'épissage ou sur un exon distal, et donnent lieu à des identifications ambiguës. On peut penser que les inconvénients de cette approche conduisant principalement à manquer des évènements potentiellement intéressants, vont tendre à s'amoindrir étant donné qu'ils ont pour origine des erreurs de séquençage, des problèmes d'assemblage des transcrits, et la taille des banques de séquences traduites. L'utilisation d'un protocole de séquençage « brin spécifique » des ARNs permettrait déjà de diminuer la taille des banques de données, car dans ce cas l'orientation des transcrits est conservée dans les librairies de cDNA (Cloonan et al., 2008). D'autre part, certains auteurs ont déjà réfléchi sur les possibilités de compacter les bases de données de séquence générées à partir de larges ensembles de données de RNA-seq (Woo et al., 2014). On peut aussi compter sur le fait que les protocoles de séquençage *de novo* à haut débit vont générer de moins en moins d'erreurs de séquençage (erreurs pendant le « base calling »), et d'assemblage des reads (fragments) de séquençage. Dans l'ensemble, la génération de banques personnalisées pour la protéogénomique, étant donné les écueils discutés dans les deux premiers chapitres, requiert des compétences pointues et une expérience dans deux domaines d'expertise : celui de la protéomique et celui de la transcriptomique. De telles approches avant d'être automatisées et appliquées à de grands projets doivent donc encore faire l'objet de mises au point.

Les perspectives ouvertes par ce travail de thèse sont nombreuses. En effet, le potentiel de découverte de nouveaux évènements codants est important dans les cellules germinales. L'approche PIT a aussi été appliquée selon les mêmes méthodes, sur d'autres types cellulaires testiculaires chez le rat : les cellules de Sertoli et les spermatogonies. Les données de protéomique générées (non présentées) possèdent les mêmes proportions d'évènements non connus (intergéniques et introniques), identifiés par spectrométrie de masse, à savoir 4 à 8 % des identifications de protéines sur la base de données « non redundant rat reference proteome ». Les filtrations appliquées selon les critères d'expression des transcrits et de leur conservation sont appliquées comme nous le présentons dans l'article du chapitre 1. Une sélection préliminaire donne lieu à la mise en évidence d'autres candidats méritant une analyse plus approfondie. Parmi eux, on peut citer le candidat intronique TCONS_00037788 dans la région : chr15:21,533,944..21,535,819 du gène codant Ddhd1 sur le génome rn4 du rat. Le transcrit de cette protéine candidate présente un profil d'expression spécifique des spermatogonies. Ce transcrit est codant dans le sens positif, alors que le gène

Ddhd1, dont un intron se trouve à cette localisation, est codant dans le sens inverse. Il s'agit donc bien d'un nouvel événement intronique codant.

Etant donné le nombre de gènes spécifiquement exprimés dans les cellules germinales, nous pouvons imaginer, à l'avenir, qu'avec l'amélioration des protocoles de séquençage et de l'assemblage du transcriptome, ainsi qu'avec les efforts d'amélioration des banques de données pour la protéogénomique, la confirmation du potentiel codant de nombreux évènements par spectrométrie de masse dans la lignée germinale va augmenter de manière significative (Brosch et al., 2011). Chez la souris, 26.000 gènes s'exprimeraient dans le testicule pendant la première vague de spermatogenèse et près de 950 lncRNAs s'exprimeraient de manière spécifique dans les cellules germinales à différents stades de différenciation (Laiho et al., 2013). Chez le rat, plus de 1.400 transcrits testiculaires non annotés et qui partagent les propriétés des lncRNAs s'accumulent dans les cellules germinales au cours de la méiose (Chalmel et al., 2014). Or nous venons de montrer que certains lncRNAs sont en réalité codants (Chocu et al., 2014). Notre contribution à la réannotation du génome du rat, du fait des nouveaux gènes « germinaux » mis en évidence, pourra dans le futur être significative.

Une future étude pourrait concerner la recherche de nouvelles isoformes s'exprimant spécifiquement dans les cellules germinales et/ou dans les cellules de Sertoli. Elles constituent en effet plus de 25 % des identifications sur la banque « non redundant rat reference proteome », et ce pour tous les types cellulaires étudiés. Ces nouvelles protéines correspondent à des séquences déduites de la catégorie des 11.837 transcrits de nouvelles isoformes potentielles, découverts par Chalmel et collaborateurs dans leur étude récente du transcriptome des cellules germinales mâles (Chalmel et al., 2014). Les nouvelles isoformes potentielles identifiées dans tous ces types cellulaires pourraient mettre en évidence des évènements d'épissage spécifiques de chaque stade de différenciation étudié. Nous pouvons peut-être espérer découvrir quelques centaines de nouvelles isoformes au cours de la différenciation des cellules germinales mâles, comme il a été rapporté chez la souris (Margolin et al., 2014). Par contre, il faut rester conscients des écueils de l'approche PIT pour l'identification de nouvelles isoformes réelles. Ces écueils consistent en la présence d'ORFs raccourcis sur des séquences écourtées et / ou erronées. Le résultat en est une identification avec un meilleur score que la séquence canonique de la protéine. Certains ajustements seront nécessaires pour utiliser au mieux une banque personnalisée dédiée à la découverte de

nouvelles isoformes. Eliminer des séquences protéiques déduites de nouveaux transcrits qui s'arrêtent avant un stop, et donc suspectes, pourrait régler en partie le problème.

Une étude qui consisterait en la comparaison des profils d'expression des protéines au cours de la spermatogenèse *versus* l'expression de leur transcrit, telle que réalisée par Gan et collaborateurs chez la souris (Gan et al., 2013) peut être envisagée afin de continuer d'explorer les mécanismes de régulation post-transcriptionnelle au cours de la spermatogenèse. Elle serait basée sur les données de quantification des transcrits par RNA-seq, et sur une quantification des protéines présentes dans les cellules germinales. Par contre, une méthode de quantification utilisant des marquages isotopiques métaboliques comme le SILAC (Ong et al., 2002) ne peut pas être utilisée avec des cellules germinales que nous ne pouvons pas cultiver et qui ne peuvent pas métaboliser les acides aminés alourdis. Une méthode de « label free » telle que le spectral count (Muntel et al., 2012) serait alors la plus appropriée afin d'obtenir des données de quantification des protéines à un niveau le plus proche possible de celui obtenu pour les transcrits, bien qu'il ne soit pas possible d'accéder à la « quantification du protéome » comme il est possible d'accéder à la quantification du transcriptome.

Le projet de séquençage du génome du rat et les ressources qu'il a générées permettent la traduction de la biologie du rat à la médecine humaine, car la triangulation entre les trois génomes souris, homme, rat, est possible et mutuellement informative. Depuis le premier projet de séquençage du génome du rat (Gibbs et al., 2004), les chercheurs ont utilisé sa séquence, ses annotations et les ressources associés, au travers d'études protéomiques appliquées à des processus biologiques. Ces annotations ne sont pas encore complètes de nos jours en ce qui concerne le génome du rat. Mais les avancées des technologies de séquençage à haut débit comme le « RNA sequencing » ont permis et permettront d'accéder à une vue globale du programme d'expression des gènes au cours d'un processus biologique donné avec de plus en plus de précision. Nos travaux ont contribué à l'amélioration de la connaissance du génome de cet organisme modèle en prouvant qu'un certain nombre d'ARNs considérés comme non codants, codent finalement pour des protéines. Comme en témoignent nos découvertes, par exemple l'erreur d'assemblage au niveau du gène Wdr62 sur le génome rn4 du rat, celui-ci est loin d'être parfaitement bien annoté.

Chez l'homme, la découverte de nouveaux évènements codants par une approche de type PIT pourrait être étendue à des projets internationaux visant à étudier le protéome humain, tels que le « Chromosome centric Human Proteome Project » (le C-HPP), qui vise à annoter

toutes les protéines codées par les gènes connus sur chaque chromosome humain (Paik et al., 2012). Le HPP est une initiative de HUPO (Human Proteome Organization) qui a pour objectif de promouvoir la protéomique à travers des collaborations internationales en favorisant le développement de nouvelles technologies et de techniques afin de mieux comprendre les pathologies humaines. Un projet en cours au laboratoire s'inscrit dans C-HPP et consiste à trouver les protéines manquantes / mal annotées dans les bases de données dont l'expression est spécifique du testicule. Il y aurait 3.844 « missing » protéines humaines de la base NexProt qui n'auraient jamais été mises en évidence par spectrométrie de masse ou par détection avec un anticorps ou qui ont une documentation inadéquate (Lane et al., 2014), dont près de 1.000 s'exprimeraient dans le testicule. La tâche prise en charge par le laboratoire est de mettre en évidence par spectrométrie de masse celles qui s'expriment dans le testicule, et correspondent aux chromosomes 2 et 14. En revanche, les actions en cours ne concernent que les protéines connues ou prédites, et une approche de type PIT pourrait permettre d'en identifier de nouvelles afin de compléter l'annotation du génome humain, et de constituer une nouvelle source pour la recherche biomédicale. C'est précisément ce à quoi se sont attachées les équipes internationales de Akhilesh Pandey et de Bernhard Küster dans le projet « The Human proteome Map » (Kim et al., 2014 ; Wilhelm et al., 2014).

En conclusion, les travaux réalisés au cours de cette thèse ont permis, grâce à la spectrométrie de masse et des stratégies omiques intégratives, la découverte de 166 protéines potentiellement impliquées dans la communication entre les cellules germinales et les cellules de Sertoli pendant la spermiogénèse ou dans les évènements liés à la formation et à la phagocytose des corps résiduels. Ils ont aussi permis la découverte de 44 nouveaux gènes codants potentiellement impliqués dans la spermiogénèse. Outre ces découvertes, ces travaux ont permis de générer des données de qualité sur le protéome des cellules germinales, des cellules de Sertoli et des corps résiduels chez le rat, et de mutualiser visuellement cet ensemble important de données sur RGV. Ces données offrent un grand potentiel de recherche future selon plusieurs axes : la recherche de nouvelles isoformes spécifiques des cellules germinales, la recherche de nouveaux évènements codants sur l'ensemble de la spermatogenèse, et la recherche de protéines partenaires entre les cellules de Sertoli et les cellules germinales.

RÉFÉRENCES

Abdolzade-Bavil, A., Hayes, S., Goretzki, L., Kröger, M., Anders, J., and Hendriks, R. (2004). Convenient and versatile subcellular extraction procedure, that facilitates classical protein expression profiling and functional protein analysis. PROTEOMICS *4*, 1397–1405.

Abel, M.H., Baker, P.J., Charlton, H.M., Monteiro, A., Verhoeven, G., De Gendt, K., Guillou, F., and O'Shaughnessy, P.J. (2008). Spermatogenesis and sertoli cell activity in mice lacking sertoli cell receptors for follicle-stimulating hormone and androgen. Endocrinology *149*, 3279–3285.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185–2195.

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.

Agami, R. (2010). microRNAs, RNA binding proteins and cancer. Eur. J. Clin. Invest. *40*, 370–374.

Ahmed, E.A., and de Rooij, D.G. (2009). Staging of mouse seminiferous tubule cross-sections. Methods Mol. Biol. Clifton NJ *558*, 263–277.

Alban, A., David, S.O., Bjorkesten, L., Andersson, C., Sloge, E., Lewis, S., and Currie, I. (2003). A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. Proteomics *3*, 36–44.

Amanai, M., Brahmajosyula, M., and Perry, A.C.F. (2006). A restricted role for sperm-borne microRNAs in mammalian fertilization. Biol. Reprod. *75*, 877–884.

Amann, R.P. (2008). The cycle of the seminiferous epithelium in humans: a need to revisit? J. Androl. *29*, 469–487.

Andéol, Y. (1996). Les premières expressions du génome embryonnaire au cours du developpement chez différentes espèces animales. Médecine/sciences *12*, 192.

Anderson, P., and Kedersha, N. (2006). RNA granules. J. Cell Biol. *172*, 803–808.

Anderson, P., and Kedersha, N. (2008). Stress granules: the Tao of RNA triage. Trends Biochem. Sci. *33*, 141–150.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796–815.

Aravin, A.A., and Bourc'his, D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev. *22*, 970–975.

Armengaud, J. (2009). A perfect genome annotation is within reach with the proteomics and genomics alliance. Curr. Opin. Microbiol. *12*, 292–300.

Armengaud, J., Marie Hartmann, E., and Bland, C. (2013). Proteogenomics for environmental microbiology. Proteomics *13*, 2731–2742.

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., and Hartmann, E.M. (2014). Non-model organisms, a species endangered by proteogenomics. J. Proteomics *105*, 5–18.

Austin, R.J., Chang, D.K., Holstein, C.A., Lee, L.W., Risler, J., Wang, J.H., Adams, L., Krusberski, N.B., and Martin, D.B. (2012). IQcat: multiplexed protein quantification by isoelectric QconCAT. Proteomics *12*, 2078–2083.

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008). Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. Science *320*, 938–941.

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. (2005). Regulation by let-7 and lin-4 miRNAs Results in Target mRNA Degradation. Cell *122*, 553–563.

Bagheri-Fam, S., Sinclair, A.H., Koopman, P., and Harley, V.R. (2010). Conserved regulatory modules in the Sox9 testis-specific enhancer predict roles for SOX, TCF/LEF, Forkhead, DMRT, and GATA proteins in vertebrate sex determination. Int. J. Biochem. Cell Biol. *42*, 472–477.

Bairoch, A., and Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res. *27*, 49–54.

Bairoch, A., and Boeckmann, B. (1993). The SWISS-PROT protein sequence data bank, recent developments. Nucleic Acids Res. *21*, 3093–3096.

Bánfai, B., Jia, H., Khatun, J., Wood, E., Risk, B., Gundling, W.E., Jr, Kundaje, A., Gunawardena, H.P., Yu, Y., Xie, L., et al. (2012). Long noncoding RNAs are rarely translated in two human cell lines. Genome Res. *22*, 1646–1657.

Bao, J., Wu, J., Schuster, A.S., Hennig, G.W., and Yan, W. (2013). Expression profiling reveals developmentally regulated lncRNA repertoire in the mouse male germline. Biol. Reprod. *89*, 107.

Barlow, D.P., Stöger, R., Herrmann, B.G., Saito, K., and Schweifer, N. (1991). The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature *349*, 84–87.

Baudet, M., Ortet, P., Gaillard, J.-C., Fernandez, B., Guérin, P., Enjalbal, C., Subra, G., Groot, A. de, Barakat, M., Dedieu, A., et al. (2010). Proteomics-based Refinement of Deinococcus deserti Genome Annotation Reveals an Unwonted Use of Non-canonical Translation Initiation Codons. Mol. Cell. Proteomics *9*, 415–426.

Bettegowda, A., and Smith, G.W. (2007). Mechanisms of maternal mRNA regulation: implications for mammalian early embryonic development. Front. Biosci. J. Virtual Libr. *12*, 3713–3726.

Bettegowda, A., and Wilkinson, M.F. (2010). Transcription and post-transcriptional regulation of spermatogenesis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *365*, 1637–1651.

Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.I., Rindone, W.P., Swindell, C.D., and Tung, C.S. (1986). The GenBank genetic sequence databank. Nucleic Acids Res. *14*, 1–4.

Blanco-Rodríguez, J., and Martínez-García, C. (1999). Apoptosis Is Physiologically Restricted to a Specialized Cytoplasmic Compartment in Rat Spermatids. Biol. Reprod. *61*, 1541–1547.

Boskovic, A., and Torres-Padilla, M.-E. (2013). How mammals pack their sperm: a variant matter. Genes Dev. *27*, 1635–1639.

Bozas, S.E., Kirszbaum, L., Sparrow, R.L., and Walker, I.D. (1993). Several vascular complement inhibitors are present on human sperm. Biol. Reprod. *48*, 503–511.

Brahmaraju, M., Shoeb, M., Laloraya, M., and Kumar, P.G. (2004). Spatio-temporal organization of Vam6P and SNAP on mouse spermatozoa and their involvement in sperm-zona pellucida interactions. Biochem. Biophys. Res. Commun. *318*, 148–155.

Bremner, W.J., Millar, M.R., Sharpe, R.M., and Saunders, P.T. (1994). Immunohistochemical localization of androgen receptors in the rat testis: evidence for stage-dependent expression and regulation by androgens. Endocrinology *135*, 1227–1234.

Breucker, H., Schäfer, E., and Holstein, A.F. (1985). Morphogenesis and fate of the residual body in human spermiogenesis. Cell Tissue Res. *240*, 303–309.

Brosch, M., Saunders, G.I., Frankish, A., Collins, M.O., Yu, L., Wright, J., Verstraten, R., Adams, D.J., Harrow, J., Choudhary, J.S., et al. (2011). Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. Genome Res. *21*, 756–767.

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. *268*, 78–94.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Cacciola, G., Chioccarelli, T., Fasano, S., Pierantoni, R., and Cobellis, G. (2013). Estrogens and Spermiogenesis: New Insights from Type 1 Cannabinoid Receptor Knockout Mice. Int. J. Endocrinol. *2013*, 501350.

Cagney, G., Park, S., Chung, C., Tong, B., O'Dushlaine, C., Shields, D.C., and Emili, A. (2005). Human Tissue Profiling with Multidimensional Protein Identification Technology. J Proteome Res *4*, 1757–1767.

Calvel, P., Kervarrec, C., Lavigne, R., Vallet-Erdtmann, V., Guerrois, M., Rolland, A.D., Chalmel, F., Jégou, B., and Pineau, C. (2009). CLPH, a novel casein kinase 2-phosphorylated disordered protein, is specifically associated with postmeiotic germ cells in rat spermatogenesis. J. Proteome Res. *8*, 2953–2965.

Calvel, P., Rolland, A.D., Jégou, B., and Pineau, C. (2010). Testicular postgenomics: targeting the regulation of spermatogenesis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *365*, 1481–1500.

Campese, A.F., Grazioli, P., de Cesaris, P., Riccioli, A., Bellavia, D., Pelullo, M., Padula, F., Noce, C., Verkhovskaia, S., Filippini, A., et al. (2014). Mouse Sertoli cells sustain de novo generation of regulatory T cells by triggering the notch pathway through soluble JAGGED1. Biol. Reprod. *90*, 53.

Cañas, B., Piñeiro, C., Calvo, E., López-Ferrer, D., and Gallardo, J.M. (2007). Trends in sample preparation for classical and second generation proteomics. J. Chromatogr. A *1153*, 235–258.

Carmell, M.A., Girard, A., van de Kant, H.J.G., Bourc'his, D., Bestor, T.H., de Rooij, D.G., and Hannon, G.J. (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev. Cell *12*, 503–514.

Carney, G.E., and Bowen, N.J. (2004). p24 proteins, intracellular trafficking, and behavior: Drosophila melanogaster provides insights and opportunities. Biol. Cell Auspices Eur. Cell Biol. Organ. *96*, 271–278.

Carreau, S., and Hess, R.A. (2010). Oestrogens and spermatogenesis. Philos. Trans. R. Soc. B Biol. Sci. *365*, 1517–1535.

Carreau, S., Bouraima-Lelong, H., and Delalande, C. (2012). Estrogen, a female hormone involved in spermatogenesis. Adv. Med. Sci. *57*, 31–36.

Castellana, N., and Bafna, V. (2010). Proteogenomics to discover the full coding content of genomes: a computational perspective. J. Proteomics *73*, 2124–2135.

Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S.P. (2008). Discovery and revision of Arabidopsis genes by proteogenomics. Proc. Natl. Acad. Sci. U. S. A. *105*, 21034–21038.

Catherman, A.D., Skinner, O.S., and Kelleher, N.L. (2014). Top Down proteomics: facts and perspectives. Biochem. Biophys. Res. Commun. *445*, 683–693.

Catizone, A., Ricci, G., Caruso, M., Ferranti, F., Canipari, R., and Galdieri, M. (2012). Hepatocyte growth factor (HGF) regulates blood-testis barrier (BTB) in adult rats. Mol. Cell. Endocrinol. *348*, 135–146.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012–2018.

Chaigne-Delalande, B., Moreau, J.-F., and Legembre, P. (2008). Rewinding the DISC. Arch. Immunol. Ther. Exp. (Warsz.) *56*, 9–14.

Chaineau, M., Danglot, L., and Galli, T. (2009). Multiple roles of the vesicular-SNARE TI-VAMP in post-Golgi and endosomal trafficking. FEBS Lett. *583*, 3817–3826.

Chalmel, F., and Primig, M. (2008). The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. BMC Bioinformatics *9*, 86.

Chalmel, F., Rolland, A.D., Niederhauser-Wiederkehr, C., Chung, S.S.W., Demougin, P., Gattiker, A., Moore, J., Patard, J.-J., Wolgemuth, D.J., Jégou, B., et al. (2007a). The conserved transcriptome in human and rodent male gametogenesis. Proc. Natl. Acad. Sci. U. S. A. *104*, 8346–8351.

Chalmel, F., Lardenois, A., and Primig, M. (2007b). Toward Understanding the Core Meiotic Transcriptome in Mammals and Its Implications for Somatic Cancer. Ann. N. Y. Acad. Sci. *1120*, 1–15.

Chalmel, F., Lardenois, A., Evrard, B., Mathieu, R., Feig, C., Demougin, P., Gattiker, A., Schulze, W., Jégou, B., Kirchhoff, C., et al. (2012). Global human tissue profiling and protein network analysis reveals distinct levels of transcriptional germline-specificity and identifies target genes for male infertility. Hum. Reprod. *27*, 3233–3248.

Chalmel, F., Lardenois, A., Evrard, B., Rolland, A.D., Sallou, O., Dumargne, M.-C., Coiffec, I., Collin, O., Primig, M., and Jégou, B. (2014). High-Resolution Profiling of Novel Transcribed Regions During Rat Spermatogenesis. Biol. Reprod. biolreprod.114.118166.

Chemes, H. (1986). The phagocytic function of Sertoli cells: a morphological, biochemical, and endocrinological study of lysosomes and acid phosphatase localization in the rat testis. Endocrinology *119*, 1673–1681.

Chen, C., Vincent, O., Jin, J., Weisz, O.A., and Montelaro, R.C. (2005). Functions of early (AP-2) and late (AIP1/ALIX) endocytic proteins in equine infectious anemia virus budding. J. Biol. Chem. *280*, 40474–40480.

Chendrimada, T.P., Finn, K.J., Ji, X., Baillat, D., Gregory, R.I., Liebhaber, S.A., Pasquinelli, A.E., and Shiekhattar, R. (2007). MicroRNA silencing through RISC recruitment of eIF6. Nature *447*, 823–828.

Cheng, C.Y., and Mruk, D.D. (2012). The Blood-Testis Barrier and Its Implications for Male Contraception. Pharmacol. Rev. *64*, 16–64.

Cheng, J., Watkins, S.C., and Walker, W.H. (2007). Testosterone activates mitogen-activated protein kinase via Src kinase and the epidermal growth factor receptor in sertoli cells. Endocrinology *148*, 2066–2074.

Chimento, A., Sirianni, R., Delalande, C., Silandre, D., Bois, C., Andò, S., Maggiolini, M., Carreau, S., and Pezzi, V. (2010). 17 beta-estradiol activates rapid signaling pathways involved in rat pachytene spermatocytes apoptosis through GPR30 and ER alpha. Mol. Cell. Endocrinol. *320*, 136–144.

Chocu, S., Calvel, P., Rolland, A.D., and Pineau, C. (2012). Spermatogenesis in mammals: proteomic insights. Syst. Biol. Reprod. Med. *58*, 179–190.

Chocu, S., Evrard, B., Lavigne, R., Rolland, A.D., Aubry, F., Jégou, B., Chalmel, F., and Pineau, C. (2014). Forty-Four Novel Protein-Coding Loci Discovered Using a Proteomics Informed by Transcriptomics (PIT) Approach in Rat Male Germ Cells. Biol. Reprod.

Chromy, B.A., Gonzales, A.D., Perkins, J., Choi, M.W., Corzett, M.H., Chang, B.C., Corzett, C.H., and McCutchen-Maloney, S.L. (2004). Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins. J. Proteome Res. *3*, 1120–1127.

Chuma, S., Hosokawa, M., Tanaka, T., and Nakatsuji, N. (2009). Ultrastructural characterization of spermatogenesis and its evolutionary conservation in the germline: germinal granules in mammals. Mol. Cell. Endocrinol. *306*, 17–23.

Clermont, Y. (1972). Kinetics of spermatogenesis in mammals: seminiferous epithelium cycle and spermatogonial renewal. Physiol. Rev. *52*, 198–236.

Cloonan, N., Forrest, A.R.R., Kolle, G., Gardiner, B.B.A., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods *5*, 613–619.

Com, E., Evrard, B., Roepstorff, P., Aubry, F., and Pineau, C. (2003). New insights into the rat spermatogonial proteome: identification of 156 additional proteins. Mol. Cell. Proteomics MCP *2*, 248–261.

Com, E., Rolland, A.D., Guerrois, M., Aubry, F., Jégou, B., Vallet-Erdtmann, V., and Pineau, C. (2006). Identification, molecular cloning, and cellular distribution of the rat homolog of minichromosome maintenance protein 7 (MCM7) in the rat testis. Mol. Reprod. Dev. *73*, 866–877.

Com, E., Melaine, N., Chalmel, F., and Pineau, C. (2014). Proteomics and integrative genomics for unraveling the mysteries of spermatogenesis: The strategies of a team. J. Proteomics.

Costoya, J.A., Hobbs, R.M., Barna, M., Cattoretti, G., Manova, K., Sukhwani, M., Orwig, K.E., Wolgemuth, D.J., and Pandolfi, P.P. (2004). Essential role of Plzf in maintenance of spermatogonial stem cells. Nat. Genet. *36*, 653–659.

Cottrell, J.S. (2011). Protein identification using MS/MS data. J. Proteomics *74*, 1842–1851.

Cupp, A.S. (2014). Sertoli Cell Based Gene Therapy? Biol. Reprod.

Dadoune, J.-P., and Démoulin, A. (1991). Structure et fonctions du testicule. 221–250.

Dastig, S., Nenicu, A., Otte, D.M., Zimmer, A., Seitz, J., Baumgart-Vogt, E., and Lüers, G.H. (2011). Germ cells of male mice express genes for peroxisomal metabolic pathways implicated in the regulation of spermatogenesis and the protection against oxidative stress. Histochem. Cell Biol. *136*, 413–425.

DeChiara, T.M., Efstratiadis, A., and Robertson, E.J. (1990). A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. Nature *345*, 78–80.

DeGracia, D.J., Jamison, J.T., Szymanski, J.J., and Lewis, M.K. (2008). Translation arrest and ribonomics in post-ischemic brain: layers and layers of players. J. Neurochem. *106*, 2288–2301.

Delfino, F.J., and Walker, W.H. (1999). NF-kappaB induces cAMP-response element-binding protein gene transcription in sertoli cells. J. Biol. Chem. *274*, 35607–35613.

Deng, W., and Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev. Cell *2*, 819–830.

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., et al. (2008). Annotating genomes with massive-scale RNA sequencing. Genome Biol. *9*, R175.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Dicker, L., Lin, X., and Ivanov, A.R. (2010). Increased power for the analysis of label-free LC-MS/MS proteomics data by combining spectral counts and peptide peak attributes. Mol. Cell. Proteomics MCP *9*, 2704–2718.

Van Dijk, E.L., Schilders, G., and Pruijn, G.J.M. (2007). Human cell growth requires a functional cytoplasmic exosome, which is involved in various mRNA decay pathways. RNA N. Y. N *13*, 1027–1035.

Djureinovic, D., Fagerberg, L., Hallström, B., Danielsson, A., Lindskog, C., Uhlén, M., and Pontén, F. (2014). The human testis-specific proteome defined by transcriptomics and antibody-based profiling. Mol. Hum. Reprod.

Dong, L.Q., Ramos, F.J., Wick, M.J., Lim, M.A., Guo, Z., Strong, R., Richardson, A., and Liu, F. (2002). Cloning and characterization of a testis and brain-specific isoform of mouse 3'-phosphoinositide-dependent protein kinase-1, mPDK-1 beta. Biochem. Biophys. Res. Commun. *294*, 136–144.

Dorrington, J.H., Fritz, I.B., and Armstrong, D.T. (1978). Control of Testicular Estrogen Synthesis. Biol. Reprod. *18*, 55–64.

Doyle, S.L., Husebye, H., Connolly, D.J., Espevik, T., O'Neill, L.A.J., and McGettrick, A.F. (2012). The GOLD domain-containing protein TMED7 inhibits TLR4 signalling from the endosome upon LPS stimulation. Nat. Commun. *3*, 707.

Drabovich, A.P., Jarvi, K., and Diamandis, E.P. (2011). Verification of Male Infertility Biomarkers in Seminal Plasma by Multiplex Selected Reaction Monitoring Assay. Mol. Cell. Proteomics MCP *10*.

Dugast, I., and Jégou, B. (1994). [Cytokines and Sertoli cell and germ cell interactions]. Contracept. Fertil. Sex. 1992 *22*, 631–634.

Dupuis, A., Hennekinne, J.-A., Garin, J., and Brun, V. (2008). Protein Standard Absolute Quantification (PSAQ) for improved investigation of staphylococcal food poisoning outbreaks. Proteomics *8*, 4633–4636.

Dym, M., and Fawcett, D.W. (1971). Further Observations on the Numbers of Spermatogonia, Spermatocytes, and Spermatids Connected by Intercellular Bridges in the Mammalian Testis. Biol. Reprod. *4*, 195–215.

Eddy, E.M. (2002). Male germ cell gene expression. Recent Prog. Horm. Res. *57*, 103–128.

Edwards, Y.H., and Grootegoed, J.A. (1983). A sperm-specific enolase. J. Reprod. Fertil. *68*, 305–310.

Elias, J.E., and Gygi, S.P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat. Methods *4*, 207–214.

Elliott, D.J., Venables, J.P., Newton, C.S., Lawson, D., Boyle, S., Eperon, I.C., and Cooke, H.J. (2000). An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. Hum. Mol. Genet. *9*, 2117–2124.

Elliott, M.R., Zheng, S., Park, D., Woodson, R.I., Reardon, M.A., Juncadella, I.J., Kinchen, J.M., Zhang, J., Lysiak, J.J., and Ravichandran, K.S. (2010). Unexpected requirement for ELMO1 in clearance of apoptotic germ cells in vivo. Nature *467*, 333–337.

Emes, A.V., Latner, A.L., and Martin, J.A. (1975). Electrofocusing followed by gradient electrophoresis: a two-dimensional polyacrylamide gel technique for the separation of proteins and its application to the immunoglobulins. Clin. Chim. Acta Int. J. Clin. Chem. *64*, 69–78.

Emmert, D.B., Stoehr, P.J., Stoesser, G., and Cameron, G.N. (1994). The European Bioinformatics Institute (EBI) databases. Nucleic Acids Res. *22*, 3445–3449.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Eulalio, A., Behm-Ansmant, I., and Izaurralde, E. (2007). P bodies: at the crossroads of post-transcriptional pathways. Nat. Rev. Mol. Cell Biol. *8*, 9–22.

Evans, E.B., Hogarth, C.A., Mitchell, D., and Griswold, M.D. (2014). Riding the Spermatogenic Wave: Profiling Gene Expression Within Neonatal Germ and Sertoli Cells During a Synchronized Initial Wave of Spermatogenesis in Mice. Biol. Reprod.

Evans, V.C., Barker, G., Heesom, K.J., Fan, J., Bessant, C., and Matthews, D.A. (2012). De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat. Methods *9*, 1207–1211.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. Science *246*, 64–71.

Fialka, I., Pasquali, C., Lottspeich, F., Ahorn, H., and Huber, L.A. (1997). Subcellular fractionation of polarized epithelial cells and identification of organelle-specific proteins by two-dimensional gel electrophoresis. Electrophoresis *18*, 2582–2590.

Filippini, F., Rossi, V., Galli, T., Budillon, A., D'Urso, M., and D'Esposito, M. (2001). Longins: a new evolutionary conserved VAMP family sharing a novel SNARE domain. Trends Biochem. Sci. *26*, 407–409.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222–D230.

Fix, C., Jordan, C., Cano, P., and Walker, W.H. (2004). Testosterone activates mitogen-activated protein kinase and the cAMP response element binding protein transcription factor in Sertoli cells. Proc. Natl. Acad. Sci. U. S. A. *101*, 10919–10924.

Flenkenthaler, F., Windschüttl, S., Fröhlich, T., Schwarzer, J.U., Mayerhofer, A., and Arnold, G.J. (2014). Secretome analysis of testicular peritubular cells: a window into the human testicular microenvironment and the spermatogonial stem cell niche in man. J. Proteome Res. *13*, 1259–1269.

Flowerdew, S.E., and Burgoyne, R.D. (2009). A VAMP7/Vti1a SNARE complex distinguishes a non-conventional traffic route to the cell surface used by KChIP1 and Kv4 potassium channels. Biochem. J. *418*, 529–540.

Force, A., Viallard, J.-L., Grizard, G., and Boucher, D. (2002). Enolase isoforms activities in spermatozoa from men with normospermia and abnormospermia. J. Androl. *23*, 202–210.

Force, A., Viallard, J.-L., Saez, F., Grizard, G., and Boucher, D. (2004). Electrophoretic characterization of the human sperm-specific enolase at different stages of maturation. J. Androl. *25*, 824–829.

Fournier, M.L., Gilmore, J.M., Martin-Brown, S.A., and Washburn, M.P. (2007). Multidimensional separations-based shotgun proteomics. Chem. Rev. *107*, 3654–3686.

França, L.R., Ogawa, T., Avarbock, M.R., Brinster, R.L., and Russell, L.D. (1998). Germ Cell Genotype Controls Cell Cycle during Spermatogenesis in the Rat. Biol. Reprod. *59*, 1371–1377.

Gan, H., Cai, T., Lin, X., Wu, Y., Wang, X., Yang, F., and Han, C. (2013). Integrative proteomic and transcriptomic analyses reveal multiple post-transcriptional regulatory mechanisms of mouse spermatogenesis. Mol. Cell. Proteomics.

Gao, F., Maiti, S., Alam, N., Zhang, Z., Deng, J.M., Behringer, R.R., Lécureuil, C., Guillou, F., and Huff, V. (2006). The Wilms tumor gene, Wt1, is required for Sox9 expression and maintenance of tubular architecture in the developing testis. Proc. Natl. Acad. Sci. U. S. A. *103*, 11987–11992.

De Gendt, K., Swinnen, J.V., Saunders, P.T.K., Schoonjans, L., Dewerchin, M., Devos, A., Tan, K., Atanassova, N., Claessens, F., Lécureuil, C., et al. (2004). A Sertoli cell-selective knockout of the androgen receptor causes spermatogenic arrest in meiosis. Proc. Natl. Acad. Sci. U. S. A. *101*, 1327–1332.

De Gendt, K., Verhoeven, G., Amieux, P.S., and Wilkinson, M.F. (2014). Research Resource: Genome-Wide Identification of AR-Regulated Genes Translated in Sertoli Cells In Vivo Using the RiboTag Approach. Mol. Endocrinol. Baltim. Md *28*, 575–591.

Gene Ontology Consortium, Blake, J.A., Dolan, M., Drabkin, H., Hill, D.P., Li, N., Sitnikov, D., Bridges, S., Burgess, S., Buza, T., et al. (2013). Gene Ontology annotations and resources. Nucleic Acids Res. *41*, D530–D535.

George, D.G., Barker, W.C., and Hunt, L.T. (1986). The protein identification resource (PIR). Nucleic Acids Res. *14*, 11–15.

Gérard, N., and Jégou, B. (1993). In-vitro influence of germ cells on Sertoli cell-secreted proteins: a two-dimensional gel electrophoresis analysis. Int. J. Androl. *16*, 285–291.

Gérard, N., Syed, V., and Jégou, B. (1992b). Lipopolysaccharide, latex beads and residual bodies are potent activators of Sertoli cell interleukin-1 alpha production. Biochem. Biophys. Res. Commun. *185*, 154–161.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493–521.

Gitlits, V.M., Toh, B.H., Loveland, K.L., and Sentry, J.W. (2000). The glycolytic enzyme enolase is present in sperm tail and displays nucleotide-dependent association with microtubules. Eur. J. Cell Biol. *79*, 104–111.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546, 563–567.

Gómez, M., Manzano, A., Figueras, A., Viñals, F., Ventura, F., Rosa, J.L., Bartrons, R., and Navarro-Sabaté, À. (2012). Sertoli-secreted FGF-2 induces PFKFB4 isozyme expression in mouse spermatogenic cells by activation of the MEK/ERK/CREB pathway. Am. J. Physiol. Endocrinol. Metab. *303*, E695–E707.

González-Iglesias, H., Álvarez, L., García, M., Escribano, J., Rodríguez-Calvo, P.P., Fernández-Vega, L., and Coca-Prados, M. (2014). Comparative proteomic study in serum of patients with primary open-angle glaucoma and pseudoexfoliation glaucoma. J. Proteomics *98*, 65–78.

Gottardo, M., Callaini, G., and Riparbelli, M.G. (2013). The cilium-like region of the Drosophila spermatocyte: an emerging flagellum? J. Cell Sci. *126*, 5441–5452.

Govin, J., Gaucher, J., Ferro, M., Debernardi, A., Garin, J., Khochbin, S., and Rousseaux, S. (2012). Proteomic strategy for the identification of critical actors in reorganization of the post-meiotic male genome. Mol. Hum. Reprod. *18*, 1–13.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. *9*, r43–r74.

Grellscheid, S., Dalgliesh, C., Storbeck, M., Best, A., Liu, Y., Jakubik, M., Mende, Y., Ehrmann, I., Curk, T., Rossbach, K., et al. (2011). Identification of evolutionarily conserved exons as regulated targets for the splicing activator tra2β in development. PLoS Genet. *7*, e1002390.

Griswold, M.D. (1995). Interactions between germ cells and Sertoli cells in the testis. Biol. Reprod. *52*, 211–216.

Griswold, M.D. (1998). The central role of Sertoli cells in spermatogenesis. Semin. Cell Dev. Biol. *9*, 411–416.

Griswold, M.D., Morales, C., and Sylvester, S.R. (1988). Molecular biology of the Sertoli cell. Oxf. Rev. Reprod. Biol. *10*, 124–161.

Grivna, S.T., Pyhtila, B., and Lin, H. (2006). MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. Proc. Natl. Acad. Sci. U. S. A. *103*, 13415–13420.

De Groot, A., Dulermo, R., Ortet, P., Blanchard, L., Guérin, P., Fernandez, B., Vacherie, B., Dossat, C., Jolivet, E., Siguier, P., et al. (2009). Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium Deinococcus deserti. PLoS Genet. *5*, e1000434.

Guillaume, E., Dupaix, A., Moertz, E., Courtens, J.-L., Jégou, B., and Pineau, C. (2000). Proteome analysis of spermatogonia: identification of a first set of 53 proteins. Proteome *1*, 1–20.

Guillaume, E., Evrard, B., Com, E., Moertz, E., Jégou, B., and Pineau, C. (2001a). Proteome analysis of rat spermatogonia: reinvestigation of stathmin spatio-temporal expression within the testis. Mol. Reprod. Dev. *60*, 439–445.

Guillaume, E., Pineau, C., Evrard, B., Dupaix, A., Moertz, E., Sanchez, J.C., Hochstrasser, D.F., and Jégou, B. (2001b). Cellular distribution of translationally controlled tumor protein in rat and human testes. Proteomics *1*, 880–889.

Guo, J., Shi, Y.-Q., Yang, W., Li, Y.-C., Hu, Z.-Y., and Liu, Y.-X. (2007). Testosterone upregulation of tissue type plasminogen activator expression in Sertoli cells : tPA expression in Sertoli cells. Endocrine *32*, 83–89.

Guo, Q., Kumar, T.R., Woodruff, T., Hadsell, L.A., DeMayo, F.J., and Matzuk, M.M. (1998). Overexpression of mouse follistatin causes reproductive defects in transgenic mice. Mol. Endocrinol. Baltim. Md *12*, 96–106.

Guryca, V., Kieffer-Jaquinod, S., Garin, J., and Masselon, C.D. (2008). Prospects for monolithic nano-LC columns in shotgun proteomics. Anal. Bioanal. Chem. *392*, 1291–1297.

Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. *17*, 994–999.

Hamm, G.H., and Cameron, G.N. (1986). The EMBL data library. Nucleic Acids Res. *14*, 5–9.

Haywood, M., Spaliviero, J., Jimemez, M., King, N.J.C., Handelsman, D.J., and Allan, C.M. (2003). Sertoli and germ cell development in hypogonadal (hpg) mice expressing transgenic follicle-stimulating hormone alone or in combination with testosterone. Endocrinology *144*, 509–517.

Heckert, L.L., and Griswold, M.D. (2002). The expression of the follicle-stimulating hormone receptor in spermatogenesis. Recent Prog. Horm. Res. *57*, 129–148.

Hedger, M.P., and Winnall, W.R. (2012). Regulation of activin and inhibin in the adult testis and the evidence for functional roles in spermatogenesis and immunoregulation. Mol. Cell. Endocrinol. *359*, 30–42.

Henzel, W.J., Watanabe, C., and Stults, J.T. (2003). Protein identification: the origins of peptide mass fingerprinting. J. Am. Soc. Mass Spectrom. *14*, 931–942.

Hess, R.A., and Renato de Franca, L. (2008). Spermatogenesis and cycle of the seminiferous epithelium. Adv. Exp. Med. Biol. *636*, 1–15.

Heyting, C., Dettmers, R.J., Dietrich, A.J., Redeker, E.J., and Vink, A.C. (1988). Two major components of synaptonemal complexes are specific for meiotic prophase nuclei. Chromosoma *96*, 325–332.

Hogarth, C.A., and Griswold, M.D. (2010). The key role of vitamin A in spermatogenesis. J. Clin. Invest. *120*, 956–962.

Holdcraft, R.W., and Braun, R.E. (2004). Androgen receptor function is required in Sertoli cells for the terminal differentiation of haploid spermatids. Development *131*, 459–467.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M.C., Wides, R., et al. (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science *298*, 129–149.

Hoogland, C., Mostaguir, K., Sanchez, J.-C., Hochstrasser, D.F., and Appel, R.D. (2004). SWISS-2DPAGE, ten years later. Proteomics *4*, 2352–2356.

Hu, Q., Noll, R.J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. (2005). The Orbitrap: a new mass spectrometer. J. Mass Spectrom. JMS *40*, 430–443.

Huang, X.-Y., and Sha, J.-H. (2011). Proteomics of spermatogenesis: from protein lists to understanding the regulation of male fertility and infertility. Asian J. Androl. *13*, 18–23.

Huang, D.W., Sherman, B.T., Zheng, X., Yang, J., Imamichi, T., Stephens, R., and Lempicki, R.A. (2009). Extracting biological meaning from large gene lists with DAVID. Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al *Chapter 13*, Unit 13.11.

Hummelke, G.C., and Cooney, A.J. (2004). Reciprocal regulation of the mouse protamine genes by the orphan nuclear receptor germ cell nuclear factor and CREMtau. Mol. Reprod. Dev. *68*, 394–407.

Hung, T., and Chang, H.Y. (2010). Long noncoding RNA in genome regulation: prospects and mechanisms. RNA Biol. *7*, 582–585.

Hunt, D.M., Saksena, S.K., and Chang, M.C. (2009). Effects of Estradiol-17/β on Reproduction in Adult Male Rats.

Hutt, D.M., Baltz, J.M., and Ngsee, J.K. (2005). Synaptotagmin VI and VIII and syntaxin 2 are essential for the mouse sperm acrosome reaction. J. Biol. Chem. *280*, 20197–20203.

Hüttenhain, R., Soste, M., Selevsek, N., Röst, H., Sethi, A., Carapito, C., Farrah, T., Deutsch, E.W., Kusebauch, U., Moritz, R.L., et al. (2012). Reproducible quantification of cancer-associated proteins in body fluids using targeted proteomics. Sci. Transl. Med. *4*, 142ra94.

Idler, R.K., and Yan, W. (2012). Control of messenger RNA fate by RNA-binding proteins: an emphasis on mammalian spermatogenesis. J. Androl. *33*, 309–337.

Iguchi, N., Tobias, J.W., and Hecht, N.B. (2006). Expression profiling reveals meiotic male germ cell mRNAs that are translationally up- and down-regulated. Proc. Natl. Acad. Sci. U. S. A. *103*, 7712–7717.

Ivanov, P.A., and Nadezhdina, E.S. (2006). [Stress granules: RNP-containing cytoplasmic bodies springing up under stress. The structure and mechanism of organization]. Mol. Biol. (Mosk.) *40*, 937–944.

Jaffe, J.D., Berg, H.C., and Church, G.M. (2004). Proteogenomic mapping as a complementary method to perform genome annotation. Proteomics *4*, 59–77.

Jégou, B. (1991). Spermatids are regulators of Sertoli cell function. Ann. N. Y. Acad. Sci. *637*, 340–353.

Jégou, B. (1993). The Sertoli-germ cell communication network in mammals. Int. Rev. Cytol. *147*, 25–96.

Jégou, B. (1995). La cellule de Sertoli: actualisation du concept de cellule nourricière. Médecine/sciences *11*, 519.

Jégou, B., Syed, V., Sourdaine, P., Byers, S., Gérard, N., Calle, J.V. de la, Pineau, C., Garnier, D.H., and Bauché, F. (1992). The Dialogue Between Late Spermatids and Sertoli Cells in Vertebrates: A Century of Research. In Spermatogenesis — Fertilization — Contraception, E. Nieschlag, and U.-F. Habenicht, eds. (Springer Berlin Heidelberg), pp. 57–95.

Jégou, B., Pineau, C., Velez de la Calle, J.F., Touzalin, A.M., Bardin, C.W., and Cheng, C.Y. (1993). Germ cell control of testin production is inverse to that of other Sertoli cell products. Endocrinology *132*, 2557–2562.

Jerome-Majewska, L.A., Achkar, T., Luo, L., Lupu, F., and Lacy, E. (2010). The trafficking protein Tmed2/p24beta(1) is required for morphogenesis of the mouse embryo and placenta. Dev. Biol. *341*, 154–166.

Jørgensen, C., Sherman, A., Chen, G.I., Pasculescu, A., Poliakov, A., Hsiung, M., Larsen, B., Wilkinson, D.G., Linding, R., and Pawson, T. (2009). Cell-specific information processing in segregating populations of Eph receptor ephrin-expressing cells. Science *326*, 1502–1509.

Kageyama, S., Liu, H., Kaneko, N., Ooga, M., Nagata, M., and Aoki, F. (2007). Alterations in epigenetic modifications during oocyte growth in mice. Reproduction *133*, 85–94.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–668.

Kan, Z., Garrett-Engele, P.W., Johnson, J.M., and Castle, J.C. (2005). Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. Nucleic Acids Res. *33*, 5659–5666.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900–903.

Karas, M., Bachmann, D., and Hillenkamp, F. (1985). Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. Anal. Chem. *57*, 2935–2939.

Katafuchi, K., Mori, T., Toshimori, K., and Iida, H. (2000). Localization of a syntaxin isoform, syntaxin 2, to the acrosomal region of rodent spermatozoa. Mol. Reprod. Dev. *57*, 375–383.

Kauma, S.W., Huff, T.F., Hayes, N., and Nilkaeo, A. (1999). Placental Fas ligand expression is a mechanism for maternal immune tolerance to the fetus. J. Clin. Endocrinol. Metab. *84*, 2188–2194.

Kawasaki, Y., Nakagawa, A., Nagaosa, K., Shiratsuchi, A., and Nakanishi, Y. (2002). Phosphatidylserine binding of class B scavenger receptor type I, a phagocytosis receptor of testicular sertoli cells. J. Biol. Chem. *277*, 27559–27566.

Keene, J.D. (2007). RNA regulons: coordination of post-transcriptional events. Nat. Rev. Genet. *8*, 533–543.

Kerr, G.E., Young, J.C., Horvay, K., Abud, H.E., and Loveland, K.L. (2014). Regulated Wnt/beta-catenin signaling sustains adult spermatogenesis in mice. Biol. Reprod. *90*, 3.

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kleene, K.C., Distel, R.J., and Hecht, N.B. (1984). Translational regulation and deadenylation of a protamine mRNA during spermiogenesis in the mouse. Dev. Biol. *105*, 71–79.

Klose, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. Humangenetik *26*, 231–243.

Kopera, I.A., Bilinska, B., Cheng, C.Y., and Mruk, D.D. (2010). Sertoli–germ cell junctions in the testis: a review of recent data. Philos. Trans. R. Soc. B Biol. Sci. *365*, 1593–1605.

Kotaja, N., Bhattacharyya, S.N., Jaskiewicz, L., Kimmins, S., Parvinen, M., Filipowicz, W., and Sassone-Corsi, P. (2006). The chromatoid body of male germ cells: Similarity with processing bodies and presence of Dicer and microRNA pathway components. Proc. Natl. Acad. Sci. U. S. A. *103*, 2647–2652.

Krämer, A., Green, J., Pollard, J., and Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. Bioinforma. Oxf. Engl. *30*, 523–530.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T.W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., et al. (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Dev. Camb. Engl. *131*, 839–849.

Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., Hudson, T.J., Sladek, R., and Majewski, J. (2008). Genome-wide analysis of transcript isoform variation in humans. Nat. Genet. *40*, 225–231.

De La Fuente, R., and Eppig, J.J. (2001). Transcriptional Activity of the Mouse Oocyte Genome: Companion Granulosa Cells Modulate Transcription and Chromatin Remodeling. Dev. Biol. *229*, 224–236.

Lagarrigue, M., Becker, M., Lavigne, R., Deininger, S.-O., Walch, A., Aubry, F., Suckau, D., and Pineau, C. (2011). Revisiting rat spermatogenesis with MALDI imaging at 20-microm resolution. Mol. Cell. Proteomics MCP *10*.

Laiho, A., Kotaja, N., Gyenesei, A., and Sironen, A. (2013). Transcriptome profiling of the murine testis during the first wave of spermatogenesis. PloS One *8*, e61558.

Lamb, D.J., Spotts, G.S., Shubhada, S., and Baker, K.R. (1991). Partial characterization of a unique mitogenic activity secreted by rat Sertoli cells. Mol. Cell. Endocrinol. *79*, 1–12.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lane, L., Bairoch, A., Beavis, R.C., Deutsch, E.W., Gaudet, P., Lundberg, E., and Omenn, G.S. (2014). Metrics for the Human Proteome Project 2013-2014 and strategies for finding missing proteins. J. Proteome Res. *13*, 15–20.

Lanes, C.F.C., Bizuayehu, T.T., de Oliveira Fernandes, J.M., Kiron, V., and Babiak, I. (2013). Transcriptome of Atlantic cod (Gadus morhua L.) early embryos from farmed and wild broodstocks. Mar. Biotechnol. N. Y. N *15*, 677–694.

Lardenois, A., Gattiker, A., Collin, O., Chalmel, F., and Primig, M. (2010). GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. Database *2010*, baq030–baq030.

Lavigne, R., Becker, E., Liu, Y., Evrard, B., Lardenois, A., Primig, M., and Pineau, C. (2012). Direct iterative protein profiling (DIPP) - an innovative method for large-scale protein detection applied to budding yeast mitosis. Mol. Cell. Proteomics *11*, M111.012682.

Leblond, C.P., and Clermont, Y. (1952). Definition of the stages of the cycle of the seminiferous epithelium in the rat. Ann. N. Y. Acad. Sci. *55*, 548–573.

Lee, H., Yi, E.C., Wen, B., Reily, T.P., Pohl, L., Nelson, S., Aebersold, R., and Goodlett, D.R. (2004). Optimization of reversed-phase microcapillary liquid chromatography for quantitative proteomics. J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci. *803*, 101–110.

Lee, S., Kwon, M.-S., Lee, H.-J., Paik, Y.-K., Tang, H., Lee, J.K., and Park, T. (2011). Enhanced peptide quantification using spectral count clustering and cluster abundance. BMC Bioinformatics *12*, 423.

Lei, Z.M., Mishra, S., Zou, W., Xu, B., Foltz, M., Li, X., and Rao, C.V. (2001). Targeted disruption of luteinizing hormone/human chorionic gonadotropin receptor gene. Mol. Endocrinol. Baltim. Md *15*, 184–200.

Lemke, G., and Rothlin, C.V. (2008). Immunobiology of the TAM receptors. Nat. Rev. Immunol. *8*, 327–336.

Lenz, P.H., Roncalli, V., Hassett, R.P., Wu, L.-S., Cieslak, M.C., Hartline, D.K., and Christie, A.E. (2014). De novo assembly of a transcriptome for Calanus finmarchicus (Crustacea, Copepoda)--the dominant zooplankter of the North Atlantic Ocean. PloS One *9*, e88589.

Lewis, A., Mitsuya, K., Umlauf, D., Smith, P., Dean, W., Walter, J., Higgins, M., Feil, R., and Reik, W. (2004). Imprinting on distal chromosome 7 in the placenta involves repressive histone methylation independent of DNA methylation. Nat. Genet. *36*, 1291–1295.

Lewis, B.P., Green, R.E., and Brenner, S.E. (2003a). Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. Proc. Natl. Acad. Sci. U. S. A. *100*, 189–192.

Lewis, J.D., Song, Y., de Jong, M.E., Bagha, S.M., and Ausió, J. (2003b). A walk though vertebrate and invertebrate protamines. Chromosoma *111*, 473–482.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Li, H., Papadopoulos, V., Vidic, B., Dym, M., and Culty, M. (1997). Regulation of rat testis gonocyte proliferation by platelet-derived growth factor and estradiol: identification of signaling mechanisms involved. Endocrinology *138*, 1289–1298.

Li, W., Guo, X.-J., Teng, F., Hou, X.-J., Lv, Z., Zhou, S.-Y., Bi, Y., Wan, H.-F., Feng, C.-J., Yuan, Y., et al. (2013). Tex101 is essential for male fertility by affecting sperm migration into the oviduct in mice. J. Mol. Cell Biol. *5*, 345–347.

Li, Y., Xue, W.-J., Wang, X.-H., Tian, X.-H., Liu, H.-B., Feng, X.-S., Ding, X.-M., Tian, P.-X., Pan, X.-M., Ding, C.-G., et al. (2012). Decreasing loss of cryopreserved-thawed rat islets by coculture with Sertoli cells. Transplant. Proc. *44*, 1423–1428.

Liang, M., Li, W., Tian, H., Hu, T., Wang, L., Lin, Y., Li, Y., Huang, H., and Sun, F. (2014). Sequential expression of long noncoding RNA as mRNA gene expression in specific stages of mouse spermatogenesis. Sci. Rep. *4*, 5966.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40*, D857–D861.

Lie, P.P.Y., Cheng, C.Y., and Mruk, D.D. (2011). Interleukin-1alpha is a regulator of the blood-testis barrier. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *25*, 1244–1253.

Ling, N., Ying, S.Y., Ueno, N., Shimasaki, S., Esch, F., Hotta, M., and Guillemin, R. (1986). Pituitary FSH is released by a heterodimer of the beta-subunits from the two forms of inhibin. Nature *321*, 779–782.

Liu, Y.-X. (2007). Involvement of Plasminogen Activator and Plasminogen Activator Inhibitor Type 1 in Spermatogenesis, Sperm Capacitation, and Fertilization. Semin. Thromb. Hemost. *33*, 029–040.

Liu, M., Hu, Z., Qi, L., Wang, J., Zhou, T., Guo, Y., Zeng, Y., Zheng, B., Wu, Y., Zhang, P., et al. (2013). Scanning of novel cancer/testis proteins by human testis proteomic analysis. Proteomics *13*, 1200–1210.

Liu, Y.X., Du, Q., Liu, K., and Fu, G.Q. (1995). Hormonal regulation of plasminogen activator in rat and mouse seminiferous epithelium. Biol. Signals *4*, 232–240.

Looso, M., Preussner, J., Sousounis, K., Bruckskotten, M., Michel, C.S., Lignelli, E., Reinhardt, R., Hoeffner, S., Krueger, M., Tsonis, P.A., et al. (2013). A de novo assembly of the newt transcriptome combined with proteomic validation identifies new protein families expressed during tissue regeneration. Genome Biol. *14*.

Lottspeich, F., and Kellermann, J. (2011). ICPL labeling strategies for proteome research. Methods Mol. Biol. Clifton NJ *753*, 55–64.

Louvi, A., Nishimura, S., and Günel, M. (2014). Ccm3, a gene associated with cerebral cavernous malformations, is required for neuronal migration. Dev. Camb. Engl. *141*, 1404–1415.

Lu, Q., Gore, M., Zhang, Q., Camenisch, T., Boast, S., Casagranda, F., Lai, C., Skinner, M.K., Klein, R., Matsushima, G.K., et al. (1999). Tyro-3 family receptors are essential regulators of mammalian spermatogenesis. Nature *398*, 723–728.

Lui, W.-Y., and Cheng, C.Y. (2008). Transcription regulation in spermatogenesis. Adv. Exp. Med. Biol. *636*, 115–132.

Lykke-Andersen, K., Gilchrist, M.J., Grabarek, J.B., Das, P., Miska, E., and Zernicka-Goetz, M. (2008). Maternal Argonaute 2 Is Essential for Early Mouse Development at the Maternal-Zygotic Transition. Mol. Biol. Cell *19*, 4383–4392.

Macleod, G., and Varmuza, S. (2013). The application of proteomic approaches to the study of mammalian spermatogenesis and sperm function. FEBS J. *280*, 5635–5651.

Maekawa, M., Kamimura, K., and Nagano, T. (1996). Peritubular myoid cells in the testis: their structure and function. Arch. Histol. Cytol. *59*, 1–13.

Le Magueresse, B., Pineau, C., Guillou, F., and Jégou, B. (1988). Influence of germ cells upon transferrin secretion by rat Sertoli cells in vitro. J. Endocrinol. *118*, R13–R16.

Makarov (2000). Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. Anal. Chem. *72*, 1156–1162.

Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., and Horning, S. (2006). Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. Anal. Chem. *78*, 2113–2120.

Mann, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal. Chem. *66*, 4390–4399.

Margolin, G., Khil, P.P., Kim, J., Bellani, M.A., and Camerini-Otero, R.D. (2014). Integrated transcriptome analysis of mouse spermatogenesis. BMC Genomics *15*, 39.

Mather, J.P., Attie, K.M., Woodruff, T.K., Rice, G.C., and Phillips, D.M. (1990). Activin stimulates spermatogonial proliferation in germ-Sertoli cell cocultures from immature rat testis. Endocrinology *127*, 3206–3214.

Matzuk, M.M., and Lamb, D.J. (2002). Genetic dissection of mammalian fertility pathways. Nat. Cell Biol. *4 Suppl*, s41–s49.

McCarrey, J.R., Geyer, C.B., and Yoshioka, H. (2005). Epigenetic regulation of testis-specific gene expression. Ann. N. Y. Acad. Sci. *1061*, 226–242.

McKinnell, C., and Sharpe, R.M. (1997). Regulation of the secretion and synthesis of rat Sertoli cell SGP-1, SGP-2 and CP-2 by elongate spermatids. Int. J. Androl. *20*, 171–179.

McLachlan, R.I., O'Donnell, L., Meachem, S.J., Stanton, P.G., de Kretser, D.M., Pratis, K., and Robertson, D.M. (2002). Identification of specific sites of hormonal regulation in spermatogenesis in rats, monkeys, and man. Recent Prog. Horm. Res. *57*, 149–179.

Meikar, O., Vagin, V.V., Chalmel, F., Sõstar, K., Lardenois, A., Hammell, M., Jin, Y., Da Ros, M., Wasik, K.A., Toppari, J., et al. (2014). An atlas of chromatoid body components. RNA N. Y. N *20*, 483–495.

Melamud, E., and Moult, J. (2009). Stochastic noise in splicing machinery. Nucleic Acids Res. *37*, 4873–4886.

Mendis-Handagama, S.M. (1997). Luteinizing hormone on Leydig cell structure and function. Histol. Histopathol. *12*, 869–882.

Meng, J., Holdcraft, R.W., Shima, J.E., Griswold, M.D., and Braun, R.E. (2005). Androgens regulate the permeability of the blood-testis barrier. Proc. Natl. Acad. Sci. U. S. A. *102*, 16696–16700.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). Long non-coding RNAs: insights into functions. Nat. Rev. Genet. *10*, 155–159.

Meyer, L.R., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Kuhn, R.M., Wong, M., Sloan, C.A., Rosenbloom, K.R., Roe, G., Rhead, B., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. *41*, D64–D69.

Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Müller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., et al. (2012). Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes. Mol. Cell. Proteomics *11*, O111.013698.

Moens, P.B. (1978). Lateral element cross connections of the synaptonemal complex and their relationship to chiasmata in rat spermatocytes. Can. J. Genet. Cytol. J. Can. Génétique Cytol. *20*, 567–579.

Monesi, V. (1964). Ribonucleic acid synthesis during mitosis and meiosis in the mouse testis. J. Cell Biol. *22*, 521–532.

Moore, G.P.M., and Lintern-Moore, S. (1978). Transcription of the Mouse Oocyte Genome. Biol. Reprod. *18*, 865–870.

Morales, C., and Clermont, Y. (1986). Receptor-mediated endocytosis of transferrin by Sertoli cells of the rat. Biol. Reprod. *35*, 393–405.

Morales, C., Clermont, Y., and Hermo, L. (1985). Nature and function of endocytosis in Sertoli cells of the rat. Am. J. Anat. *173*, 203–217.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Muntel, J., Hecker, M., and Becher, D. (2012). An exclusion list based label-free proteome quantification approach using an LTQ Orbitrap. Rapid Commun. Mass Spectrom. RCM *26*, 701–709.

Nakagawa, T., Nabeshima, Y.-I., and Yoshida, S. (2007). Functional identification of the actual and potential stem cell compartments in mouse spermatogenesis. Dev. Cell *12*, 195–206.

Nakamura, N., Dai, Q., Williams, J., Goulding, E.H., Willis, W.D., Brown, P.R., and Eddy, E.M. (2013). Disruption of a spermatogenic cell-specific mouse enolase 4 (eno4) gene causes sperm structural defects and male infertility. Biol. Reprod. *88*, 90.

Nakamura, T., Yao, R., Ogawa, T., Suzuki, T., Ito, C., Tsunekawa, N., Inoue, K., Ajima, R., Miyasaka, T., Yoshida, Y., et al. (2004). Oligo-astheno-teratozoospermia in mice lacking Cnot7, a regulator of retinoid X receptor beta. Nat. Genet. *36*, 528–533.

Nakanishi, Y., and Shiratsuchi, A. (2004). Phagocytic removal of apoptotic spermatogenic cells by Sertoli cells: mechanisms and consequences. Biol. Pharm. Bull. *27*, 13–16.

Nanduri, B., Lawrence, M.L., Vanguri, S., Pechan, T., and Burgess, S.C. (2005). Proteomic analysis using an unfinished bacterial genome: the effects of subminimum inhibitory concentrations of antibiotics on Mannheimia haemolytica virulence factor expression. Proteomics *5*, 4852–4863.

Nesvizhskii, A.I. (2010). A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. J. Proteomics *73*, 2092–2123.

Nesvizhskii, A.I., and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics MCP *4*, 1419–1440.

Nesvizhskii, A.I., Vitek, O., and Aebersold, R. (2007). Analysis and validation of proteomic data generated by tandem mass spectrometry. Nat. Methods *4*, 787–797.

Nguyen Chi, M., Chalmel, F., Agius, E., Vanzo, N., Khabar, K.S.A., Jégou, B., and Morello, D. (2009). Temporally regulated traffic of HuR and its associated ARE-containing mRNAs from the chromatoid body to polysomes during mouse spermatogenesis. PloS One *4*, e4900.

Nicholls, P.K., Harrison, C.A., Walton, K.L., McLachlan, R.I., O'Donnell, L., and Stanton, P.G. (2011). Hormonal regulation of sertoli cell micro-RNAs at spermiation. Endocrinology *152*, 1670–1683.

Nogueira, F.C.S., Palmisano, G., Schwämmle, V., Campos, F.A.P., Larsen, M.R., Domont, G.B., and Roepstorff, P. (2012). Performance of isobaric and isotopic labeling in quantitative plant proteomics. J. Proteome Res. *11*, 3046–3052.

Nynca, J., Arnold, G.J., Fröhlich, T., Otte, K., and Ciereszko, A. (2014). Proteomic identification of rainbow trout sperm proteins. Proteomics.

Oakes, C.C., La Salle, S., Smiraglia, D.J., Robaire, B., and Trasler, J.M. (2007). Developmental acquisition of genome-wide DNA methylation occurs prior to meiosis in male germ cells. Dev. Biol. *307*, 368–379.

O'Brien, D.A., Gabel, C.A., and Eddy, E.M. (1993). Mouse Sertoli cells secrete mannose 6-phosphate containing glycoproteins that are endocytosed by spermatogenic cells. Biol. Reprod. *49*, 1055–1065.

O'Donnell, L., McLachlan, R.I., Wreford, N.G., de Kretser, D.M., and Robertson, D.M. (1996). Testosterone withdrawal promotes stage-specific detachment of round spermatids from the rat seminiferous epithelium. Biol. Reprod. *55*, 895–901.

O'Donnell, L., Robertson, K.M., Jones, M.E., and Simpson, E.R. (2001). Estrogen and spermatogenesis. Endocr. Rev. *22*, 289–318.

O'Donnell, L., Nicholls, P.K., O'Bryan, M.K., McLachlan, R.I., and Stanton, P.G. (2011). Spermiation: The process of sperm release. Spermatogenesis *1*, 14–35.

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis of proteins. J. Biol. Chem. *250*, 4007–4021.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics MCP *1*, 376–386.

Onoda, M., and Djakiew, D. (1990). Modulation of Sertoli cell secretory function by rat round spermatid protein(s). Mol. Cell. Endocrinol. *73*, 35–44.

Onoda, M., and Djakiew, D. (1993). A 24,500-Da Protein Derived from Rat Germ Cells Is Associated with Sertoli Cell Secretory Function. Biochem. Biophys. Res. Commun. *197*, 688–695.

Onoda, M., Djakiew, D., and Papadopoulos, V. (1991). Pachytene spermatocytes regulate the secretion of Sertoli cell protein(s) which stimulate Leydig cell steroidogenesis. Mol. Cell. Endocrinol. *77*, 207–216.

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363.

Orth, J.M., Gunsalus, G.L., and Lamperti, A.A. (1988). Evidence from Sertoli cell-depleted rats indicates that spermatid number in adults depends on numbers of Sertoli cells produced during perinatal development. Endocrinology *122*, 787–794.

O'Shaughnessy, P.J. (2014). Hormonal control of germ cell development and spermatogenesis. Semin. Cell Dev. Biol.

O'Shaughnessy, P.J., Abel, M., Charlton, H.M., Hu, B., Johnston, H., and Baker, P.J. (2007). Altered expression of genes involved in regulation of vitamin A metabolism, solute transportation, and cytoskeletal function in the androgen-insensitive tfm mouse testis. Endocrinology *148*, 2914–2924.

O'Shaughnessy, P.J., Verhoeven, G., De Gendt, K., Monteiro, A., and Abel, M.H. (2010). Direct action through the sertoli cells is essential for androgen stimulation of spermatogenesis. Endocrinology *151*, 2343–2348.

O'Shaughnessy, P.J., Monteiro, A., and Abel, M. (2012). Testicular development in mice lacking receptors for follicle stimulating hormone and androgen. PloS One *7*, e35136.

Oshiro, G., Wodicka, L.M., Washburn, M.P., Yates, J.R., 3rd, Lockhart, D.J., and Winzeler, E.A. (2002). Parallel identification of new genes in Saccharomyces cerevisiae. Genome Res. *12*, 1210–1220.

Overton, G.C., Aaronson, J.S., Haas, J., and Adams, J. (1994). QGB: a system for querying sequence database fields and features. J. Comput. Biol. J. Comput. Mol. Cell Biol. *1*, 3–14.

Oyama, T., Sasagawa, S., Takeda, S., Hess, R.A., Lieberman, P.M., Cheng, E.H., and Hsieh, J.J. (2013). Cleavage of TFIIA by Taspase1 activates TRF2-specified mammalian male germ cell programs. Dev. Cell *27*, 188–200.

Paik, Y.-K., Jeong, S.-K., Omenn, G.S., Uhlen, M., Hanash, S., Cho, S.Y., Lee, H.-J., Na, K., Choi, E.-Y., Yan, F., et al. (2012). The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat. Biotechnol. *30*, 221–223.

Palmer, M.R., McDowall, M.H., Stewart, L., Ouaddi, A., MacCoss, M.J., and Swanson, W.J. (2013). Mass spectrometry and next-generation sequencing reveal an abundant and rapidly evolving abalone sperm protein. Mol. Reprod. Dev. *80*, 460–465.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat. Genet. *40*, 1413–1415.

Paronetto, M.P., Zalfa, F., Botti, F., Geremia, R., Bagni, C., and Sette, C. (2006). The nuclear RNA-binding protein Sam68 translocates to the cytoplasm and associates with the polysomes in mouse spermatocytes. Mol. Biol. Cell *17*, 14–24.

Parvinen, M. (1982). Regulation of the seminiferous epithelium. Endocr. Rev. *3*, 404–417.

Parvinen, M. (2005). The chromatoid body in spermatogenesis. Int. J. Androl. *28*, 189–201.

Pawar, H., Sahasrabuddhe, N.A., Renuse, S., Keerthikumar, S., Sharma, J., Kumar, G.S.S., Venugopal, A., Sekhar, N.R., Kelkar, D.S., Nemade, H., et al. (2012). A proteogenomic approach to map the proteome of an unsequenced pathogen – Leishmania donovani. PROTEOMICS *12*, 832–844.

Pearse, R.V., Drolet, D.W., Kalla, K.A., Hooshmand, F., Bermingham, J.R., and Rosenfeld, M.G. (1997). Reduced fertility in mice deficient for the POU protein sperm-1. Proc. Natl. Acad. Sci. U. S. A. *94*, 7555–7560.

Pellegrini, M., Grimaldi, P., Rossi, P., Geremia, R., and Dolci, S. (2003). Developmental expression of BMP4/ALK3/SMAD5 signaling pathway in the mouse testis: a potential role of BMP4 in spermatogonia differentiation. J. Cell Sci. *116*, 3363–3372.

Perey, B., Clermont, Y., and Leblond, C. (1961). The wave of the seminiferous epithelium in the rat. Am J Anat *108*, 47–77.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS *20*, 3551–3567.

Pesce, M., Gross, M.K., and Schöler, H.R. (1998a). In line with our ancestors: Oct-4 and the mammalian germ. BioEssays News Rev. Mol. Cell. Dev. Biol. *20*, 722–732.

Pesce, M., Wang, X., Wolgemuth, D.J., and Schöler, H. (1998b). Differential expression of the Oct-4 transcription factor during mouse germ cell differentiation. Mech. Dev. *71*, 89–98.

Peschansky, V.J., and Wahlestedt, C. (2014). Non-coding RNAs as direct and indirect modulators of epigenetic regulation. Epigenetics Off. J. DNA Methylation Soc. *9*, 3–12.

Piechura, H., Oeljeklaus, S., and Warscheid, B. (2012). SILAC for the study of mammalian cell lines and yeast protein complexes. Methods Mol. Biol. Clifton NJ *893*, 201–221.

Pineau, C., Velez de la Calle, J.F., Pinon-Lataillade, G., and Jégou, B. (1989). Assessment of testicular function after acute and chronic irradiation: further evidence for an influence of late spermatids on Sertoli cell function in the adult rat. Endocrinology *124*, 2720–2728.

Pineau, C., Sharpe, R.M., Saunders, P.T., Gérard, N., and Jégou, B. (1990). Regulation of Sertoli cell inhibin production and of inhibin alpha-subunit mRNA levels by specific germ cell types. Mol. Cell. Endocrinol. *72*, 13–22.

Pineau, C., Le Magueresse, B., Courtens, J.L., and Jégou, B. (1991). Study in vitro of the phagocytic function of Sertoli cells in the rat. Cell Tissue Res. *264*, 589–598.

Pineau, C., Syed, V., Bardin, C.W., Jégou, B., and Cheng, C.Y. (1993). Germ cell-conditioned medium contains multiple factors that modulate the secretion of testins, clusterin, and transferrin by Sertoli cells. J. Androl. *14*, 87–98.

Pineau, C., Syed, V., Bardin, C.W., Jégou, B., and Cheng, C.Y. (1993b). Identification and partial purification of a germ cell factor that stimulates transferrin secretion by Sertoli cells. Recent Prog. Horm. Res. *48*, 539–542.

Pinon-Lataillade, G., Vélez de la Calle, J.F., Viguier-Martinez, M.C., Garnier, D.H., Folliot, R., Maas, J., and Jégou, B. (1988). Influence of germ cells upon Sertoli cells during continuous low-dose rate gamma-irradiation of adult rats. Mol. Cell. Endocrinol. *58*, 51–63.

Pinon-Lataillade, G., Viguier-Martinez, M.C., Touzalin, A.M., Maas, J., and Jégou, B. (1991). Effect of an acute exposure of rat testes to gamma rays on germ cells and on Sertoli and Leydig cell functions. Reprod. Nutr. Dev. *31*, 617–629.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., et al. (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. *29*, 742–749.

Qian, X., Mruk, D.D., Cheng, Y.-H., Tang, E.I., Han, D., Lee, W.M., Wong, E.W.P., and Cheng, C.Y. (2014). Actin binding proteins, spermatid transport and spermiation. Semin. Cell Dev. Biol. *30*, 75–85.

Rabilloud, T. (1996). Solubilization of proteins for electrophoretic analyses. ELECTROPHORESIS *17*, 813–829.

Rainey, M.A., George, M., Ying, G., Akakura, R., Burgess, D.J., Siefker, E., Bargar, T., Doglio, L., Crawford, S.E., Todd, G.L., et al. (2010). The endocytic recycling regulator EHD1 is essential for spermatogenesis and male fertility in mice. BMC Dev. Biol. *10*, 37.

Rajkovic, M., Iwen, K.A.H., Hofmann, P.J., Harneit, A., and Weitzel, J.M. (2010). Functional cooperation between CREM and GCNF directs gene expression in haploid male germ cells. Nucleic Acids Res. *38*, 2268–2278.

Ramalho-Santos, J., Moreno, R.D., Wessel, G.M., Chan, E.K., and Schatten, G. (2001). Membrane trafficking machinery components associated with the mammalian acrosome during spermiogenesis. Exp. Cell Res. *267*, 45–60.

Rana, T.M. (2007). Illuminating the silence: understanding the structure and function of small RNAs. Nat. Rev. Mol. Cell Biol. *8*, 23–36.

Rannikki, A.S., Zhang, F.P., and Huhtaniemi, I.T. (1995). Ontogeny of follicle-stimulating hormone receptor gene expression in the rat testis and ovary. Mol. Cell. Endocrinol. *107*, 199–208.

Rato, L., Socorro, S., Cavaco, J.E.B., and Oliveira, P.F. (2010). Tubular Fluid Secretion in the Seminiferous Epithelium: Ion Transporters and Aquaporins in Sertoli Cells. J. Membr. Biol. *236*, 215–224.

Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. (2000). Genome annotation assessment in Drosophila melanogaster. Genome Res. *10*, 483–501.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. Science *293*, 1089–1093.

Renuse, S., Chaerkady, R., and Pandey, A. (2011). Proteogenomics. Proteomics *11*, 620–630.

Robertson, K.M., O'Donnell, L., Jones, M.E., Meachem, S.J., Boon, W.C., Fisher, C.R., Graves, K.H., McLachlan, R.I., and Simpson, E.R. (1999). Impairment of spermatogenesis in mice lacking a functional aromatase (cyp 19) gene. Proc. Natl. Acad. Sci. U. S. A. *96*, 7986–7991.

Rolland, A.D., Evrard, B., Guitton, N., Lavigne, R., Calvel, P., Couvet, M., Jégou, B., and Pineau, C. (2007). Two-dimensional fluorescence difference gel electrophoresis analysis of spermatogenesis in the rat. J. Proteome Res. *6*, 683–697.

Rolland, A.D., Jégou, B., and Pineau, C. (2008). Testicular development and spermatogenesis: harvesting the postgenomics bounty. Adv. Exp. Med. Biol. *636*, 16–41.

Rolland, A.D., Lavigne, R., Dauly, C., Calvel, P., Kervarrec, C., Freour, T., Evrard, B., Rioux-Leclercq, N., Auger, J., and Pineau, C. (2013). Identification of genital tract markers in the human seminal plasma using an integrative genomics approach. Hum. Reprod. Oxf. Engl. *28*, 199–209.

Romero, Y., Meikar, O., Papaioannou, M.D., Conne, B., Grey, C., Weier, M., Pralong, F., De Massy, B., Kaessmann, H., Vassalli, J.-D., et al. (2011). Dicer1 depletion in male germ cells leads to infertility due to cumulative meiotic and spermiogenic defects. PloS One *6*.

De Rooij, D.G., and de Boer, P. (2003). Specific arrests of spermatogenesis in genetically modified and mutant mice. Cytogenet. Genome Res. *103*, 267–276.

Rossi, P., and Dolci, S. (2013). Paracrine mechanisms involved in the control of early stages of Mammalian spermatogenesis. Front. Endocrinol. *4*, 181.

Rotem-Yehudar, R., Galperin, E., and Horowitz, M. (2001). Association of insulin-like growth factor 1 receptor with EHD1 and SNAP29. J. Biol. Chem. *276*, 33054–33060.

Rubiano-Labrador, C., Bland, C., Miotello, G., Guérin, P., Pible, O., Baena, S., and Armengaud, J. (2014). Proteogenomic insights into salt tolerance by a halotolerant alpha-proteobacterium isolated from an Andean saline spring. J. Proteomics *97*, 36–47.

Russell, L.D., Saxena, N.K., and Turner, T.T. (1989). Cytoskeletal involvement in spermiation and sperm transport. Tissue Cell *21*, 361–379.

Sadoul, R. (2006). Do Alix and ALG-2 really control endosomes for better or for worse? Biol. Cell Auspices Eur. Cell Biol. Organ. *98*, 69–77.

Saez, J.M., Tabone, E., Perrard-Sapori, M.H., and Rivarola, M.A. (1986). Paracrine role of Sertoli cells. Med. Biol. *63*, 225–236.

Saito, K., O'Donnell, L., McLachlan, R.I., and Robertson, D.M. (2000). Spermiation failure is a major contributor to early spermatogenic suppression caused by hormone withdrawal in adult rats. Endocrinology *141*, 2779–2785.

Sanglier, S., Leize, E., Van Dorsselaer, A., and Zal, F. (2003). Comparative ESI-MS study of approximately 2.2 MDa native hemocyanins from deep-sea and shore crabs: from protein oligomeric state to biotope. J. Am. Soc. Mass Spectrom. *14*, 419–429.

Sasso-Cerri, E. (2009). Enhanced ERbeta immunoexpression and apoptosis in the germ cells of cimetidine-treated rats. Reprod. Biol. Endocrinol. RBE *7*.

Sassone-Corsi, P. (2002). Unique chromatin remodeling and transcriptional regulation in spermatogenesis. Science *296*, 2176–2178.

Sato, M., Yoshimura, S., Hirai, R., Goto, A., Kunii, M., Atik, N., Sato, T., Sato, K., Harada, R., Shimada, J., et al. (2011a). The Role of VAMP7/TI-VAMP in Cell Polarity and Lysosomal Exocytosis in vivo. Traffic *12*, 1383–1393.

Sato, T., Aiyama, Y., Ishii-Inagaki, M., Hara, K., Tsunekawa, N., Harikae, K., Uemura-Kamata, M., Shinomura, M., Zhu, X.B., Maeda, S., et al. (2011b). Cyclical and patch-like GDNF distribution along the basal surface of Sertoli cells in mouse and hamster testes. PloS One *6*, e28367.

Schier, A.F. (2007). The Maternal-Zygotic Transition: Death and Birth of RNAs. Science *316*, 406–407.

Schlecht, U., Demougin, P., Koch, R., Hermida, L., Wiederkehr, C., Descombes, P., Pineau, C., Jégou, B., and Primig, M. (2004). Expression profiling of mammalian male meiosis and gametogenesis identifies novel candidate genes for roles in the regulation of fertility. Mol. Biol. Cell *15*, 1031–1043.

Schmid, R., Grellscheid, S.N., Ehrmann, I., Dalgliesh, C., Danilenko, M., Paronetto, M.P., Pedrotti, S., Grellscheid, D., Dixon, R.J., Sette, C., et al. (2013). The splicing landscape is globally reprogrammed during male meiosis. Nucleic Acids Res. *41*, 10170–10184.

Schmidt, A., Kellermann, J., and Lottspeich, F. (2005). A novel strategy for quantitative proteomics using isotope-coded protein labels. Proteomics *5*, 4–15.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115.

Schultz, N., Hamra, F.K., and Garbers, D.L. (2003). A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. Proc. Natl. Acad. Sci. U. S. A. *100*, 12201–12206.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Seisenberger, S., Peat, J.R., and Reik, W. (2013). Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells. Curr. Opin. Cell Biol. *25*, 281–288.

Serada, S., and Naka, T. (2014). Screening for Novel Serum Biomarker for Monitoring Disease Activity in Rheumatoid Arthritis Using iTRAQ Technology-Based Quantitative Proteomic Approach. Methods Mol. Biol. Clifton NJ *1142*, 99–110.

Serang, O., Paulo, J., Steen, H., and Steen, J.A. (2013). A non-parametric cutout index for robust evaluation of identified proteins. Mol. Cell. Proteomics *12*, 807–812.

Sharon, M., Witt, S., Glasmacher, E., Baumeister, W., and Robinson, C.V. (2007). Mass Spectrometry Reveals the Missing Links in the Assembly Pathway of the Bacterial 20 S Proteasome. J. Biol. Chem. *282*, 18448–18457.

Sheets, M.D., Fox, C.A., Hunt, T., Vande Woude, G., and Wickens, M. (1994). The 3'-untranslated regions of c-mos and cyclin mRNAs stimulate translation by regulating cytoplasmic polyadenylation. Genes Dev. *8*, 926–938.

Sheynkman, G.M., Shortreed, M.R., Frey, B.L., and Smith, L.M. (2013). Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. Mol. Cell. Proteomics *12*, 2341–2353.

Shi, X., Opi, S., Lugari, A., Restouin, A., Coursindel, T., Parrot, I., Perez, J., Madore, E., Zimmermann, P., Corbeil, J., et al. (2010). Identification and biophysical assessment of the molecular recognition mechanisms between the human haemopoietic cell kinase Src homology domain 3 and ALG-2-interacting protein X. Biochem. J. *431*, 93–102.

Shima, J.E., McLean, D.J., McCarrey, J.R., and Griswold, M.D. (2004). The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. Biol. Reprod. *71*, 319–330.

Shupe, J., Cheng, J., Puri, P., Kostereva, N., and Walker, W.H. (2011). Regulation of Sertoli-germ cell adhesion and sperm release by FSH and nonclassical testosterone signaling. Mol. Endocrinol. Baltim. Md *25*, 238–252.

Sigillo, F., Pernod, G., Kolodie, L., Benahmed, M., and Le Magueresse-Battistoni, B. (1998). Residual bodies stimulate rat Sertoli cell plasminogen activator activity. Biochem. Biophys. Res. Commun. *250*, 59–62.

Silva, J.C., Gorenstein, M.V., Li, G.-Z., Vissers, J.P.C., and Geromanos, S.J. (2006). Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol. Cell. Proteomics MCP *5*, 144–156.

Simorangkir, D.R., Ramaswamy, S., Marshall, G.R., Pohl, C.R., and Plant, T.M. (2009). A selective monotropic elevation of FSH, but not that of LH, amplifies the proliferation and differentiation of spermatogonia in the adult rhesus monkey (Macaca mulatta). Hum. Reprod. Oxf. Engl. *24*, 1584–1595.

Sinkevicius, K.W., Laine, M., Lotan, T.L., Woloszyn, K., Richburg, J.H., and Greene, G.L. (2009). Estrogen-dependent and -independent estrogen receptor-alpha signaling separately regulate male fertility. Endocrinology *150*, 2898–2905.

Siu, M.K.Y., Wong, C.-H., Lee, W.M., and Cheng, C.Y. (2005). Sertoli-germ cell anchoring junction dynamics in the testis are regulated by an interplay of lipid and protein kinases. J. Biol. Chem. *280*, 25029–25047.

Siuti, N., and Kelleher, N.L. (2007). Decoding protein modifications using top-down mass spectrometry. Nat. Methods *4*, 817–821.

Skerget, S., Rosenow, M., Polpitiya, A., Petritis, K., Dorus, S., and Karr, T.L. (2013). The Rhesus macaque (Macaca mulatta) sperm proteome. Mol. Cell. Proteomics MCP *12*, 3052–3067.

Skinner, M.K., and Fritz, I.B. (1985). Structural characterization of proteoglycans produced by testicular peritubular cells and Sertoli cells. J. Biol. Chem. *260*, 11874–11883.

Son, C.G., Bilke, S., Davis, S., Greer, B.T., Wei, J.S., Whiteford, C.C., Chen, Q.-R., Cenacchi, N., and Khan, J. (2005). Database of mRNA gene expression profiles of multiple human organs. Genome Res. *15*, 443–450.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. Cell Rep. *3*, 2179–2190.

Steegmaier, M., Oorschot, V., Klumperman, J., and Scheller, R.H. (2000). Syntaxin 17 Is Abundant in Steroidogenic Cells and Implicated in Smooth Endoplasmic Reticulum Membrane Dynamics. Mol. Biol. Cell *11*, 2719–2731.

Størvold, G.L., Landskron, J., Strozynski, M., Arntzen, M.Ø., Koehler, C.J., Kalland, M.E., Taskén, K., and Thiede, B. (2013). Quantitative profiling of tyrosine phosphorylation revealed changes in the activity of the T cell receptor signaling pathway upon cisplatin-induced apoptosis. J. Proteomics *91*, 344–357.

Strappazzon, F., Torch, S., Chatellard-Causse, C., Petiot, A., Thibert, C., Blot, B., Verna, J.-M., and Sadoul, R. (2010). Alix is involved in caspase 9 activation during calcium-induced apoptosis. Biochem. Biophys. Res. Commun. *397*, 64–69.

Sun, J., Wang, W., Hundertmark, C., Zeng, A.-P., Jahn, D., and Deckwer, W.-D. (2006). A protein database constructed from low-coverage genomic sequence of Bacillus megaterium and its use for accelerated proteomic analysis. J. Biotechnol. *124*, 486–495.

Syed, V., and Hecht, N.B. (1997). Up-regulation and down-regulation of genes expressed in cocultures of rat Sertoli cells and germ cells. Mol. Reprod. Dev. *47*, 380–389.

Syed, V., Stéphan, J.P., Gérard, N., Legrand, A., Parvinen, M., Bardin, C.W., and Jégou, B. (1995). Residual bodies activate Sertoli cell interleukin-1 alpha (IL-1 alpha) release, which triggers IL-6 production by an autocrine mechanism, through the lipoxygenase pathway. Endocrinology *136*, 3070–3078.

Szalinski, C.M., Labilloy, A., Bruns, J.R., and Weisz, O.A. (2014). VAMP7 Modulates Ciliary Biogenesis in Kidney Cells. PloS One *9*, e86425.

Takashima, S., Takehashi, M., Lee, J., Chuma, S., Okano, M., Hata, K., Suetake, I., Nakatsuji, N., Miyoshi, H., Tajima, S., et al. (2009). Abnormal DNA methyltransferase expression in mouse germline stem cells results in spermatogenic defects. Biol. Reprod. *81*, 155–164.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534–538.

Tang, F., Kaneda, M., O'Carroll, D., Hajkova, P., Barton, S.C., Sun, Y.A., Lee, C., Tarakhovsky, A., Lao, K., and Surani, M.A. (2007). Maternal microRNAs are essential for mouse zygotic development. Genes Dev. *21*, 644–648.

Tanner, S., Shen, Z., Ng, J., Florea, L., Guigó, R., Briggs, S.P., and Bafna, V. (2007). Improving gene annotation using peptide mass spectrometry. Genome Res. *17*, 231–239.

Tauzin, S., Chaigne-Delalande, B., Selva, E., Khadra, N., Daburon, S., Contin-Bordes, C., Blanco, P., Le Seyec, J., Ducret, T., Counillon, L., et al. (2011). The naturally processed CD95L elicits a c-yes/calcium/PI3K-driven cell migration pathway. PLoS Biol. *9*, e1001090.

Taverner, T., Hernández, H., Sharon, M., Ruotolo, B.T., Matak-Vinković, D., Devos, D., Russell, R.B., and Robinson, C.V. (2008). Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. Acc. Chem. Res. *41*, 617–627.

The UniProt Consortium (2014). Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. *42*, D191–D198.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A.K.A., and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. Anal. Chem. *75*, 1895–1904.

Toebosch, A.M., Robertson, D.M., Klaij, I.A., de Jong, F.H., and Grootegoed, J.A. (1989). Effects of FSH and testosterone on highly purified rat Sertoli cells: inhibin alpha-subunit mRNA expression and inhibin secretion are enhanced by FSH but not by testosterone. J. Endocrinol. *122*, 757–762.

Tracy, M.R., and Hedges, S.B. (2000). Evolutionary history of the enolase gene family. Gene *259*, 129–138.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Trasler, J.M. (2009). Epigenetics in spermatogenesis. Mol. Cell. Endocrinol. *306*, 33–36.

Tsukamoto, H., Yoshitake, H., Mori, M., Yanagida, M., Takamori, K., Ogawa, H., Takizawa, T., and Araki, Y. (2006). Testicular proteins associated with the germ cell-marker, TEX101: involvement of cellubrevin in TEX101-trafficking to the cell surface during spermatogenesis. Biochem. Biophys. Res. Commun. *345*, 229–238.

Tsuruta, J.K., Eddy, E.M., and O'Brien, D.A. (2000). Insulin-like growth factor-II/cation-independent mannose 6-phosphate receptor mediates paracrine interactions during spermatogonial development. Biol. Reprod. *63*, 1006–1013.

Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. Nat. Genet. *36*, 1296–1300.

Unlü, M., Morgan, M.E., and Minden, J.S. (1997). Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis *18*, 2071–2077.

Vasudevan, S., and Steitz, J.A. (2007). AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. Cell *128*, 1105–1118.

Venables, J.P., Dalgliesh, C., Paronetto, M.P., Skitt, L., Thornton, J.K., Saunders, P.T., Sette, C., Jones, K.T., and Elliott, D.J. (2004). SIAH1 targets the alternative splicing factor T-STAR for degradation by the proteasome. Hum. Mol. Genet. *13*, 1525–1534.

Venter, E., Smith, R.D., and Payne, S.H. (2011). Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PloS One *6*, e27587.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Verhoeven, G., Willems, A., Denolet, E., Swinnen, J.V., and Gendt, K.D. (2010). Androgens and spermatogenesis: lessons from transgenic mouse models. Philos. Trans. R. Soc. B Biol. Sci. *365*, 1537–1556.

Vourekas, A., Zheng, Q., Alexiou, P., Maragkakis, M., Kirino, Y., Gregory, B.D., and Mourelatos, Z. (2012). Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nat. Struct. Mol. Biol. *19*, 773–781.

Walker, W.H. (2010). Non-classical actions of testosterone and spermatogenesis. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *365*, 1557–1569.

Wang, D., Lou, J., Ouyang, C., Chen, W., Liu, Y., Liu, X., Cao, X., Wang, J., and Lu, L. (2010). Ras-related protein Rab10 facilitates TLR4 signaling by promoting replenishment of TLR4 onto the plasma membrane. Proc. Natl. Acad. Sci. U. S. A. *107*, 13806–13811.

Wang, J., Xia, Y., Wang, G., Zhou, T., Guo, Y., Zhang, C., An, X., Sun, Y., Guo, X., Zhou, Z., et al. (2014a). In-depth proteomic analysis of whole testis tissue from the adult rhesus macaque. Proteomics.

Wang, R.-S., Yeh, S., Chen, L.-M., Lin, H.-Y., Zhang, C., Ni, J., Wu, C.-C., di Sant'Agnese, P.A., deMesy-Bentley, K.L., Tzeng, C.-R., et al. (2006). Androgen receptor in sertoli cell is essential for germ cell nursery and junctional complex formation in mouse testes. Endocrinology *147*, 5624–5633.

Wang, S.-F., Tsao, C.-H., Lin, Y.-T., Hsu, D.K., Chiang, M.-L., Lo, C.-H., Chien, F.-C., Chen, P., Chen, Y.-M.A., Chen, H.-Y., et al. (2014b). Galectin-3 promotes HIV-1 budding via association with Alix and Gag p6. Glycobiology.

Wang, X., Slebos, R.J.C., Wang, D., Halvey, P.J., Tabb, D.L., Liebler, D.C., and Zhang, B. (2012). Protein identification using customized protein sequence databases derived from RNA-Seq data. J. Proteome Res. *11*, 1009–1017.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Warnken, U., Schleich, K., Schnölzer, M., and Lavrik, I. (2013). Quantification of High-Molecular Weight Protein Platforms by AQUA Mass Spectrometry as Exemplified for the CD95 Death-Inducing Signaling Complex (DISC). Cells *2*, 476–495.

Warren, A.S., Archuleta, J., Feng, W.-C., and Setubal, J.C. (2010). Missing genes in the annotation of prokaryotic genomes. BMC Bioinformatics *11*, 131.

Webster, K.E., O'Bryan, M.K., Fletcher, S., Crewther, P.E., Aapola, U., Craig, J., Harrison, D.K., Aung, H., Phutikanit, N., Lyle, R., et al. (2005). Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. Proc. Natl. Acad. Sci. U. S. A. *102*, 4068–4073.

Welsh, M., Saunders, P.T.K., Atanassova, N., Sharpe, R.M., and Smith, L.B. (2009). Androgen action via testicular peritubular myoid cells is essential for male fertility. FASEB J. *23*, 4218–4230.

Whitehouse, C.M., Dreyer, R.N., Yamashita, M., and Fenn, J.B. (1985). Electrospray interface for liquid chromatographs and mass spectrometers. Anal. Chem. *57*, 675–679.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587.

Winter, A.G., Wildenhain, J., and Tyers, M. (2011). BioGRID REST Service, BiogridPlugin2 and BioGRID WebGraph: new tools for access to interaction data at BioGRID. Bioinformatics *27*, 1043–1044.

Woo, S., Cha, S.W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., and Bafna, V. (2014). Proteogenomic database construction driven from large scale RNA-seq data. J. Proteome Res. *13*, 21–28.

Wrobel, G., and Primig, M. (2005). Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. Reproduction *129*, 1–7.

Xie, J., Lee, J.-A., Kress, T.L., Mowry, K.L., and Black, D.L. (2003). Protein kinase A phosphorylation modulates transport of the polypyrimidine tract-binding protein. Proc. Natl. Acad. Sci. U. S. A. *100*, 8776–8781.

Xing, X.-B., Li, Q.-R., Sun, H., Fu, X., Zhan, F., Huang, X., Li, J., Chen, C.-L., Shyr, Y., Zeng, R., et al. (2011). The discovery of novel protein-coding features in mouse genome based on mass spectrometry data. Genomics *98*, 343–351.

Xiong, W., Chen, Y., Wang, H., Wang, H., Wu, H., Lu, Q., and Han, D. (2008). Gas6 and the Tyro 3 receptor tyrosine kinase subfamily regulate the phagocytic function of Sertoli cells. Reproduction *135*, 77–87.

Xiong, W., Wang, H., Wu, H., Chen, Y., and Han, D. (2009). Apoptotic spermatogenic cells can be energy sources for Sertoli cells. Reprod. Camb. Engl. *137*, 469–479.

Xu, Q., Lin, H.-Y., Yeh, S.-D., Yu, I.-C., Wang, R.-S., Chen, Y.-T., Zhang, C., Altuwaijri, S., Chen, L.-M., Chuang, K.-H., et al. (2007). Infertility with defective spermatogenesis and steroidogenesis in male mice lacking androgen receptor in Leydig cells. Endocrine *32*, 96–106.

Yaman, R., and Grandjean, V. (2006). Timing of entry of meiosis depends on a mark generated by DNA methyltransferase 3a in testis. Mol. Reprod. Dev. *73*, 390–397.

Yanagiya, A., Delbes, G., Svitkin, Y.V., Robaire, B., and Sonenberg, N. (2010). The poly(A)-binding protein partner Paip2a controls translation during late spermiogenesis in mice. J. Clin. Invest. *120*, 3389–3400.

Yefimova, M.G., Sow, A., Fontaine, I., Guilleminot, V., Martinat, N., Crepieux, P., Canepa, S., Maurel, M.-C., Fouchécourt, S., Reiter, E., et al. (2008). Dimeric transferrin inhibits phagocytosis of residual bodies by testicular rat Sertoli cells. Biol. Reprod. *78*, 697–704.

Yergey, J., Heller, D., Hansen, G., Cotter, R.J., and Fenselau, C. (1983). Isotopic distributions in mass spectra of large molecules. Anal. Chem. *55*, 353–356.

Yu, Y.E., Zhang, Y., Unni, E., Shirley, C.R., Deng, J.M., Russell, L.D., Weil, M.M., Behringer, R.R., and Meistrich, M.L. (2000). Abnormal spermatogenesis and reduced fertility in transition nuclear protein 1-deficient mice. Proc. Natl. Acad. Sci. U. S. A. *97*, 4683–4688.

Zavalin, A., Todd, E.M., Rawhouser, P.D., Yang, J., Norris, J.L., and Caprioli, R.M. (2012). Direct imaging of single cells and tissue at sub-cellular spatial resolution using transmission geometry MALDI MS. J. Mass Spectrom. JMS *47*, 1473–1481.

Zhang, M.Q. (2002). Computational prediction of eukaryotic protein-coding genes. Nat. Rev. Genet. *3*, 698–709.

Zhang, H., Yin, Y., Wang, G., Liu, Z., Liu, L., and Sun, F. (2014). Interleukin-6 disrupts blood-testis barrier through inhibiting protein degradation or activating phosphorylated ERK in Sertoli cells. Sci. Rep. *4*, 4260.

Zhang, Y., Tang, W., Zhang, H., Niu, X., Xu, Y., Zhang, J., Gao, K., Pan, W., Boggon, T.J., Toomre, D., et al. (2013). A network of interactions enables CCM3 and STK24 to coordinate UNC13D-driven vesicle exocytosis in neutrophils. Dev. Cell *27*, 215–226.

Zhong, J., Cui, Y., Guo, J., Chen, Z., Yang, L., He, Q.-Y., Zhang, G., and Wang, T. (2014). Resolving chromosome-centric human proteome with translating mRNA analysis: a strategic demonstration. J. Proteome Res. *13*, 50–59.

Zhou, J.-Y., Schepmoes, A.A., Zhang, X., Moore, R.J., Monroe, M.E., Lee, J.H., Camp, D.G., Smith, R.D., and Qian, W.-J. (2010). Improved LC-MS/MS spectral counting statistics by recovering low-scoring spectra matched to confidently identified peptide sequences. J. Proteome Res. *9*, 5698–5704.

Zhou, J.-Y., Dann, G.P., Shi, T., Wang, L., Gao, X., Su, D., Nicora, C.D., Shukla, A.K., Moore, R.J., Liu, T., et al. (2012). Simple sodium dodecyl sulfate-assisted sample preparation method for LC-MS-based proteomics applications. Anal. Chem. *84*, 2862–2867.

ANNEXES

I. Communications scientifiques

A. Articles et revues à comité de lecture

Forty-four novel protein-coding loci discovered using a PIT approach in rat male germ cells. Sophie Chocu, Bertrand Evrard, Régis Lavigne, Antoine D Rolland, Florence Aubry, Bernard Jégou, Frédéric Chalmel*, and Charles Pineau*. (submitted) MS ID#: BIOLREPROD/2014/122416.

* Equal contribution.

(Article accepté sous réserve de révisions mineures dans la revue Biology Of Reproduction)

C2orf62 and TTC17 Are Involved in Actin Organization and Ciliogenesis in Zebrafish and Human. Bontems, F., Fish, R.J., Borlat, I., Lembo, F., Chocu, S., Chalmel, F., Borg, J.-P., Pineau, C., Neerman-Arbez, M., Bairoch, A., et al. (2014). PLoS ONE *9*, e86476.

Spermatogenesis in mammals: proteomic insights. Chocu, S., Calvel, P., Rolland, A.D., and Pineau, C. (2012). Syst Biol Reprod Med *58*, 179–190.

La protéomique, un outil puissant pour comprendre la spermatogenèse normale et pathologique. Chocu, S., Calvel, P., Rolland, A.D., and Pineau, C. (2012). MT médecine de la reproduction *14*, 272–286.

Human Sperm Proteome reveals DCDC2C as a new microtubule associated protein of the sperm flagellum

Fanny Jumeau, Francisco-Jose Fernandez-Gomez, Céline Carpentier, Sophie Chocu, Hélène Obriot, Sabiha Eddarkaoui, Meryem Tardivel, Johann Hachani, Sophie Duban-Deweer, Frédéric Halgand, Frédéric Chalmel, Claire-Marie Dhaenens, Marie-Laure Caillet-Boudin, Jean-Marc Rigot, Luc Buée, Charles Pineau, Nicolas Sergeant* & Valérie Mitchell *

* Equal contribution.

(article en preparation).

B. Communications orales

- **Journée des Jeunes chercheurs de l'IRSET, 4 Fév 2014, Rennes.**

  Titre de la présentation: Identifying novel molecular actors of spermatogenesis in the rat using Proteomics Informed by Transcriptomics.

  Travail présenté par: Sophie Chocu.

- **18 th European Testis Workshop, 14-17 mai 2014, Dannemark.**

  Titre de la présentation: 'Omics' and systems biology approaches in studies of testis function and dysfunction - Novel protein-coding loci uncovered by a PIT approach in rat male germ cells.

  Travail présenté par: Frédéric Chalmel.

Abstract (18 th European Testis Workshop)

Introduction. The testis is often considered as one of the organs, if not the organ, that expresses the higher number of tissue-specific genes. In this regard it is a fertile ground for the discovery of yet unidentified transcipt isoforms, splicing events or even genes. Recently, several groups including us have identified hundreds of Novel Unannotated Transcripts (NUTs) expressed in testicular cells thanks to RNA-seq analyses1-4. While a majority of NUTs share features of non-coding RNAs, it is also very likely that some of them actually correspond to novel protein-coding genes.

Objective: to identify among these novel transcriptional events those that do code for proteins. To address this issue we used a Proteomics Informed by Transcriptomics (PIT) approach5 that combines RNA-seq with Shotgun proteomics analysis of isolated rat spermatocytes and spermatids.

Results. 99,438 transcripts (54.142 loci) were reconstructed including 32,024 long unannotated transcripts (29,668 loci) (Figure 1). 4.458 of them were significantly detected in spermatocytes (SPC) and/or spermatids (SPT). In silico translation of these unannotated transcripts yielded thousands of potential open reading frames among which 126 (86 loci) were detected by Shotgun LC-MS/MS in SPC or SPT. Of these, 90 NUTs (65 loci) were also preferentially transcribed in meiotic and/or post-meiotic germ cells as compared to Sertoli cells and spermatogonia.

After manual inspection of these potential novel protein-coding loci we selected one candidate for further experimental validations. This locus is present on the human chromosome 1, is conserved in all mammalian species and potentially codes for a 90 amino acids protein. Finally, we produced an antibody raised against the corresponding recombinant protein and confirmed the candidate protein expression on rat testicular sections by immunohistochemistry. The protein displayed a cytoplasmic localization in elongated spermatids from stages XI to XIX.

Conclusion. The PIT strategy allowed us to identify 126 novel proteins expressed during rat spermatogenesis among which 90 displayed a preferential expression pattern in the germline. As a proof of concept, we already validated one candidate and antibodies raised against a couple of additional promising candidates are currently under production .

We demonstrate that the PIT technique is a powerful tool to discover novel protein-coding genes, even in the annotated genomes of model organisms. This study paves the way for further functional investigations of the selected candidates which may play important roles in mammalian spermatogenesis.

C. Présentations affichées

- **Journée des Jeunes chercheurs de l'IRSET, 27 novembre 2012, Rennes;**
  Titre du poster: Analyse à grande échelle de l'expression des protéines germinales au cours de la spermatogenèse chez les mammifères.

- **7ème école d'été européenne de Protéomique: "Advanced Proteomics", 4 au 10 août 2013, Brixen (Sud Tyrol, Italie);**
  Titre du poster: Identifying novel molecular actors of spermatogenesis in the rat using Proteomics Informed by Transcriptomics (PIT).

- **Congrès Européen de Protéomique EuPA 2013, du 14 au 17 octobre 2013, Saint Malo;**
  Résumés des posters présentés au congrès Eupa 2013, Saint Malo :

Identifying novel molecular actors of spermatogenesis in the rat using Proteomics Informed by Transcriptomics

Sophie Chocu [1,2]*, Frédéric Chalmel [2]*, Régis Lavigne [1,2] , Bertrand Evrard[2], Florence Aubry[2], Bernard Jégou[2] and Charles Pineau [1,2]

[1]*Proteomics Core facility Biogenouest and* [2]*IRSET – Inserm U1085, Campus de Beaulieu, 35042 Rennes, France*
*: These authors made an equal contribution

Spermatogenesis is a highly sophisticated process involved in transmission of genetic heritage. Until they become mature spermatozoa, germ cells go through several successive steps including proliferation, meiosis and differentiation. The complexity of the communication network that takes place during this process is unique and depends on the coordinated and regulated expression of specific molecular actors including transcripts and proteins. The aim of this study was to identify those transcripts and proteins for which regulated expression during rat spermatogenesis predicts step-specific functions. A strategy combining RNA-seq and Shotgun proteomics analyses was used with the aim to identify novel uncharacterized protein coding genes, transcribed and translated in rat germ cells.

High-throughput RNA sequencing was performed on isolated rat testicular cell types (Chalmel *et al.* in preparation). In this analysis, 77'490 non-redundant transcripts were reconstructed including 27'917 and 10'062 intronic and intergenic novel testicular unannotated transcripts (TUTS). These 77'490 transcripts were translated into the 6 reading frames to constitute a set of putative protein sequences. A combined-set of non-redundant protein sequences was created by merging annotated or predicted protein sequences from public databases (UniProt, Ensembl) to the putative protein sequences resulting from the translation of RNA-seq transcripts detected in testicular cell samples. Total protein extracts from pachytene spermatocytes, early spermatids and residual bodies were trypsin-digested and further analysed by nanoLC-MS/MS on a LTQ-Orbitrap XL mass spectrometer. Spectral

data were queried against this custom database for peptide characterization and protein identification.

As much as 1871 non-redundant proteins were detected in pachytene spermatocytes, round spermatids and residual bodies. Among these, 10% correspond to translated regions of TUTs. In order to isolate the most robust protein-coding TUT candidates, we refined our process by further selecting: 1) highly detectable transcripts differentially expressed in rat germ cells, and 2) TUTs with a total transcript length ≥ 200 bp. Among the 164 TUTs for which at least one peptide was unambiguously identified by LC-MS/MS, 32 candidates pass these stringent criteria including 3 top-priority putative proteins conserved from human to rodents whose expression during spermatogenesis was further studied using molecular biology and biochemistry tools (RT-PCR, ISH, gene cloning and recombinant protein expression for antibody production, western blotting and IHC).

Here, we demonstrate the power of combining RNA-seq and shotgun proteomics analyses for identifying novel unannotated protein-coding genes transcribed and translated in rat germ cells during mammalian spermatogenesis. This genome wide proteomics strategy could be applied to other biological processes for discovering novel protein-coding genes and could significantly contribute to the annotation of genomes. No doubt it could also open interesting perspectives if wisely used in the context of both human c-HPP and B/D HPP initiatives.

Characterization of the Sertoli cell secretome using Shotgun proteomics and Integrative Genomics

Loren Méar1, Blandine Guével1, Sophie Chocu1, Régis Lavigne1, Frédéric Chalmel2, Charles Pineau1
1Proteomics Core Facility Biogenouest, IRSET- Inserm U1085, Campus de Beaulieu, F-35042 Rennes, France
2IRSET - Inserm U1085, Campus de Beaulieu, F-35042 Rennes, France

In the mammalian testis, Sertoli cells constitute the seminiferous epithelium and interact directly with developing germ cells throughout spermatogenesis. Sertoli cells serve as a support for germ cells and keep the seminiferous epithelium integrity by shaping extracellular matrix components and specialized cell junctions. Sertoli cells are also involved in the movement of germ cells and in the release of spermatozoa from the seminiferous epithelium. An important function of Sertoli cells is the secretion of nutrients, testis fluid, and of numerous proteins, which have not yet been totally identified. Interestingly, this secretory function is crucial for the success of spermatogenesis.

Total proteins extracts were prepared from Sertoli cells monolayers in primary cultures and further analyzed by LC MS/MS, prior to protein identification thanks to a *Shotgun* strategy. Using an integrative genomics approach, a subset of 143 potentially secreted proteins was selected and scored out of the newly established Sertoli cell proteome repertoire. The AMEN suite of tools (Chalmel & Primig 2008) was used for data mining, together with information from the Secreted Protein Database (SPD).

Many protein secreted by Sertoli cells are suspected to play a crucial role on germ cell function. Potential partners of Sertoli cell-secreted proteins - *i.e.*, germ cell plasma membrane and cell surface proteins - were selected and ranked using interactomics AMEN modules and protein network data available *via* certified public repositories. A couple of potentially interacting protein candidates were selected for further studies: the Neural Wiskott-Aldrich syndrome protein (N-WASP) and Heat shock protein HSP 90-alpha. Expression of both HSP90 with N-WASP proteins was yet described within the testis. Nevertheless, their interaction was only documented in the brain.

This *in silico* prediction strongly suggests that N-WASP/HSP90 could interact in the seminiferous tubules and thus play a role in the control of germ cell development by Sertoli cells. Co-expression and physical interaction of these two proteins partners were further validated using the Duolink™ method and by immunohistochemistry on rat testis sections.

Our results provide new insights into the Sertoli cell-germ cell crosstalk by predicting novel interacting protein partners. We also demonstrate that such an integrative genomics strategy is relevant to study cell secretomes and can easily be extended to any type of cellular model.

References :

Chalmel F. The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. BMC Bioinformatics 2008, 9:86.

Human Sperm Proteome reveals DCDC2C as a new microtubule associated protein of sperm flagellum

Fanny Jumeau [1,2], Francisco-Jose Fernandez-Gomez [2], Céline Carpentier [2], Hélène Obriot [2], Sabiha Eddarkaoui [2], Sophie Duban-Deweer [3], Johann Hachani [3], Frédéric Halgand [4], Frédéric Chalmel [4], Sophie Chocu [4], Claire-Marie Dhaenens [2], Marie-Laure Caillet-Boudin [2], Jean-Marc Rigot [1], Marie-Claire Peers [1], Charles Pineau [4], Luc Buée [2], Nicolas Sergeant [2] & Valérie Mitchell [1]

1 Institut de Biologie de la Reproduction – CHRU Lille, EA 4308 Gametogenesis and Quality of the Gamete, F-59045 Lille
2 Inserm UMR 837-1, Alzheimer & Tauopathies, Jean-Pierre Aubert Research Center, Univ. Lille 2, F-59045 Lille
3 Plate-forme Protéomique de l'Artois CAPA, Univ. de l'Artois, F-62300 Lens
4 Plate-forme Protéomique Biogenouest, IRSET Inserm U1085, F-35000 Rennes

During the sperm maturation in epididymis and in female genital tractus, spermatozoa interacts with its environment. These are translated by proteome modification and proteomics represents a relevant approach to understand sperm physiology. Flagellum permits to spermatozoa to progress through female genital tractus to reach the oocyte. Cytoskeleton is composed by a microtubular network and is implicated in sperm motility. In this study, we focused on proteins related to the microtubular network regulation and specifically DCDC2C (DoubleCortin Domain Containing 2C), a potential microtubules associated protein. Spermatozoa were isolated from semen of 20 normozoospermic individuals (WHO 2010 criteria) in our reproductive biology department. Sperm proteins were extracted, separated by 1D/2D SDS-PAGE and identified by mass spectrometry. Dataset obtained were analyzed with AMEN software. Interest protein was selected among overexpressed protein family from sperm flagellum and yet not described in testis or spermatozoa. DCDC2C expression was assessed by RT-PCR, immunoblotting and immunochemistry. In testis, DCDC2C was detected by immunoblotting, RT-PCR and localized in late state of spermatogenesis (long spermatid and spermatozoa). In spermatozoa, DCDC2C was strongly detected in immunoblotting and localized in terminal piece of sperm flagellum.

Altogether, our results suggest that proteome analysis of human spermatozoa may be a powerfull tool to further assess and understand sperm physiology. A functional flagellum is necessary for sperm motility considered as a sperm quality marker in assisted reproductive medicine. Among flagellar protein, DCDC2C expression was characterized for the first time in human testis and spermatozoa. DCDC2C belongs to doublecortin domain containing protein family implied in microtubules stabilization. DCDC2C could be implied in sperm motility failure and further investigations are needed to define the role of DCDC2C.

II. Activités d'encadrement

- Encadrement de Fanny Jumeau, (jeune Docteur, elle a soutenu sa thèse à Lille en septembre dernier); pour l'analyse de données de protéomique Shotgun, choix de nouvelles isoformes de la double cortine exprimées dans le spermatozoïde humain.

- Encadrement de Loren Méar, stagiaire de Master 1 SCMV, Université de Rennes 1 ; pour l'isolement et la mise en culture de cellules de Sertoli de rats de 20 jours, et pour l'analyse de données de protéomique, choix de candidats interactants à l'aide du logiciel AMEN.

III. Activité de vulgarisation scientifique

Participation aux **Doctoriales Bretagne 2012, 14ème édition** : Rencontres jeunes chercheurs et acteurs socio-economiques, du 3 au 7 décembre 2012, Lorient;

Titre du poster : « Décryter la spermatogenèse chez les mammifères ». Etude protéomique des cellules testiculaires.

C2orf62 and TTC17 Are Involved in Actin Organization and Ciliogenesis in Zebrafish and Human

Franck Bontems[1]*, Richard J. Fish[2], Irene Borlat[1], Frédérique Lembo[3,4,5,6], Sophie Chocu[7], Frédéric Chalmel[7], Jean-Paul Borg[3,4,5,6], Charles Pineau[7], Marguerite Neerman-Arbez[2], Amos Bairoch[1,8], Lydie Lane[1,8]*

1 Department of Human Protein Sciences, Faculty of Medicine, University of Geneva, Geneva, Switzerland, **2** Department of Genetic Medicine and Development, Faculty of Medicine, University of Geneva, Geneva, Switzerland, **3** CRCM - Inserm U1068, Marseille, France, **4** Institut Paoli-Calmettes, Marseille, France, **5** CNRS UMR7258, Marseille, France, **6** Aix-Marseille University, Marseille, France, **7** IRSET - Inserm U1085, Rennes, France, **8** SIB-Swiss Institute of Bioinformatics, Geneva, Switzerland

Abstract

Vertebrate genomes contain around 20,000 protein-encoding genes, of which a large fraction is still not associated with specific functions. A major task in future genomics will thus be to assign physiological roles to all open reading frames revealed by genome sequencing. Here we show that C2orf62, a highly conserved protein with little homology to characterized proteins, is strongly expressed in testis in zebrafish and mammals, and in various types of ciliated cells during zebrafish development. By yeast two hybrid and GST pull-down, C2orf62 was shown to interact with TTC17, another uncharacterized protein. Depletion of either C2orf62 or TTC17 in human ciliated cells interferes with actin polymerization and reduces the number of primary cilia without changing their length. Zebrafish embryos injected with morpholinos against C2orf62 or TTC17, or with mRNA coding for the C2orf62 C-terminal part containing a RII dimerization/docking (R2D2) – like domain show morphological defects consistent with imperfect ciliogenesis. We provide here the first evidence for a C2orf62-TTC17 axis that would regulate actin polymerization and ciliogenesis.

Citation: Bontems F, Fish RJ, Borlat I, Lembo F, Chocu S, et al. (2014) C2orf62 and TTC17 Are Involved in Actin Organization and Ciliogenesis in Zebrafish and Human. PLoS ONE 9(1): e86476. doi:10.1371/journal.pone.0086476

Editor: Yulia Komarova, University of Illinois at Chicago, United States of America

Received June 21, 2013; **Accepted** December 9, 2013; **Published** January 27, 2014

Copyright: © 2014 Bontems 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: JPB's lab is supported by La Ligue Contre le Cancer (Label Ligue 2010), EUCAAD (FP7 program), and Institut Paoli-Calmettes. AB's lab is supported by the Faculty of Medicine of the University of Geneva. 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: Lydie.lane@isb-sib.ch (LL); Franck.bontems@unige.ch (FB)

Introduction

Cilia are centriole-derived projections from the cell surface present in all vertebrates, in invertebrates such as *Drosophila* spp. or *Caenorhabditis elegans*, in protists like *Tetrahymena thermophila*, and in the green alga *Chlamydomonas reinhardtii* [1]. They are made of a microtubule cytoskeleton, the axoneme, surrounded by a membrane that contains numerous receptors and ion channels [2]. Classically, motile cilia are distinguished from primary/immotile cilia based on the structure of their axoneme: motile cilia have nine outer microtubule doublets with a central pair of microtubules, whereas primary cilia lack the central doublet. Motile cilia from multi-ciliated epithelial cells mediate fluid flows inside the organism through mechanically coordinated beating. They are responsible for functions such as mucus clearance in the trachea, egg removal in the fallopian tube or cerebrospinal fluid flow in the brain. Motile monocilia that direct the fluid flow in the embryonic node (Kupffer's vesicle in Zebrafish (*Danio rerio*)) are responsible for the left/right asymmetrical organization of the vertebrate body [3][4][5][6][7]. Flagella of protozoans or sperm cells are long motile cilia that allow the motility of entire cells.

A vast range of vertebrate cells are able to assemble a single, immotile primary cilium when they exit from the cell cycle. These primary cilia perform important sensing functions during development and in adult homeostasis. In fish, they are found in the neuromasts of the lateral line to sense mechanical changes in water, in the kidney to sense mechanic flow in the renal ducts, and in the brain [8]. Specialized forms of primary cilia are found in sensory cells such as olfactory sensory neurons or retinal cells [9], where they concentrate and organize sensory signaling molecules.

Regardless of its structure, the axoneme is anchored at the cell surface by the basal body, made of nine triplets of microtubules surrounding the cartwheel. The basal body of the primary cilium derives from the mother centriole during specific phases of the cell cycle. The basal bodies of multi-ciliated epithelial cells are produced de novo by centriole multiplication [10]. In both cases, the transfer of centrioles to the apical plasma membrane depends on the actin network [11].

Recent proteomic, genomic and bioinformatic analyses have allowed cataloguing of over 1,000 centrosome-, basal body- or cilia/flagella-associated proteins, collectively referred to as the "ciliome", that can be explored in specialized databases such as Centrosomedb [12], the Ciliome Database [13], Ciliaproteome [14], and Cildb [15]. Defects in some of these genes cause disorders called ciliopathies [16], which encompass many symptoms including sensory defects, developmental delay, obesity, diabetes, kidney anomalies, skeletal dysplasia, situs invertus, genital and fertility problems. Recently, dozens of additional genes involved in ciliogenesis have been identified using RNA interference on human cells [17][18]. However, some causative

genes in these diseases are still missing. We postulated that candidate genes could be found among the thousands of human proteins that are only predicted from transcriptomic data and still await experimental validation and characterization [19].

Among a set of 5,200 poorly characterized human proteins - according to neXtProt annotation[20], 1,049 were found by BLAST analysis to have a phylogenetic profile compatible with an involvement in ciliogenesis, *i.e.* conserved in vertebrates but not in the following organisms devoid of cilia: *Escherichia coli*, *Bacillus subtilis*, *Methanocaldococcus* sp., *Saccharomyces cerevisiae* and *Dictyostelium discoideum* [1].

The zebrafish model has been extensively used to study ciliogenesis [8][7][21]. It allows different approaches compared to mammalian models, including a fast and readily observable development, relatively easy establishment of transgenic reporters and the possibility to modulate protein expression by injection of morpholinos (MOs) at early developmental stages [22]. Because MO strategies are generally more promising for genes that are well conserved and have no functional equivalog, proteins having at least one paralog in human or zebrafish or whose sequences were divergent between zebrafish and human were eliminated from the preliminary set. That led to a final set of 283 poorly characterized proteins with no paralog and a phylogenetic profile compatible with a role in ciliogenesis.

This report presents the first characterization of one of these proteins, C2orf62, in zebrafish embryo and human cell lines. C2orf62 is present in ciliated cells throughout zebrafish embryonic development. In human, it is expressed in ciliated tissues, notably in testis at the moment when sperm cell flagella are formed. C2orf62 was found to interact with TTC17, another uncharacterized protein. We show that C2orf62 and TTC17 are involved in primary ciliogenesis in human cells and modulate actin polymerization. Moreover, down-regulation of C2orf62 or TTC17 by MO in zebrafish embryos or overexpression of C2orf62 C-terminus induces morphological features classically associated with ciliogenesis defects.

Results

C2orf62 is a candidate ciliogenesis gene

We reasoned that proteins involved in ciliogenesis would start to be expressed in zebrafish at 10–12 hours post fertilization (hpf), during the formation of the ciliated Kupffer's vesicle [4][6][7]. A RT-PCR screen was performed on 120 genes from the list of candidates established by bioinformatics (see introduction), and 18 matched this expression criterion (R.J.F., unpublished data). We chose to start our validation work with C2orf62 because Switzerland is committed to investigate human chromosome 2 proteins within the HUPO Chromosome-Centric Human Proteome Project, whose aim is to annotate all human proteins [23].

C2orf62 is a 387 amino-acid (aa) protein for which no information concerning function or interacting partners is available in the literature. It is not mentioned in any of the ciliome databases [15][12]. To our knowledge, it has never been detected by mass spectrometry; its existence has only been validated at transcript level in the brain (BC052750). Using an antibody against aa 217–305 (HPA044818), the Human Protein Atlas [24] team reported an enrichment in heart myocytes and bone marrow cells, but with uncertain reliability. The COSMIC database [25] reports three somatic mutations (R133W, I168V and S162Y) associated with rectal adenocarcinoma samples. Whereas C2orf62 has no annotated functional domain, Pfam [26] analysis shows that the aa 327–352 region is just below the threshold of detection by Pfam of the docking and dimerization domain of protein kinase A II-alpha (R2D2, PF02197) (Fig. 1A). This domain is found at the N-terminus of regulatory subunits of protein kinase A and at the N-terminus of four AKAP-binding proteins found in ciliated cells and highly expressed in testes, where it provides the dimerization interface and the binding site for A-kinase-anchoring proteins (AKAPs) [27][28][29] (Fig. 1B).

C2orf62 orthologs are confined to Metazoa: no ortholog was found by TBLASTN in plants, Fungi, Amoebozoa, Alveolata, or Stramenopiles. C2orf62 is well-conserved in Chordates (all vertebrates and Ciona), Echinoderms (Sea urchin), Hemichordata (Acorn worm), Insecta (Drosophila, Body louse) and Plathelminthes (Schistosoma) (Fig. 1A). The sequences of zebrafish and human proteins share 36% identity. Gene synteny is partially conserved, with PNKD as upstream adjacent gene in both organisms (Fig. 1C).

According to UniGene, the mouse ortholog Gm216 is specifically expressed in brain, testis and nasopharynx (www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG = Mm&CID = 310460). The Drosophila ortholog CG13243 is specifically expressed in adult testis (flybase.org/reports/FBgn0028903.html), and the protein has been identified in sperm [30]. The high conservation level of C2orf62 across metazoans and its restricted expression in mouse and Drosophila ciliated tissues prompted us to analyze its expression profile in zebrafish.

zC2orf62 is expressed in ciliated cells during embryonic development

In the absence of expression data for zC2orf62 (zgc:153063) in the Zfin database [31], we sought to confirm our initial RT-PCR results by quantitative RT-PCR (RT-qPCR) (Figs 2A, S1). As expected, zC2orf62 mRNA is not detectable at 6 hpf and clearly expressed after 12 hpf, in a development timing that corresponds to optic vesicle and Kupffer's vesicle development and neural keel formation [32][3][4]. zC2orf62 is also expressed between 0 and 4 hpf during maternal and early zygotic periods [33], and strongly in adult testis (about 240 times more than in ovary and embryo).

zC2orf62 expression pattern was investigated by in situ hybridization at 24 and 48 hpf (Fig. 2B). At 24 hpf, zC2orf62 is expressed in the extremity of neural tube formation inside the tail, in brain and in pronephric ducts (future kidney) (Fig. 2B top). At 48 hpf, zC2orf62 expression is restricted to brain and olfactory pits. No expression in pronephric ducts and neural tube can be observed (Fig. 2B down). To further characterize zC2orf62 expression during development, reporter transgenic fish lines expressing EGFP under the control of zC2orf62 promoter were established (Fig. 2C). Fluorescence starts to be observed around 12 hpf in Kupffer's vesicle. At 28 hpf, fluorescence was visualized in brain, neural tube and pronephric ducts, consistent with in situ results, as well as in the olfactory placode and the eye (Fig. 2C). At 48 hpf, expression is detected in the ciliated cells of the olfactory placode, ear, neuromasts and pronephric ducts, visualized by acetyl-tubulin labeling (Figs 2C, S2A). At 96 hpf, fluorescence was observed in neuromasts, olfactory sensory neurons, and in the ear. Within the ears, the EGFP signal was restricted to sensory patches (anterior and posterior maculae, anterior, lateral and posterior cristae), which contain ciliated hair cells [34] (Figs 2C, S2B).

Taken together, expression profiles obtained with transgenic reporter fish and by in situ hybridization are well correlated and show early zC2orf62 expression in Kupffer vesicle, neural tube, pronephric ducts and brain, followed by expression in sensory structures of the olfactory placode, eye, ear and neuromasts.

Unfortunately, our antibody against human C2orf62 does not cross-react with zC2orf62, preventing us from confirming the expression profile of zC2orf62 at protein level, and defining its

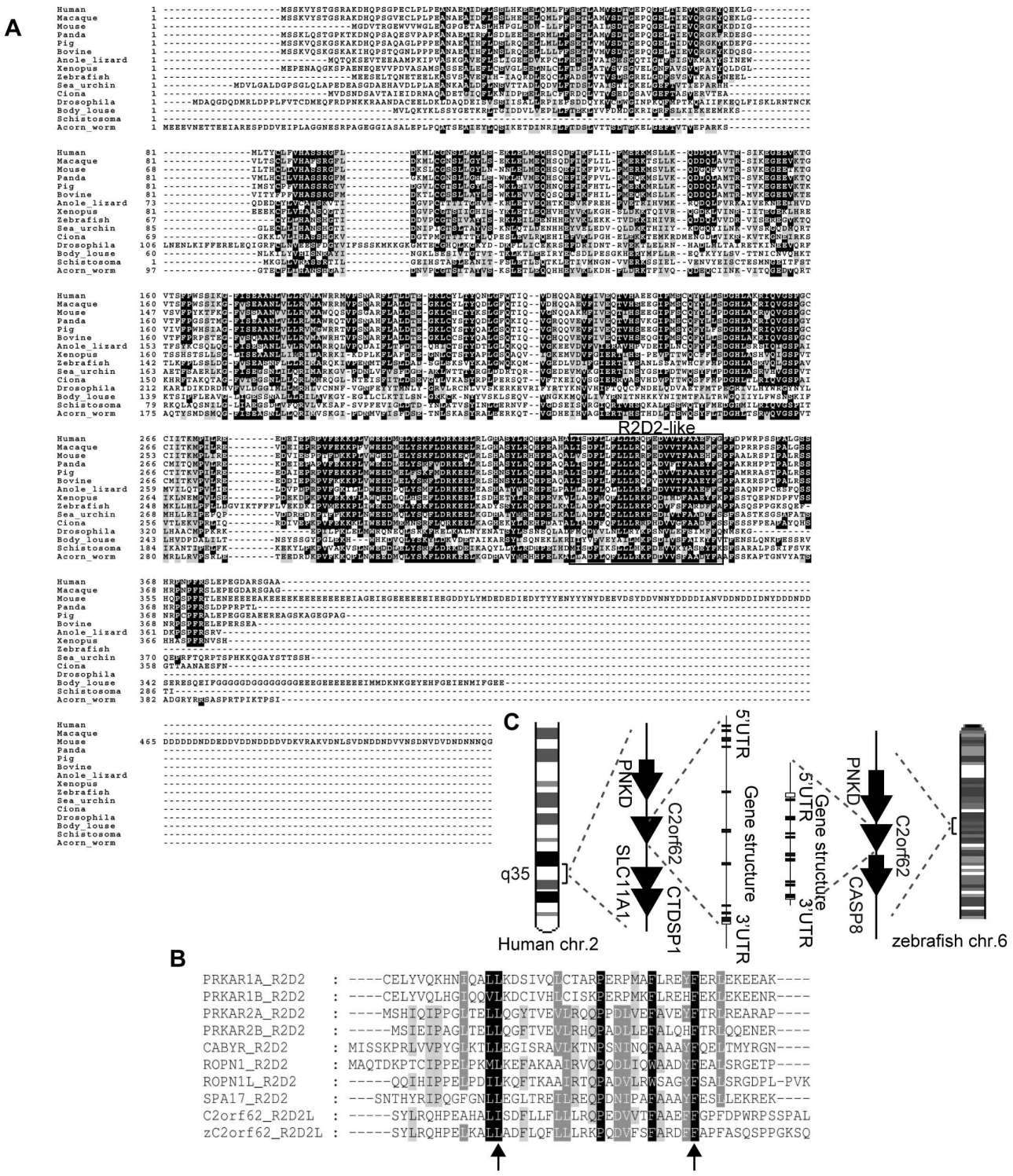

Figure 1. C2orf62 is widely conserved in metazoans. (**A**) Multiple protein sequence alignment of C2orf62 orthologs. Accession numbers of sequences are: **Human**, Q7Z7H3; **Macaque**, F7CRL9; **Mouse**, B9EKE5; **Panda**, XP_002913749.1; **Bovine**, E1BGR4; **Pig**, NP_001177149; **Anole lizard**, XP_003214975; **Xenopus**, Q0IHI3; **Zebrafish**, Q08CH6; **Acorn worm**, XP_002734427.1; **Sea urchin**, XP_793449.3; **Ciona**, XP_002131511.1; **Body louse**, E0VIE2; **Drosophila**, Q9V3B1; **Schistosoma**, G4VER9. Boxed is the R2D2 (PF02197)-like domain (**B**) C2orf62 and zC2orf62 genomic structures and synteny. (**C**) Multiple protein sequence alignment of the R2D2-like domains of human and zebrafish C2orf62, the R2D2 domains of PKA regulatory subunits PRKAR1A, PRKAR1B, PRKAR2A, PRKAR2B and the R2D2 domains of CABYR, ROPN1, ROPN1L and SPA17. The conserved residues that were mutated in zC2orf62 for the functional analysis shown in Fig. S6D are indicated by arrows.
doi:10.1371/journal.pone.0086476.g001

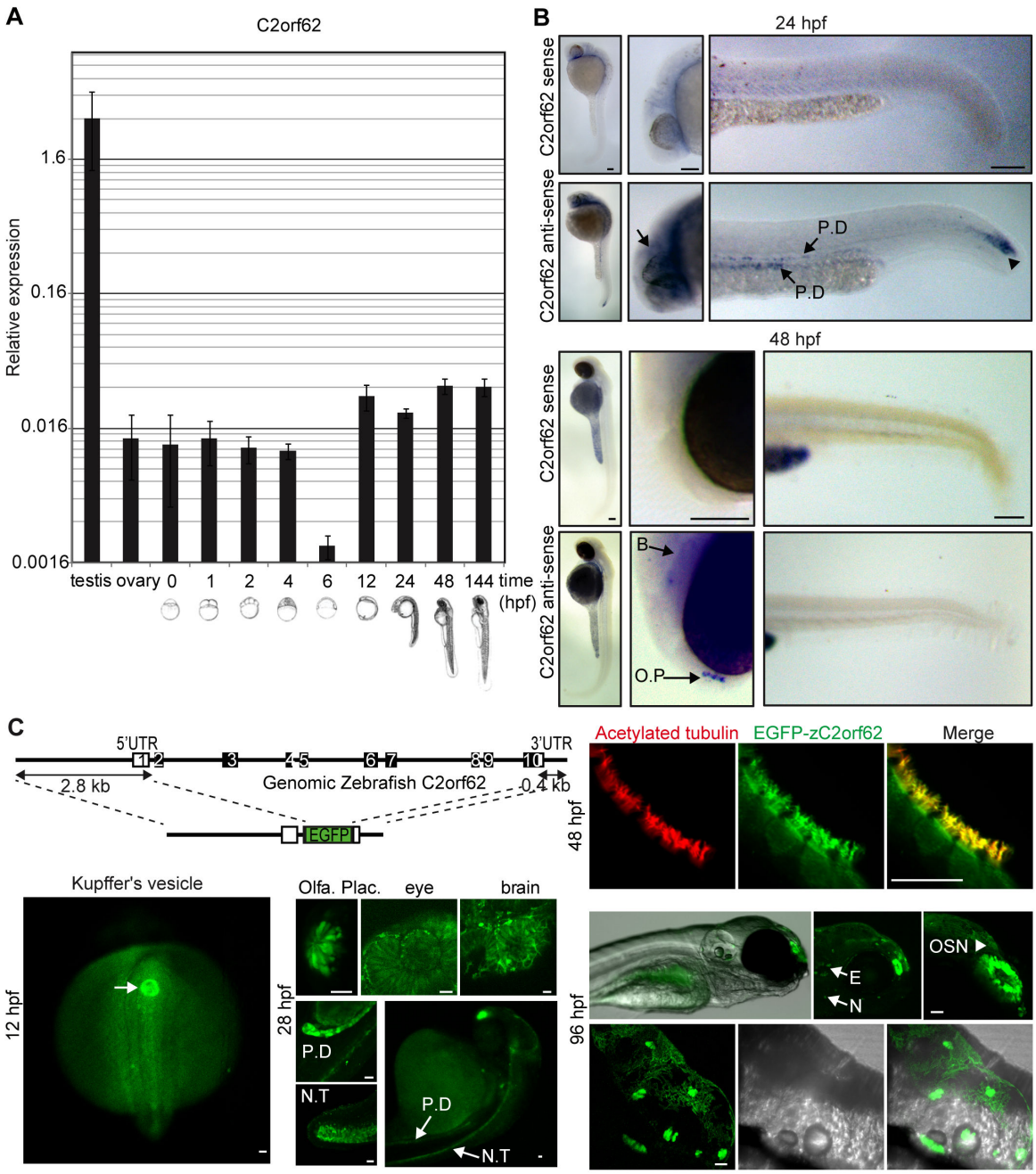

Figure 2. zC2orf62 is expressed in ciliated structures during embryonic development and highly expressed in adult testis. (**A**) zC2orf62 mRNA levels measured by RT-qPCR at various developmental stages. zC2orf62 is highly expressed in adult testis and during almost all embryonic development. It is down-regulated at shield stage (6 hpf) and re-expressed at tail bud (12 hpf) when Kupffer's vesicle forms. (**B**) Whole mount in situ hybridization of zC2orf62 mRNA at 24 hpf and 48 hpf. At 24 hpf, zC2orf62 mRNA is detected at the end of neural tube formation on the tail extremity (arrowhead), in pronephric ducts (P.D.) and in the brain (arrows). At 48 hpf, zC2orf62 is expressed in forebrain (B) and olfactory pits (O.P.) (arrows). Scale bars, 100 μm. (**C**) A construct containing zC2orf62 5′-UTR, 3′-UTR and potential regulatory sequences in which zC2orf62 coding sequence was replaced by EGFP was generated and inserted randomly inside the zebrafish genome. Resulting transgenic EGFP-zC2orf62 reporter fish were observed at 12 hpf using a fluorescent stereomicroscope and at 28, 48 and 96 hpf using a confocal microscope. EGFP is expressed at 12 hpf in Kupffer's vesicle, and at 28 hpf in olfactory placode, eyes, brain, neural tube (N.T), and pronephric ducts (P.D). At 48 hpf, EGFP is expressed in ciliated cells of the olfactory organ (positive for α-acetylated tubulin (red)). At 96 hpf, EGFP is also expressed in neuromast cells (N), olfactory sensory neurons (OSN) and in the ears (E and lower enlarged panel). Scale bars, 25 μm. See also Figs S1, S2.
doi:10.1371/journal.pone.0086476.g002

subcellular location in ciliated cells. Therefore, C2orf62 expression was analyzed in mammalian tissues.

C2orf62 is expressed in mammalian germ cells and ciliated cells

By RT-PCR, we showed that C2orf62 is highly expressed in human testis, placenta, prostate and lung, moderately in ovary and brain, and undetectable in other tissues (Fig. 3A). Genome-wide expression profiling in rat (manuscript in preparation, Chalmel et al.) shows a stronger expression in testis and ovary than in other tissues (Fig. 3C). In rat testis, C2orf62 is highly expressed in pachytene spermatocytes and round spermatids compared to spermatogonia and somatic cells (Fig. 3C), which indicates a specific expression in meiotic and post-meiotic germ cells.

Immunohistochemistry on rat testis shows that C2orf62 protein is enriched in the cytoplasm of spermatocytes at the pachytene stage and concentrated in elongating spermatids (Fig. 3D main section-stage VII). Signal is also visible in the cytoplasm of elongated spermatids (Fig. 3E–stage IX), but there is no accumulation in late spermatids prior to their release into the lumen (stage VII). Signal is absent in Sertoli cells, but unexpectedly present in Leydig cells (Fig. 3D). A similar localization was found in human testis with a high staining in the cytoplasm of round and elongating spermatids and a lower one in pachytene spermatocytes. An intense staining was also visible in the cytoplasm of Leydig cells (Fig. 3F).

C2orf62 is expressed in the human cilia-forming cell lines HEK293T [35], PANC-1 [36] and hTERT-RPE1 [37] but not in HeLa [38], Huh-7 and HOS cell lines devoid of cilia (Fig. 3B). C2orf62 mRNA was also detected in HepG2 cells. Hepatocytes are classically considered to be devoid of cilia, but some cancerous cell lines derived from hepatocytes do form cilia [39]. We verified that, in our hands, HepG2 cells formed cilia upon serum starvation, like PANC-1 and hTERT-RPE1 cells (Fig. 3B).

Although our antibody against human C2orf62 works well in tissues like testis, which express C2orf62 strongly, we failed to detect a specific signal in hTERT-RPE1 or PANC-1 cell lines (data not shown). Therefore, a C2orf62 overexpression strategy was chosen to precise C2orf62 subcellular location in ciliated hTERT-RPE1 cells. V5-C2orf62 shows variable localization in the cytoplasm, nucleus and F-actin rich zones of the plasma membrane, but is always excluded from cilia, as shown by the lack of co-localization with acetylated tubulin (Fig. 4A,B). Same results were obtained in PANC-1 cells (data not shown).

The variability in the observed V5-C2orf62 localization in fixed cells might be explained by a dynamic localization cycle of the protein. To assess this point, overexpressed EGFP-C2orf62 was observed by time lapse microscopy. EGFP-C2orf62 was observed in the whole cell but concentrated in dynamic plasma membrane protrusions that appear and disappear in less than 5 min (Movie S1).

C2orf62 interacts with TTC17 (Tetratricopeptide repeat protein 17)

Since C2orf62 is expressed in brain (Fig. 3A) and potentially involved in development, we screened a human fetal brain cDNA library by yeast two-hybrid using full-length C2orf62 as bait, in order to identify possible partners. Among the 10 interacting protein-coding clones (Table S1), 3 were selected as relevant: PRKRA, which plays a role in ciliogenesis [18], CEP192, which is involved in cell cycle and cilia formation [40][41] and TTC17 because some proteins containing tetratricopeptide repeats are involved in intraflagellar transport or cilia formation [42][43][44]. The full-length sequence of PRKRA and the sequences of the C2orf62-interacting regions of CEP192 (aa 1501–1941) and TTC17 (aa 945–1041) were produced as GST fusion proteins and used to precipitate V5-C2orf62 from a HEK293T cell lysate. Only TTC17 interacted with C2orf62 under these conditions (Fig. 4C).

Since TTC17 has not been characterized yet, we analysed its expression in mammals in order to investigate whether this interaction could be relevant in ciliated cells.

Ttc17 is ubiquitously expressed in rat tissues. Within rat testes, Ttc17 mRNA is detectable in germ cells as well as in somatic cells (Fig. S3A). TTC17 is also expressed in every tested human cell line (Fig. 4D). According to the Human Protein Atlas (HPA) [24], TTC17 protein is mainly located in respiratory epithelium, fallopian tube and epididymis, and localizes to both the cytosol and the plasma membrane in different cell lines (HPA038508). Using the same antibody, we found that TTC17 distribution in mammalian testis closely matches the distribution of C2orf62, being undetectable in Sertoli cells, present in germ cells and enriched in spermatocytes, both in rat and in human (Fig. S3B).

Although the HPA038508 antibody gave a strong nuclear background signal, a specific filamentous staining that was abolished by siRNA against TTC17 could be observed in hTERT-RPE1 cells (Fig. 4E). The staining was more intense in PANC-1 cells and completely abolished by siRNA. The observed localization (Fig. 4E) is similar to what was reported by HPA in A-431 and U-251 MG cells. Interestingly, the signal detected in hTERT-RPE1 cells was strongly enhanced in mitotic cells, from metaphase to telophase (Fig. S3C).

TTC17 and C2orf62 were co-expressed in hTERT-RPE-1 cells as mCherry and EGFP fusion proteins, respectively, and observed by confocal microscopy. As previously described using EGFP-C2orf62 alone (Movie S1), both proteins were found to localize in discrete parts of the cytoplasm and in cell protrusions (Fig. 4F). Colocalization was confirmed by FRET analysis (Fig. 4F), which indicates that C2orf62 and TTC17 may interact in ciliated cells.

C2orf62 and TTC17 are involved in ciliogenesis in human cells

The possible implication of C2orf62 and TTC17 in ciliogenesis was investigated by RNA interference in human ciliated cells. Ciliogenesis was induced by serum starvation 24 h after transfection with control siRNA or with siRNA against C2orf62, TTC17 or MAPRE1/EB1, a microtubule (MT) plus-end-tracking protein involved in several microtubule-dependent cellular processes, including primary cilia assembly [45], mitosis and cell migration [46]. siRNA against C2orf62, TTC17 or MAPRE1/EB1 reduced the number of ciliated cells obtained after serum starvation, both in hTERT-RPE1 cells (Fig. 5A,B) and in PANC-1 cells (data not shown). The combination of siRNA against C2orf62 and TTC17 has a stronger effect than a double dose of siRNA against C2orf62 (Fig. 5B), which suggests that C2orf62 and TTC17 may contribute to the same biological function.

Ciliogenesis in hTERT-RPE1 cells can only be induced by serum deprivation if cells are spatially confined [47]. To test if the observed effects of siRNA on ciliogenesis could be indirectly due to an decrease in cell density, we counted cells 48 h and 72 h after siRNA transfection. siRNA against TTC17 had no effect at either time point. siRNA against C2orf62 had no effect at 48 h and slightly reduced cell numbers to 80% of controls 72 h after transfection, as did siRNA against MAPRE1/EB1 (Fig. S4A). Similar effects on cell growth were previously described on COLO 320 cells using siRNA against MAPRE1/EB1 [46]. In our ciliogenesis experiments, serum starvation was performed

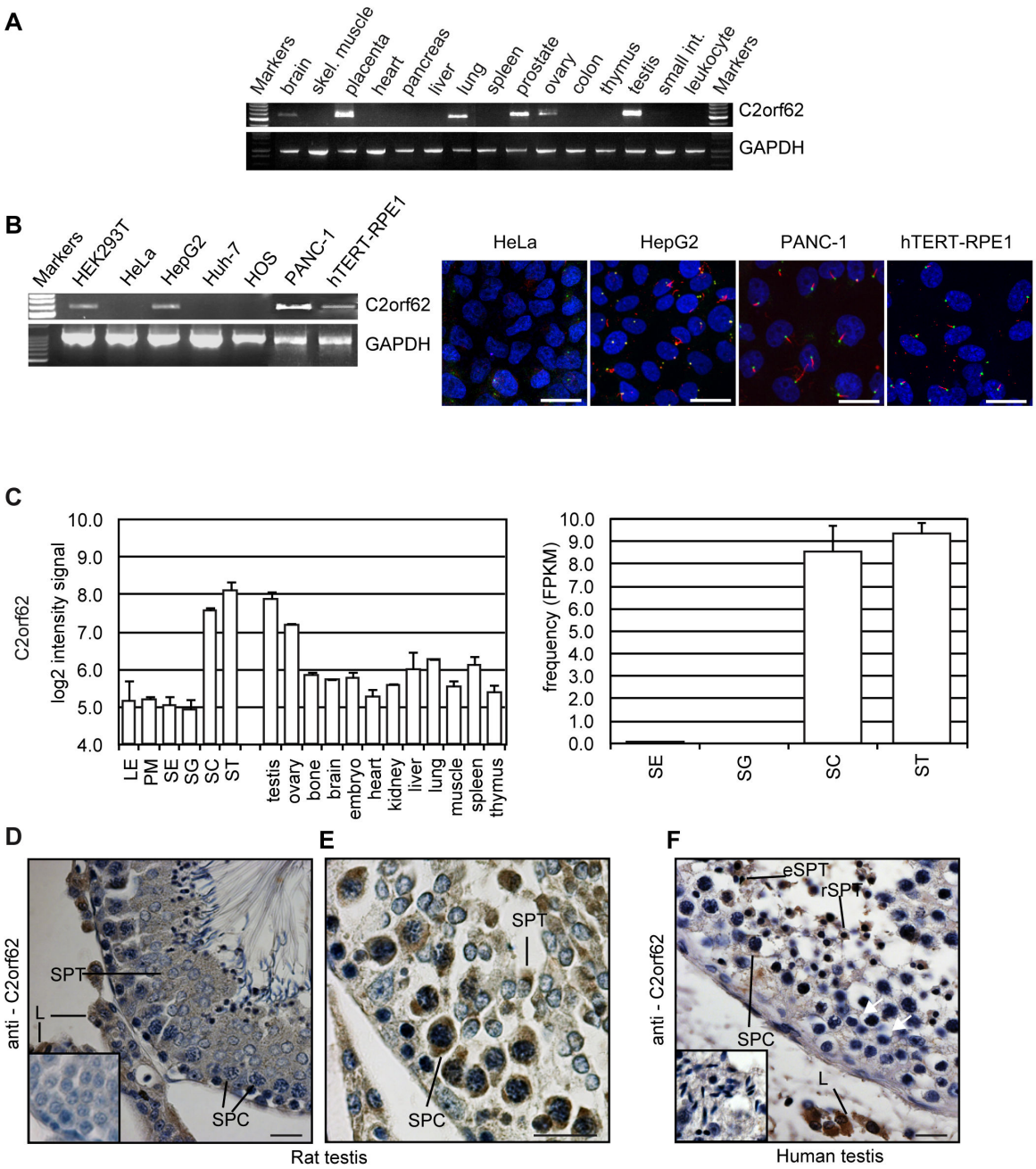

Figure 3. **Mammalian C2orf62 is expressed in ciliated tissues and cell lines.** (**A**) PCR performed on a pre-normalized tissue cDNA panel (Clontech) showing that C2orf62 is highly expressed in testis, placenta, prostate and lung, and moderately in ovary and brain. GAPDH was assessed in parallel for comparison. (**B**) **Left panel**: RT-PCR showing C2orf62 expression in HEK293T, HepG2, PANC-1 and hTERT-RPE1 but not in HeLa, Huh-7 and HOS cells. **Right panel**: HeLa, HepG2, PANC-1 and hTERT-RPE1 cells were serum-starved for 48 hours (hTERT-RPE1 cells) or 72 hours (the 3 other cell lines). Cilia were visualized by immunofluorescence using anti α-acetylated tubulin (red), centrioles by immunofluorescence using anti pericentrin (green) and nuclei by DAPI coloration (blue). In contrast to HepG2, PANC-1 and hTERT-RPE1 cells, HeLa cells were unable to grow cilia. Scale bars, 25 μm. (**C**) Histograms showing that C2orf62 transcript expression levels based on Affymetrix Rat Exon 1.0 ST (left) and Illumina RNA-seq (right) are higher in rat testis and ovary than in other tissues, and higher in spermatocytes (SC) and spermatids (ST) than in somatic testicular cells [Leydig (LE), peritubular myoid (PM) and Sertoli (SE)] and in spermotogonia (SG). (**D,E**) Transverse sections of a testis from an adult rat showing C2orf62 immunoreactivity (**D**) in the cytoplasm of pachytene spermatocytes (SPC) and of early and elongating spermatids (SPT) at stage VII of spermatogenesis (with strong staining in Leydig cells (L)), and (**E**) at higher magnification, in the cytoplasm of pachytene spermatocytes (SPC) and early spermatids (SPT), here at stage IX of spermatogenesis. (**F**) Transverse section of a human testis showing C2orf62 immunoreactivity in the cytoplasm of pachytene spermatocytes (SPC), early/round (rSPT) and elongating spermatids (eSPT). A strong staining in the cytoplasm of Leydig cells (L) is also visible. Inserts: negative controls with preimmune serum. Scale bars, 50 μm.
doi:10.1371/journal.pone.0086476.g003

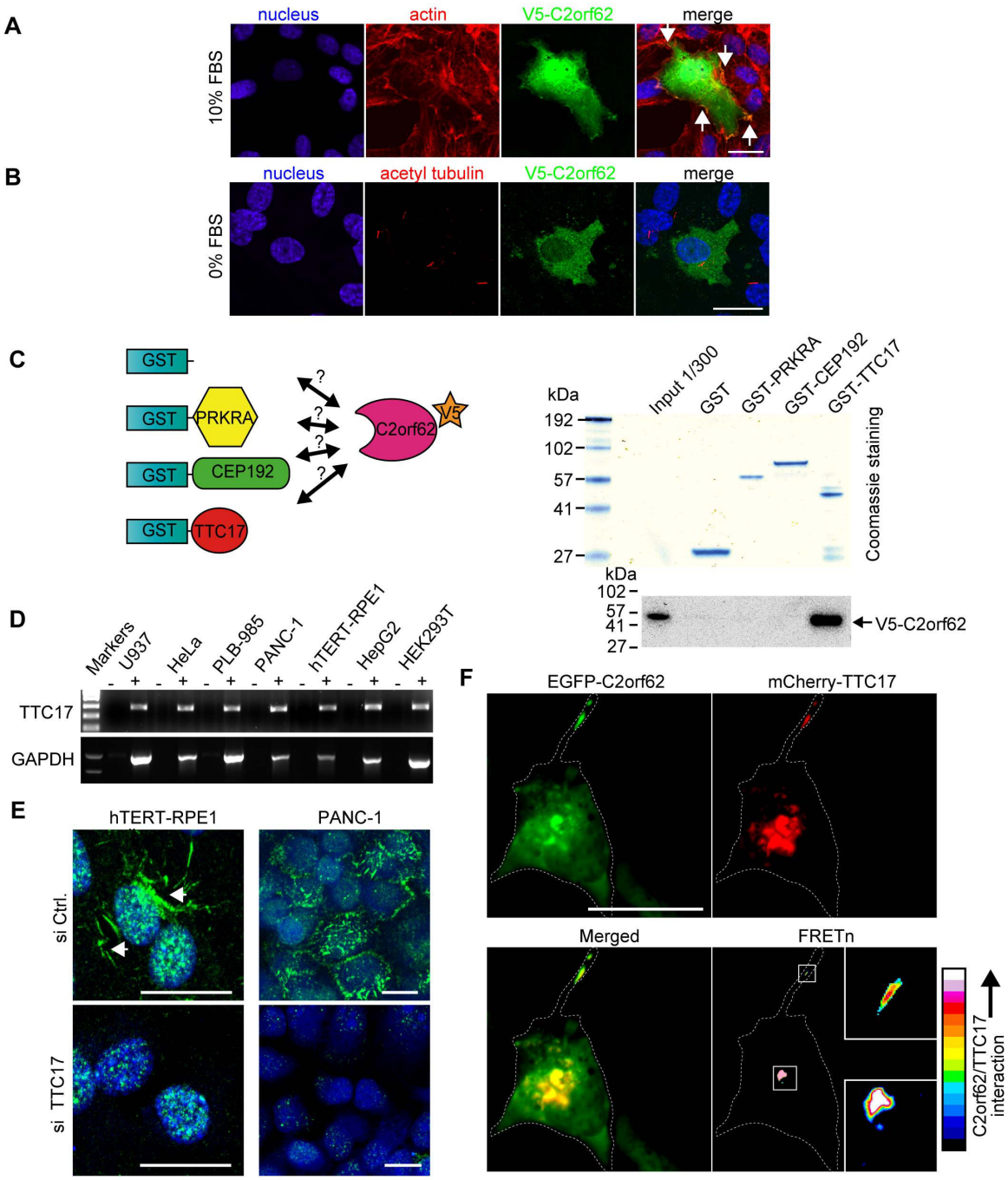

Figure 4. C2orf62 interacts with TTC17 in mammalian cells. (**A**) When overexpressed in hTERT-RPE1 cells grown at 10% FBS, V5-C2orf62 localizes in the cytoplasm, nucleus and in F-actin-rich zones of the plasma membrane (arrows). See also EGFP-C2orf62 live cell imaging in movie S1. (**B**) Ciliogenesis was induced by serum starvation in hTERT-RPE1 cells transfected with V5-C2orf62. V5-C2orf62 (green) was not detected in cilia visualized using anti α-acetylated tubulin (red). Scale bars, 25 μm. (**C**) Beads coated with GST, PRKRA-GST, CEP192-GST and TTC17-GST were used for pull-down experiments on V5-C2orf62-transfected HEK293T cell lysates. V5-C2orf62 only bound to TTC17-GST beads. See also Table S1. (**D**) RT-PCR showing TTC17 expression in every tested cell line (U937, HeLa, PLB-985, PANC-1, hTERT-RPE1, HepG2 and HEK293T). GAPDH was assessed in parallel for comparison. (**E**) Subcellular localization of endogenous TTC17 was studied by immunofluorescence on hTERT-RPE1 and PANC-1 cells transfected either with a control siRNA (si Ctrl) or with siRNA against TTC17 (si TTC17). Despite a strong background staining in the nucleus, a specific staining that disappeared with si TTC17 treatment could be seen in hTERT-RPE1 cells (see also Fig. S3). The staining was more intense, with less background, in PANC-1 cells. Scale bars, 25 μm. (**F**) hTERT-RPE1 cells were co-transfected with EGFP-C2orf62 and mCherry-TTC17 (red) and observed 24h later by confocal microscopy. Both fusion proteins were enriched in the same areas, including cell surface protrusions. The pseudocolor images represent FRET signals corrected for bleed-through using the normalized FRET method (FRETn), with white and red indicating higher interaction levels. Colocalisation was detectable by FRETn, suggesting a stable interaction. Scale bars, 25 μm.
doi:10.1371/journal.pone.0086476.g004

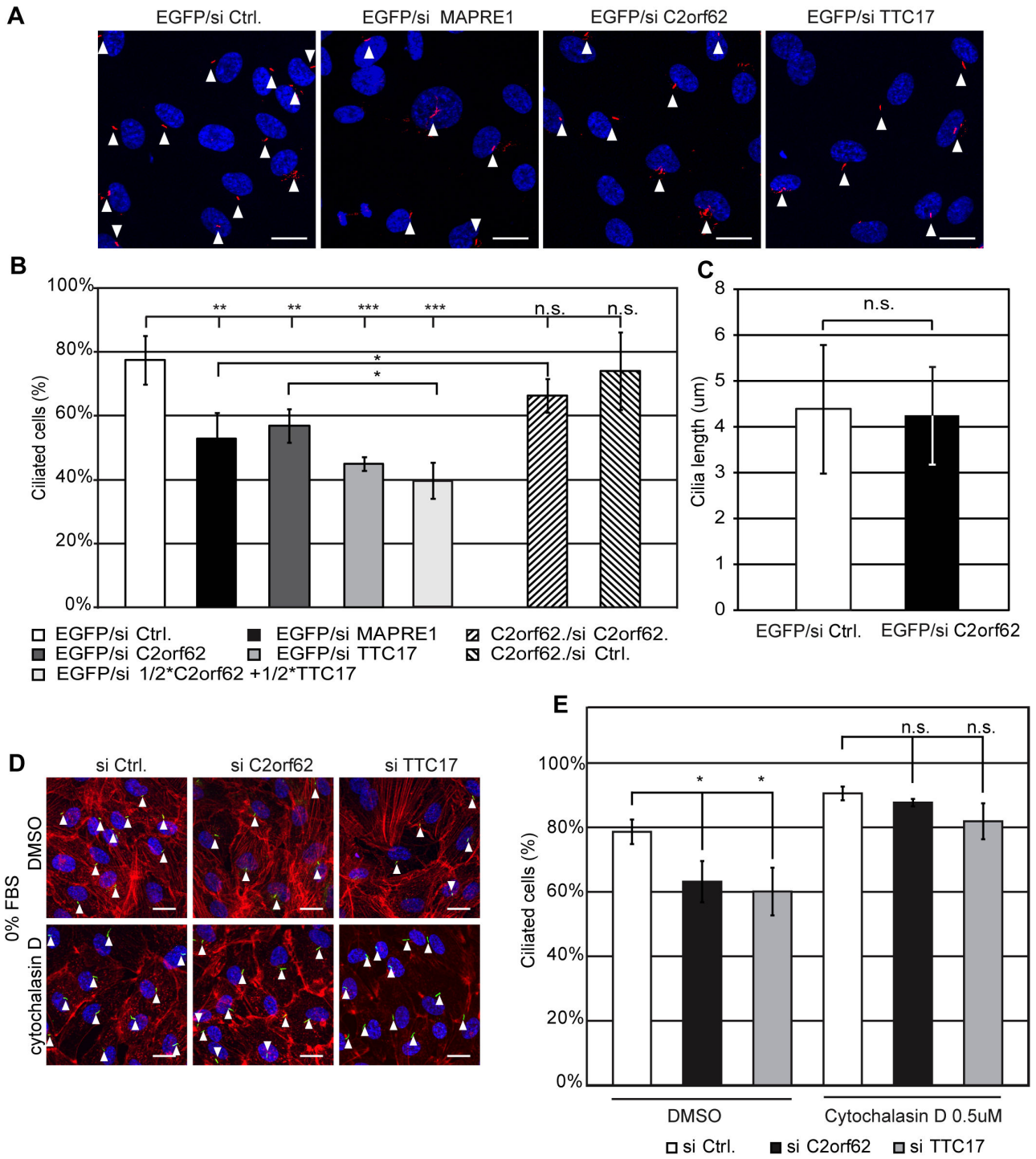

Figure 5. C2orf62 and TTC17 act on ciliogenesis by modulating actin polymerization. Ciliogenesis was induced by serum starvation in hTERT-RPE1 cells transfected with combinations of siRNA and plasmids as indicated **(A-C)** Cells transiently expressing EGFP or C2orf62 were transfected with siRNA against C2orf62, MAPRE1 or TTC17 (50 μM), with both siRNA against C2orf62 and TTC17 (25 μM each), or with a control siRNA. Cilia were visualized by immunofluorescence using anti α-acetylated tubulin (red) and nuclei by DAPI (blue). (**A**) Cells transfected with siRNA against MAPRE1, C2orf62 or TTC17 have less cilia than cells transfected with control siRNA. Scale bars, 25 μm. (**B**) The bar graph shows the percentages of ciliated cells obtained in a minimum of 3 independent experiments with 100 cells counted per condition in each experiment. C2orf62, TTC17 and MAPRE1 siRNA significantly reduce cilia numbers, and effects of C2orf62 and TTC17 siRNA are synergic. Effect of C2orf62 siRNA can be rescued by C2orf62 overexpression (C2orf62 siRNA/Ctrl siRNA**P = 0.0026; MAPRE1 siRNA /Ctrl siRNA **P = 0.0034; TTC17 siRNA/Ctrl siRNA ***P = 0.0005; ½C2orf62 + ½TTC17 siRNA/Ctrl siRNA ***P = 0.0004; C2orf62 + C2orf62 siRNA/EGFP + C2orf62 siRNA *P = 0.035; ½C2orf62 + ½TTC17 siRNA/C2orf62 siRNA *P = 0.0127). **(C)** C2orf62 siRNA does not affect cilia length. (**D, E**) Cells transfected with 50 μM siRNA against C2orf62 or TTC17 or with a control siRNA were serum-starved in the presence or absence of 500 nM cytochalasin D. Cilia were visualized by immunofluorescence using anti-acetylated

α-tubulin antibody (green), F-actin by phalloidin (red) and nuclei by DAPI (blue). Scale bars, 25 μm. (**D**) Both F-actin polymerization and the reduction in cilia numbers induced by siRNA are reversed by cytochalasin D. (**E**) The bar graph shows the percentages of ciliated cells obtained in each condition. siRNA against C2orf62 or TTC17 significantly reduce the number of cilia in DMSO-treated cells (C2orf62 siRNA/Ctrl siRNA *P = 0.032; TTC17 siRNA/Ctrl siRNA *P = 0.031). Cells treated with Cytochalasin D exhibit more cilia than cells treated with vehicle (DMSO) in all conditions (C2orf62 siRNA/C2orf62 siRNA + CytD *P = 0.018; C2orf62 siRNA/C2orf62 siRNA + CytD *P = 0.019; Ctrl siRNA/Ctrl siRNA + CytD P = 0–016). In the presence of Cytochalasin D, siRNA against C2orf62 or TTC17 do not affect the number of cilia. See also effects of C2orf62 siRNA on PANC-1 cells in Fig. S4B.
doi:10.1371/journal.pone.0086476.g005

24 hours after siRNA transfection. At this time point, none of the siRNA had an effect on cell density, excluding the possibility that the observed defect of ciliogenesis would indirectly result from a decreased density of cells at the time of serum starvation. Furthermore, the length of cilia in cells treated with siRNA against C2orf62 was similar to the length of cilia in control cells (4.4 um +/− 1.4 versus 4.25 um +/− 1.1) (Fig. 5C), and similar to the reported length of cilia when starvation is done on confined hTERT-RPE1 cells (5 um) [47]. The lack of effect on siRNA against C2orf62 on cilia length suggests that C2orf62 would be involved in cilia initiation rather than in cilia elongation.

C2orf62 and TTC17 modulate actin polymerization in ciliated human cells

Since actin dynamics have been shown to affect ciliogenesis [17], we tested whether actin structures could mediate some of the observed effects of siRNA against C2orf62 and TTC17 on primary cilium growth. hTERT-RPE1 cells with down-regulated C2orf62 or TTC17 were labeled for F-actin and alpha-acetylated tubulin after serum starvation. The reduction in cilia number induced by siRNA against TTC17 or C2orf62 was accompanied by extensive actin polymerization (Fig. 5D). Similar effects were observed using siRNA against C2orf62 on PANC-1 cells (Fig. S4B). Interestingly, siRNA against C2orf62 also induced moderate actin polymerization on unstarved cells (Fig. S4B). These observations suggest that siRNA against C2orf62 and TTC17 may prevent cilia initiation by modifying actin architecture.

Cytochalasin D is an F-actin destabilizer that enhances ciliogenesis at submicromolar concentrations [17][48]. We verified that cytochalasin D (500 nM) increases ciliogenesis in hTERT-RPE1 cells, as previously described [47]. Whereas siRNA against TTC17 or C2orf62 inhibited ciliogenesis in DMSO-treated hTERT-RPE1 cells, they were without effect in presence of cytochalasin D (Figs 5D,E), confirming that C2orf62 and TTC17 may act on ciliogenesis by interfering with F-actin dynamics.

zC2orf62 and zTTC17 knock-down induce ciliogenesis defects in vivo

In order to confirm the role of C2orf62 and TTC17 in ciliogenesis in vivo, a MO strategy carefully designed to avoid off-target effects [22] was employed. Two MOs were designed for each gene. C2orf62_1 MO targets C2orf62 start codon, C2orf62_2 MO targets the splice site between intron 2 and exon 3 of zC2orf62 pre-mRNA, TTC17_1 MO targets the splice site between exon 1 and intron 1 of zTTC17 pre-mRNA, TTC17_2 MO targets the splice site between exon 2 and intron 2 of zTTC17 pre-mRNA. At 5 ng, C2orf62_1 MO abolished the fluorescence in embryos injected with zC2orf62-EGFP mRNA (Fig. S5D), demonstrating its efficiency. The inhibitory effect of C2orf62_2 MO on zC2orf62 splicing was checked by RT-PCR on embryos injected with 5 ng C2orf62_2 MO (Fig. S5E).

At 24 hpf, injection of C2orf62_1 MO, C2orf62_2 MO, TTC17_1 MO and TTC17_2 MO resulted in similar phenotypic traits with various severity degrees (Fig. 6A), from minor developmental delay to loss of polarity. These traits include a curved body, a lack of defined brain structures or necrosis in the developing brain, and eye formation defects. The severity of the phenotypic traits was dose-dependent (Fig. 6A and data not shown). Simultaneous injection of C2orf62_1 MO and C2orf62_2 MO at low doses (1.25 ng) gave a synergistic effect (Fig. S5A). MO may provoke off-target effects by inducing the p53-dependent cell death program, which can be avoided by p53 co-knockdown [49]. Co-injection of p53 MO did not change the phenotypic traits induced by C2orf62_2 MO, apart from a slightly reduced brain necrosis (Fig. S5B).

Importantly, the phenotype induced by 10 ng C2orf62_2 MO could be partially alleviated by co-injection of 2 ng zC2orf62 mRNA (Fig. S5C). However, zC2orf62 mRNA itself induced a severe phenotype in 14.5% of embryos (Fig. S5C), which prevented us to test if higher doses could rescue the C2orf62_2 MO-induced phenotypic traits totally. Although it was designed to target zC2orf62 splicing and not its coding sequence, C2orf62_2 MO decreased the fluorescence in embryos injected with zC2orf62-EGFP mRNA (Fig. S5D). This unanticipated inhibitory effect of C2orf62_2 MO on zC2orf62 mRNA translation may explain the modest rescue effect of zC2orf62 mRNA.

To check whether the observed phenotypes at 24 hpf were specific developmental defects and not a global delay in fish development, C2orf62_1 MO-injected embryos were also observed at 48 hpf (Fig. S6B) and 72 hpf (Fig. 6B). Like at 24 hpf, the severity of the phenotypes was dose-dependent, from curly-tipped tail, small cardiac edema and head defects, to strong curvature of the whole body, bigger heart edema, and brain necrosis (Fig. 6B, Fig. S6B). We also verified that the phenotypes induced by injection of TTC17_2 MO at 72 hpf was similar to the phenotypes induced by injection of C2orf62_1 MO. Indeed, injection of either MO at 5 ng resulted in about 50% of severe-to-lethal phenotypes at 72 h (Fig. 6B).

To complete the characterization of C2orf62 and TTC17 MO effects, in situ hybridization experiments using *myosin heavy chain 6* (*myh6*) as heart marker [50] were performed at 30 hpf on embryos injected with 2.5 ng C2orf62_1 MO, 5 ng TTC17_2 MO or 5 ng Ctrl. MO. 65% of C2orf62_1 MO-injected embryos and 52% of TTC17_2 injected embryos have heart positioning defects, and 20% of embryos injected with either C2orf62_1 or TTC17_2 display situs inversus, which is a rare event (<1%) in controls (Fig. 6C).

Additional in situ hybridization experiments using *no tail* (*ntla*) and *sonic hedgehog* (*shha*) as markers for the notochord and floorplate [4][6][51][5] were performed at 24 hpf on embryos, either uninjected (n.i.), or injected with 2.5 ng C2orf62_1 MO or Ctrl. MO. More than 92% of C2orf62_1 MO injected embryos display an undulation of the notochord and floorplate at 24 hpf (Fig. S6A).

Reduced axis length, decreased head size, curved-down body shape, heart position defects and a curled posterior tail were reported at 28–30 hpf in embryos injected with MOs against Cordon-bleu, a protein involved in the development of motile cilia [52]. Abnormal body curvature and heart edema were reported at 48 hpf as consequences of the downregulation of nephrocystin-4, a protein involved in ciliogenesis [53]. Cardiac edema and notochord and tail defects were observed at 72 hpf upon downregulation of meckelin, also required for ciliogenesis [54].

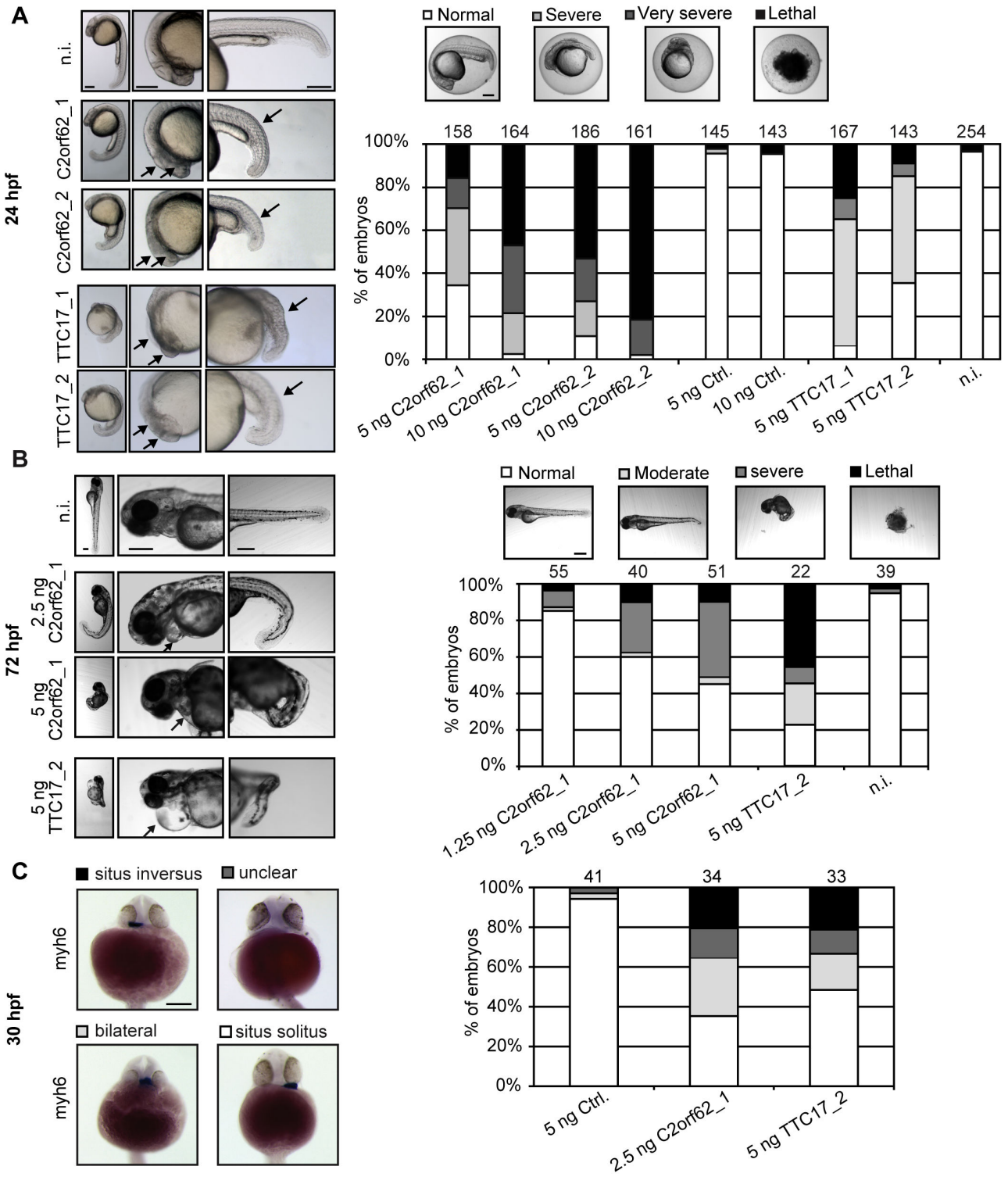

Figure 6. Knockdown of zC2orf62 or zTTC17 in embryos results in body curvature, head defects, and heart positioning defects. Zebrafish embryos were either not injected (n.i.) or injected with different quantities of Ctrl. MO, C2orf62_1 MO, C2orf62_2 MO, TTC17_1 MO or TTC17_2 MO and observed under a stereomicroscope. Scale bars, 200 μm. (**A**) Embryos injected with 5 ng C2orf62_1 MO, C2orf62_2 MO, TTC17_1 MO or TTC17_2 MO display the same phenotype at 24 hpf characterized by a curved body, a lack of defined brain structures or necrosis in the developing brain, and eye formation defects (arrows). Phenotypic traits were classified into four categories, each being associated with one grey code, as reported in the bar graph: normal (white, indistinguishable from controls), severe (clear grey, defect of brain structure formation and short body axis), very severe (dark grey, indistinguishable left/right or front/back polarity), and lethal (black). The bar graph shows the percentages of embryos in each category. Injection of 10 ng C2orf62_1 or C2orf62_2 MO or of 5 ng TTC17_1 MO results in >97% severe-to-lethal phenotypes. Injection of 5 ng C2orf62_1 or TTC17_2 MO results in about 65% severe-to-lethal phenotypes. See also the synergy between C2orf62_1 MO and

C2orf62_2 MO in fig. S5. (**B**) Embryos were either not injected (n.i.), or injected with 1.25 to 5 ng C2orf62_1 MO or with 5 ng TTC17_2 MO, and observed at 72 hpf. Embryos injected with 5 ng C2orf62_1 or TTC17_2 MO display a curved body, incorrectly defined brain structures and heart edema (arrows). Injected embryos were classified into four categories, each being associated with one grey code, as reported in the bar graph: normal (white, indistinguishable from controls), moderate (clear grey, little body curvature and small heart edema), severe (dark grey, strong body curvature, brain defect morphology and edema), and lethal (black). The bar graph shows the percentages of embryos in each category. Injection of 5 ng C2orf62_1 or TTC17_2 MO results in about 50% severe-to-lethal phenotypes. (**C**) Embryos were labeled for *myh6* at 30 hpf and classified into four categories according to the position of their heart (ventral views, scale bars, 200 μm): situs solitus (white), situs inversus (black), bilateral/midline positioning (light grey) and unclear/undeterminable positioning (dark grey). The bar graph shows the percentage of embryos in each category. About 20% of C2orf62_1 or TTC17_2 MO-injected embryos display situs inversus. See Figs S5 and S6 for further characterization of C2orf62_1 and C2orf62_2 MO.
doi:10.1371/journal.pone.0086476.g006

The phenotypes induced by zC2orf62 and TTC17 downregulation at 24–72 hpf are similar to those phenotypes associated with ciliogenesis impairment, suggesting a functional involvement of zC2orf62 and zTTC17 in ciliogenesis in vivo.

In order to directly assess the effect of zC2orf62 knock-down on ciliogenesis, ciliated cells within the olfactory organ were evidenced by α-acetylated tubulin staining. At 48 hpf, embryos injected with 2.5 ng C2orf62_1 MO clearly show a reduction in the number of ciliated cells within the olfactory organ (Fig. S6C).

Altogether, these data show that zC2orf62 and zTTC17 knockdown induce ciliogenesis defects in vivo, confirming the results obtained on human cell lines.

zC2orf62 C-terminal overexpression induces ciliogenesis defects in vivo

Although C2orf62 has no annotated functional domain, the aa 327–352 region is just below the threshold of detection by Pfam of the docking and dimerization domain of protein kinase A II-alpha (R2D2, PF02197) (Fig. 1A). In order to test whether this domain could be important for the function of zC2orf62, we injected embryos with mRNA coding for the zC2orf62 C-terminal part (aa 276–356) and observed them at 24 hpf. As shown in figure S6D, about 75% of injected embryos displayed similar features than embryos injected with C2orf62 MO, including axis length reduction, decreased head size, and curved-down body shape.

L13A and F36A mutations were shown to abolish PRKAR2B dimerization and binding to AKAP75 [55]. Because these residues are conserved in C2orf62 (Fig. 1B), we mutated corresponding zC2orf62 residues (Leu-318 and Phe-341) into alanines and performed mRNA injection experiments. In contrast to injection of wild type C-terminus mRNA, injection of L318A/F341A C-terminus mRNA only had minor effect, since 80% of embryos injected with the mutant construct were normal (Fig. S6D). Taken together, these data suggest that zC2orf62 is involved in ciliogenesis through its R2D2-like domain.

Discussion

In this report, we show that zC2orf62 expression is predominant in tissues rich in motile cilia (Kupffer's vesicle, testis, lung, kidney) and in tissues rich in sensory/immotile cilia (olfactory pits, eye, ears, brain). The distribution of C2orf62 in mammalian tissues was similar, with highest expression levels found in testis. Expression of C2orf62 in human cell lines correlated with their ability to form cilia upon serum starvation. These observations suggest that C2orf62 could be associated with ciliogenesis.

Cilia/flagella are ancestral eukaryotic structures. Whereas proteins involved in their core assembly are conserved across species, regulatory proteins appeared in a stepwise manner during evolution [40][1]. For example, CCP110, a centriolar protein that negatively controls the first steps of primary ciliogenesis [56], is conserved in metazoans but absent in *C. elegans* [57]. Likewise, C2orf62 is absent in *C. elegans*, suggesting a regulatory role in ciliogenesis rather than a direct function in cilia assembly. This is corroborated by the fact that, in ciliated human cells, C2orf62 is not localized on the cilium itself. Furthermore, in mammalian testis, C2orf62 is not present in mature, flagellate sperm cells, but in spermatocytes and in round and elongating spermatids where flagella formation is initiated [58]. These results suggest that C2orf62 would be important for processes occurring prior to or during cilia formation rather than for cilia maintenance or motility.

The early steps of primary ciliogenesis consist in the transfer of the basal body to the plasma membrane in a process that depends on the actin network [11]. Known regulators of actin polymerization have been recently shown to act on ciliogenesis [17][47], and an inhibitory role of branched F-actin in ciliogenesis through modulation of ciliogenic vesicle trafficking is now recognized [48]. In human cells, C2orf62 was found to interact with TTC17, another uncharacterized protein. Down-regulation of either protein in hTERT-RPE1 or PANC-1 cells reduces the number of ciliated cells upon serum starvation while promoting actin polymerization. These effects were not seen in presence of cytochalasin D, an F-actin destabilizer known to induce ciliogenesis. These results provide evidence that C2orf62 acts together with TTC17 to favor primary ciliogenesis by negatively regulating actin polymerization.

Interestingly, in *Xenopus* embryos, c2orf62 was recently found to be strongly upregulated (9.3 fold) by multicilin (MCI), a protein required for multiciliate cell formation in diverse tissues [59]. Other genes found to be upregulated in this screen include foxj1, which is required for motile ciliogenesis and left-right patterning [60], its downstream genes tektins [61] and CCDC78 [61][62], and several genes encoding centriole components. This observation suggests that C2orf62 may be involved in the regulation of motile ciliogenesis as well.

Zebrafish embryos depleted of either C2orf62 or TTC17 display a set of morphological features typical for imperfect motile ciliogenesis such as curly tail, heart positioning defects and cardiac edema [54][53][52] as well as a reduction in the number of primary ciliated cells in the olfactory organ. Thus, C2orf62 and TTC17 probably act on both motile and primary cilia formation.

In contrast to C2orf62, TTC17 is not exclusively distributed in human ciliated cells. TTC17 was found to be ubiquitously expressed in mammals, suggesting that it probably mediates additional cellular functions. Our preliminary data indicate that TTC17 expression is increased in human cells during mitosis. Although siRNA against TTC17 had no gross effect on cell proliferation rates, we cannot exclude that TTC17 may be involved in a cellular process that is linked to cell division. In addition, we showed that zC2orf62 was expressed in the maternal and early zygotic developmental periods of zebrafish, at a time when there is no ciliogenesis but extensive cell divisions [32]. In human cells, siRNA against C2orf62 slightly decreased cell proliferation rates, similarly to siRNA against MAPRE1. Therefore, we cannot exclude that some of the phenotypes induced by

C2orf62 or TTC17 MOs in Zebrafish embryos could be due to cell proliferation defects. It will be particularly important to test the effects of C2orf62 or TTC17 MOs at 0–4 hpf in other F-actin-dependent processes such as gastrulation and epiboly [63][64][65][66].

Injection of mRNA coding for the C-terminal part of zC2orf62 containing the R2D2-like domain induced the same morphological features at 24 hpf than injection of MOs against zC2orf62, suggesting an important function for the R2D2-like domain. R2D2 domains are found at the N-terminus of PKA regulatory subunits, and at the N-terminus of four proteins conserved in ciliated organisms and highly expressed in testes and ciliated cells: CABYR, ROPN1, ROPN1L/ASP and SPA17 [28] [29]. R2D2 domains homodimerize and interact with a short amphipathic helix motif located on AKAP proteins [67][29]. R2D2 domains of PKA RII alpha and RII beta have been shown to target these proteins to the centrosome, microtubules and actin via interaction with AKAPs [68]. Peptides that disrupt interactions between R2D2 domains and AKAPs impair sperm motility [69], suggesting an important function for R2D2 domain-containing proteins in cilia or flagella function. This was recently confirmed for ROPN1 and ROPN1L [70][71].We showed that mutations of Leu-318 and Phe-341, predicted to be required for dimerization and AKAP binding based on homology with other R2D2 domains [55][67], prevented the developmental effects induced by the mRNA coding for the C-terminal part of zC2orf62. Therefore, a functional R2D2-like domain seems to be required for C2orf62 action on ciliogenesis in vivo, although we did not find any interaction between C2orf62 and an AKAP. The precise mechanism of action of this atypical C-terminal R2D2-like domain in ciliogenesis processes remains to be investigated. It could regulate C2orf62 oligomerization state and interaction with other proteins such as TTC17, or target C2orf62 to specific subcompartments enriched in cytoskeletal elements.

Taken together, our observations allow us to add C2orf62 and TTC17 to the growing list of genes directly or indirectly involved in the regulation of the formation of primary and motile cilia in Vertebrates. Due to the versatile role of cilia in human development and physiology, cilia defects can cause multiple severe diseases. Defects in motile cilia can lead to embryonic death, hydrocephalus, heart heterotaxy, or to primary ciliary dyskinesia (PCD), a complex disease with respiratory dysfunction and reproductive sterility, whereas defects in sensory cilia can result in altered development of limbs, polydactyly, retinal degeneration, nephropathies and cognitive impairement [72]. The molecular causes of motile and sensory ciliopathies are still incompletely defined, which complicates their diagnosis. For example, known mutations only account for 50% of all PCD cases. Based on our results, we propose to add C2orf62 and TTC17 to the list of candidate genes for ciliopathies of unknown etiology.

Materials and Methods

Zebrafish maintenance and cell culture

Zebrafish animal experimentation was approved by the Ethical Committee for Animal Experimentation of the Geneva University Medical School and the Canton of Geneva Animal Experimentation Veterinary authority. Wild-type (WT) AB and Casper [73] zebrafish were maintained in standard conditions (27°C, 500 µS, pH 7.5). Embryos obtained by natural intercrossing were staged according to morphology [32]. Ovaries and testes were excised from 3 month-old zebrafish euthanized using Tricaine.

Human cells were grown at 37°C in 5% CO2 in either DMEM with glutamax (HEK293T, HeLa, HepG2 and PANC-1) or DMEM/F12 (hTERT-RPE1) [74], and supplemented with 10% heat-inactivated fetal bovine serum (Gibco, 10270).

Antibodies and staining reagents

Immunofluorescence studies on embryos were performed using rabbit anti-GFP (1:500, Life Technologies), rabbit anti-pericentrin (1:1000, Abcam, ab4448), mouse anti-acetylated tubulin (1:250, Sigma, T7451), and goat anti-rabbit and anti-mouse antibodies (1/400, Life Technologies). Immunofluorescence studies on cells were performed using mouse anti-acetylated tubulin (1:5000), anti-V5 (1:500), anti TTC17 (1:150) and goat anti-mouse and anti-rabbit (1/600) antibodies. Nuclei were stained with 4',6-Diamidino-2-Phenylindole (DAPI), and F-actin with Alexa-fluor 594 phalloidin (1/300, Life Technologies).

Immunohistochemistry experiments were performed using rabbit anti-C2orf62 and anti-TTC17 (1:1000, HPA044818 and HPA038508, kind gifts of M. Uhlen), and non-immune serum (1:1000) as negative control.

Plasmids

C2orf62 cDNA was obtained from Life Technologies (IOH28795). PRKRA, CEP192 (aa 1501-1941) and TTC17 (aa 945-1041) cDNAs were generated by PCR on a mixture of human cDNA (Clontech). cDNAs were subcloned into pENTR/SD/D-TOPO or pENTR/TEV/D-TOPO (Life Technologies) and transferred into EGFP, mCherry, GST or V5 pDEST plasmids using Gateway BP Clonase II. For yeast two-hybrid, the C2orf62 cDNA was cloned into pDBa.

zC2orf62 cDNA obtained by PCR on a mixture of embryo cDNAs was subcloned into pCS2$^+$, pCS2$^+$ EGFP and pCRII TOPO plasmids (Life Technologies). The 3′ and 5′-flanking zC2orf62 DNA regions were amplified by PCR. The reporter plasmid was constructed by inserting EGFP between the 2.8 kb upstream and 0.4 kb downstream sequences of zC2orf62 in pT2KXIG△in (Fig. 2C).

zC2orf62 C-ter (aa 276–356) cDNA was obtained by PCR from zC2orf62 pCS2+. L318A/F341A zC2orf62 C-ter (aa 276–356) cDNA was obtained by PCR overlap extension using primers harboring the mutations. Both cDNAs were cloned into pCS2+ vector.

Genome-wide expression profiling

The Affymetrix Rat Exon 1.0 ST GeneChip dataset includes six enriched populations of testicular cell types (Leydig, peritubular myoid and Sertoli cells, spermatogonia, pachytene spermatocytes and round spermatids) in triplicates, complemented with twelve tissues: ovary, bone marrow, brain, embryo, heart, kidney, liver, lung, muscle, spleen, thymus (3 samples each), and testis (7 samples). GeneChip data were normalized using the Robust Multi-Array Average method [75]. The RNA-seq dataset (Illumina's protocol) includes Sertoli cells, spermatogonia, pachytene spermatocytes and round spermatids (in duplicates). The Tuxedo Suite [76] was used to map RNA-seq-derived reads on the genome, and to assemble and quantify transcripts.

RT-PCR and RT-qPCR

RNA extracted with RNeasy Qiagen columns was treated with DNase (Ambion) and checked with an Agilent Bioanalyser. Reverse transcription was carried out from 2 µg RNA using random hexamers (Promega) and SuperScript III Reverse Transcriptase (Life Technologies). The thermal profile used for

PCR on human cells and tissue cDNAs (Clontech) was: 95°C for 20 s, 58°C for 40 s and 72°C for 34 s, 37 cycles.

q-PCR on zebrafish cDNAs was performed in triplicates in 10 µl samples containing 1µl SYBR green reagent, 200 nM oligonucleotides and 1/20 of total cDNA (50°C for 2 min, 95°C for 10 min, 40 cycles of 95°C for 15 s and 60°C for 1 min) using either Amplicon 1 primers zC2orf62_335F (TGGAGCAGTGTTTGTTTGCAG) and zC2orf62_385R (TGCCTGAATCAGACACGGTC) or Amplicon 2 primers zC2orf62_186F (AGCCTGTTGAAGGATTAAGCTGTTA) and zC2orf62_263R (TGAGTTAATTCACTTTCCTCCATGTC).

Relative levels of RNAs were calculated on the basis of △CT (Cycle Threshold variation) and normalized to the geometric mean of *beta-actin*, *ef1-alpha* [77] and *odc1* RNA levels.

Microinjection

C2orf62_1 MO (5′-GTGCTGCAAATGACAGCATAAGTGA-3′), C2orf62_2 MO (5′-CTTTCCTCCATGTCTTAAAAACTCC-3′), TTC17_1 MO (5′- GACACACTCGCTCACCTGCTGCTGT-3′), TTC17_2 MO (5′-GAACATGAGGGTTAAAATCACCTCT-3′) and Ctrl MO (5′-CCTCTTACCTCAGTTACAATTTATA-3′) (Gene Tools) were dissolved in nuclease-free water and their concentrations determined with NanoDrop.

zC2orf62 mRNA was in vitro transcribed from the zC2orf62 $pCS2^+$ plasmid using the mMessage mMachine System (Ambion) and purified by phenol/chloroform extraction.

MO and mRNA were injected at 1–2 cell stages using 0.1% phenol red in 0.5–2 nl Danieau buffer.

To generate EGFP-zC2orf62 reporter fish, AB and Casper embryos were microinjected with 1 nl of a solution composed of 35 ng/µl Tol2 transposase mRNA, 25 ng/µl EGFP reporter plasmid, and 0.1% phenol red as described [78], resulting in generation number 0 (G0). G1 and G2 embryos were established by successively mating adults with WT fish. EGFP-zC2orf62 carriers were identified by fluorescence stereomicroscopy, and imaged using either a Leica DFC340FX digital camera with Leica software LAF 2.1 or a Zeiss LSM 510 META confocal microscope with LSM Viewer software. ImageJ (rsb.info.nih.gov/ij/) was used for image processing.

Whole-mount in situ hybridization

RNA probes for zC2orf62, *shha*, *ntla* and *myh6* were in vitro transcribed with SP6 and T7 RNA polymerases (Promega) from linearized pCRII TOPO plasmids, and labeled using DIG RNA labeling kit (Roche). Dechorionated embryos were fixed overnight with 4% paraformaldehyde in PBS at 4°C and stored in methanol at –20°C. Embryos were rehydrated in PBS/methanol, washed in 0.1% Tween 20, treated with H2O2 during 30 min, and permeabilized using Proteinase K (10 µg/ml, Roche). Embryos were hybridized overnight at 50°C, washed in increasing concentrations of methanol, and clarified in glycerol as described [79]. Imaging was performed using a MZ16FA stereomicroscope and a DFC420 camera (Leica).

Immunohistochemistry

Immunohistochemical experiments were performed on testes from 90 dpp male Sprague-Dawley rats and on human testes, fixed in Bouin's fixative and embedded in paraffin, as described [80]. Human testes were obtained from patients undergoing therapeutic orchidectomy for metastatic prostate carcinoma. The protocol was approved by the Ethical Committee of Rennes, France (Authorization n°DC-2010-1155 - June 15 2011) and written informed consent was obtained from all donors. Thin sections (5 µm) were deparaffined, rehydrated, and incubated for 1 hr at 80°C in citrate buffer (10 mM pH 6.0) with 0.05% Tween 20 for antigen retrieval. After saturation for 30 min with 1% BSA in PBS, the sections were incubated overnight at 48C with the rabbit polyclonal anti-C2orf62 or anti-TTC17 antibodies used both at a final dilution of 1:1000, in PBS containing 0.1% Tween-20 (v/v) and 1% BSA (PBST-BSA). After several washes in PBS, sections were incubated for 45 min with a secondary biotinylated mouse anti-rabbit antibody (Dako, Trappes, France) at a final dilution of 1:500 in PBS-BSA. Samples were subsequently washed in PBS and incubated for an additional 30 min with a streptavidin-peroxidase complex (Dako) at a dilution of 1:500 in PBS. Immunoreaction was revealed with a diaminobenzidine solution (Sigma). Finally, sections were counterstained with Masson hemalun, dehydrated, and mounted in Eukitt (Labnord, Villeneuve d'Asq, France).

Cell transfection and imaging

Cells grown on coverslips were transfected with plasmids using Fugene (Roche), and 6 hr later with siRNA (Ambion) against C2orf62 (5′-AGACCAUCCAGGUAGACCAtt-3′; s51680), TTC17 (5′-CAGUGAUGAUUAUUCUACAtt-3′; s31447), MAPRE1 (5′-CCUGUGGACAAAUUAGUAAtt-3′; s22674) or with control siRNA (AM-4611) using lipofectamine RNAiMAX Reagent (Life Technologies). Serum starvation was performed 24 hrs later during 72 hr (PANC-1) or 48 hr (hTERT-RPE1). Cells could be treated during the last 8 hrs with 0.5 uM Cytochalasin D (Sigma C8273) or DMSO.

Cells were fixed for 20 min in 4% paraformaldehyde, treated with NH_4Cl during 20 min, permeabilized with 0.1% Triton X-100 in PBS (PBS-T) for 20 min, and incubated in PBS-T with 3% BSA during 1 hr. The coverslips were incubated overnight at 4°C with the primary antibodies, washed in PBS-T, incubated with secondary antibodies, and mounted on slides.

Quantification of cilia was performed on randomly selected cells, based on acetylated tubulin labeling. ImageJ was used for image processing. Cilia lengths were measured following acquisition of *z*-stacks. 3D depictions of cilia were reconstructed using Imaris (Bitplane scientific software).

To estimate hTERT-RPE1 cell proliferation rates, equivalent numbers of cells were plated in 12 well plates, transfected with siRNA 24 hours later, and counted using a hemocytometer 48 or 72 hours after the transfection.

For time-lapse imaging, cells were cultured in 96 well black plates with clear bottom (Costar 3603) and transfected with plasmids using Fugene. Imaging was performed every 2 min during 16 hrs on ImageXpress Micro from Molecular Devices equipped with transmitted light at 37°C and 5% CO_2.

For FRET experiments, cells were cultured in individual culture dishes (Fluorodish FD35–100) and transfected with plasmids using Xtreme gene (Roche). Imaging was performed in a controlled atmosphere chamber at 60X magnification under oil immersion, using a Nikon A1R confocal microscope running with NIS element AR imaging software v.4.11.01. Using ImageJ, images were successively corrected for background, realigned, converted in 32 bits, and smoothed as described [81]. Then, a threshold and a ratio (mCherry/GFP) were applied. FRET signals were corrected for bleed-through by measuring the donor spectral bleed-trough into the acceptor channel and the direct acceptor excitation, following three spectral configurations as described (Padilla-Parra and Tramier, 2012), and presented as normalized FRET (FRETn). Pseudocolors images were generated on the basis of 16 colors, with the lowest FRET intensity in black and the highest FRET intensities in white and red.

Yeast two-hybrid

The C2orf62 pDBa (Leu) plasmid was transformed in MaV203 yeast strain (MATα; leu2-3,112; trp1-901; his3△200; ade2-101; gal4△, gal80△ SPAL10UASGAL1::URA3, GAL1::lacZ, GAL1::His3@LYS2, can1R,cyh2R) as described [82]. This bait did not show self-activation and was further used for screening. MaV203 cells were transformed with human fetal brain cDNA library (cloned into pEXP502-AD (Trp), Proquest libraries™, Life Technologies), plated onto Synthetic Complete (SC) medium minus Leucine (-L), minus Tryptophan (-W), minus Histidine (-H) + 25 mM 3-amino-1,2,4-triazole (3-AT), and incubated at 30°C for 4–5 days. Positive clones were patched onto SC-WHL + 3-AT in 96-well plates, incubated for 3 days at 30°C and transferred in liquid SC-WL for 3 days at 30°C with agitation to normalize the yeast cell concentration used for the phenotypic assay. Cells were then diluted 1/20 in water, spotted onto selective medium (-WHL+25 mM 3-AT or -WUL) and incubated at 30°C for 4 to 5 days. To perform the β-galactosidase assay, undiluted yeast cells were spotted onto YPD (yeast extract peptone dextrose) medium plates with nitrocellulose filters, and β-galactosidase activity was evaluated one day after. Only the interactors that were positive for the three phenotypes tested (growth on -WHL+25 mM 3-AT, or -WUL or β-galactosidase) were further analyzed.

GST pull down

48 hrs after transfection with C2orf62 pDEST-V5, HEK293T cells were lysed in extraction medium (0.1 M PIPES pH 6.8; 5 mM $MgCl_2$; 150 mM NaCl, 1% Nonidet P40) containing protease inhibitors (Roche).

E. coli BL21 were transformed with the PRKRA-GST, TTC17-GST and CEP192-GST plasmids by heat shock. GST fusion proteins were induced by 1 mM IPTG and extracted by sonication (4 times 15 s) in lysis buffer (10 mM Tris-HCl (pH 8.0); 5 mM $MgCl_2$; 0,15 M NaCl; 1% Triton; 5 mM DTT). After centrifugation at 10,000 *g* for 10 min, the supernatants were incubated with swelled glutathione–agarose beads (Sigma, G4510) during 60 min before extensive washings. V5-C2orf62 cell lysates previously cleared on beads (1 mg protein) were incubated on GST-fusion proteins beads for 2 hrs at 4°C.

Following SDS–PAGE, bound proteins were stained with Coomassie or analysed by Western blot using mouse anti-V5 antibody (AbD Serotec; 1:1000) and alkaline phosphatase conjugated goat anti-mouse antibody (Jackson ImmunoResearch; 1:5000).

Sequence analysis

BLASTP and TBLASTN analysis were performed on UniProtKB (UniProt consortium, 2012) release 2012_07 and the NCBI Reference Sequences (RefSeq) [83]. Multiple sequence alignments were performed using MUSCLE 3.6 with the default parameters [84]. C2orf62 and zC2orf62 sequences were aligned on genomes using BLAT [85] on the UCSC genome browser (genome.ucsc.edu). Domain analysis was performed using Pfam [26].

Statistics

Statistical analyses were performed with Microsoft Excel. Error bars are means +/− s.d. of values from at least 3 independent experiments. Statistical significance (P-value) was calculated using a two-tailed paired t-test (for rescue of MO effects), or a two-sample with unequal variance t-test (for experiments on cells), and is indicated by * $P<0.05$, ** $P< 0.01$, *** $P<0.001$ or n.s. (not significant).

Supporting Information

Figure S1 (related to Fig. 2). zC2orf62 expression pattern during development. Reverse Transcription-quantitative Polymerase Chain Reaction (*RT-qPCR*) performed with Amplicon 2 of zC2orf62 shows the same expression pattern than in Fig. 2. zC2orf62 mRNA is expressed essentially in testis (100 times more than in ovary), but also during almost all development. It is down-regulated at shield stage (6 hpf) and reexpressed at tail bud (12 hpf) when Kupffer's vesicle forms.
(TIF)

Figure S2 (related to Fig. 2). zC2orf62 is expressed in ciliated cells during embryonic development. Transgenic EGFP-zC2orf62 reporter zebrafish were observed using a confocal microscope. (**A**) At 48 hpf, immunostaining of α-acetylated tubulin (red) and EGFP (green) show a co-localization (yellow) in olfactory placode (OP), neuromast cells (N), ear (E) and pronephric ducts (P.D), and in cilia of olfactory sensory neurons (OSN). (**B**) At 96 hpf, EGFP is expressed specifically in ciliated neuromast cells and in hair cell-containing structures of the ear (cristae and macula). Scale bars, 25 μm.
(TIF)

Figure S3 (related to Figs. 3 and 4). Analysis of TTC17 expression in rat and human tissues and cell lines. (**A**) Histograms displaying Ttc17 transcript levels in rat somatic testicular cells [Leydig (LE), peritubular myoid (PM) and Sertoli (SE)], male germ cells [spermatogonia (SG), spermatocytes (SC) and spermatids (ST)] and twelve normal tissues based on the Affymetrix Rat Exon 1.0 ST (left) and Illumina RNA-seq (right) datasets. (**B**) Transverse sections of adult rat testis (left) and human testis (right) showing immunoreactivity in cells with TTC17 antibody. In rat, TTC17 is detected in the germ cell lineage from preleptotene spermatocytes to elongating spermatids, here at stage II-III of spermatogenesis (left panel). Please note the strong staining in both the Leydig cells and capillary endothelium. TTC17 is detected in the cytoplasm of pachytene spermatocytes (SPC), early (rSPT) and elongating spermatids (eSPT) at stage II-III of spermatogenesis (right panel). In human, TTC17 is detected in the cytoplasm of spermatocytes and spermatids. Inserts**:** negative controls using preimmune serum. Scale bars, 50 μm. (**C**) Subcellular localization of endogenous TTC17 was studied by immunofluorescence on hTERT-RPE1 cells transfected either with a control siRNA (si Ctrl.) or with siRNA against TTC17 (si TTC17). Despite a strong background staining in the nucleus, a specific staining (arrows) that disappeared with si TTC17 treatment could be seen in non-dividing cells. This staining was considerably enhanced during mitosis, from metaphase to telophase. Scale bars, 10 μm.
(TIF)

Figure S4 (related to Fig. 5). Effects of C2orf62 siRNA knockdown on cell proliferation and actin polymerization. (A) hTERT-RPE1 cells were counted using a hematocytometer 48 h and 72 h after transfection with the indicated siRNA. The bar graphs show the ratios of proliferation rates of cells transfected with siRNA against MAPRE1, C2orf62 or TTC17 as compared to cells transfected with control siRNA, measured in 4 independent experiments. None of the tested siRNA affected cell proliferation when it was assayed 48 h after transfection, Both si C2orf62 and si MAPRE1 slightly inhibit cell proliferation when it was assayed 72 h after transfection (C2orf62 siRNA/ Ctrl siRNA *P = 0.034; MAPRE1 siRNA/Ctrl siRNA *P = 0.031). (**B**) PANC-1 cells were transfected with siRNA against C2orf62 or with a control siRNA. Cells were either serum-starved for 72 hours (0.5%

FBS) or left in 10% FBS medium, and F-actin was stained using Alexa fluor 594 phalloidin. Scale bars, 50 μm. C2orf62 knockdown results in enhanced actin polymerization in both serum conditions.
(TIF)

Figure S5 (related to Fig. 6). Evaluation of the specificity of C2orf62 morpholinos. Zebrafish embryos were injected with different quantities of C2orf62_1 or C2orf62_2 MO and observed under a stereomicroscope. (**A-C**) Phenotypic traits at 24 hpf were classified into four categories, each being associated with one grey code, as reported in Fig. 6A: normal (white, indistinguishable from controls), severe (clear grey, defect of brain structure formation and short body axis), very severe (dark grey, indistinguishable left/right or front/back polarity), and lethal (black). The bar graphs show the percentage of embryos in each category. (**A**) Although the injection of 1.25 ng C2orf62_1 or C2orf62_2 MO has no effect on embryo morphology, the co-injection of 1.25 ng C2orf62_1 MO and 1.25 ng C2orf62_2 MO results in 47% of severe-to-lethal phenotypes, showing a strong synergy between both MOs. (**B**) Embryos were injected with 5 or 10 ng C2orf62_2 MO together with 0 (buffer alone), 15 or 30 ng P53 MO. Coinjection of P53 MO does not significantly change the phenotypes induced by injection of C2orf62_2 MO alone. (**C**) Embryos were injected with 10 ng C2orf62_2 MO and/or 2 ng full-length zC2orf62 mRNA. The percentage of normal (white) C2orf62_2 MO-injected embryos is significantly higher in the presence of mRNA (1.96% for C2orf62_2 MO alone and 12.53% for C2orf62_2 MO + RNA, $P = 0.0044$, Student's *t* test). (**D**) Embryos were co-injected with 680 pg C2orf62-EGFP *mRNA* and 5 ng Ctrl. MO, C2orf62_1 MO or C2orf62_2 MO and observed 6 h later under a fluorescence stereomicroscope. EGFP signal was abrogated in all C2orf62_1 MO-injected embryos (star marked), and unexpectedly in 33% of C2orf62_2 MO-injected embryos. (**E**). RT-PCR analysis of total RNA extracted from 24 hpf embryos after injection of 5 ng of Ctrl. MO or C2orf62_2 MO shows that C2orf62_2 MO alters zC2orf62 splicing.
(TIF)

Figure S6 (related to Fig. 6). Knockdown of zC2orf62 results in notochord undulation and ciliogenesis defects in the olfactory organ. (**A**) In situ hybridization of *ntla and shha* at 24 hpf shows that more than 90% of C2orf62_1 MO-injected embryos have an undulated notochord, in contrast to controls (lateral views, scale bars, 200 μm). (**B**) Embryos were either not injected (n.i.) or injected with 5 or 10 ng C2orf62_1 MO and observed at 48 hpf. C2orf62_1 MO-injected embryos display a curved body, heart edema (arrow), and a small head with incorrectly defined brain structures, small eyes and small ears (dashed circle). Scale bars, 200 μm. (**C**) Embryos were injected either with 2.5 ng Ctrl. MO or C2orf62_1 MO, and ciliated cells from the olfactory organ (OO) were visualized using anti α-acetylated tubulin at 48 hpf. The olfactory organs of C2orf62_1 MO-injected embryos are smaller and display less ciliated cells than controls. Scale bars, 50 μm. (**D**) Embryos were injected with 850 pg mRNA coding for the C-terminal part (aa 276-356) of zC2orf62,either wild type (WT C-ter) or alanine-mutated on Leu-318 and Phe-341 (L318A/F341A C-ter), and observed 24 h later. Phenotypic traits were classified into four categories, each being associated with one grey code, as reported in Fig. 6A: normal (white, indistinguishable from controls), severe (clear grey, defect of brain structure formation and short body axis), very severe (dark grey, indistinguishable left/right or front/back polarity), and lethal (black). The bar graph shows the percentage of embryos in each category. The percentage of normal (white) embryos is significantly higher for L318A/F341A C-ter mRNA-injected embryos than for WT C-ter mRNA-injected embryos (25.7% for WT C-ter mRNA and 80.6% for L318A/F341A C-ter mRNA $P = 0.0078$, Student's *t* test).
(TIF)

Table S1 (related to Fig. 4). Yeast two-hybrid results. 10 positive clones obtained in the yeast two-hybrid screen were identified by sequencing. Clones labeled as "Not relevant" are typical yeast two-hybrid artefacts.
(DOCX)

Movie S1 (related to Fig. 4). EGFP-C2orf62 localization in hTERT-RPE1 cells. The movie shows the dynamic localization of EGFP-C2orf62 (in green) in hTERT-RPE1 cells 2.25 hours after transfection. The EGFP-C2orf62 fusion protein is localized in the cytoplasm, nucleus and in plasma membrane protrusions.
(DOCX)

Acknowledgments

We thank Mathias Uhlen (Human Protein Atlas, Sweden) for antibodies against C2orf62 and TTC17; Corinne Di Sanza and Luciana Romano from Marguerite Neeman-Arbez's lab for help with zebrafish experiments and fish facility, Joan Goulley and Anne Charollais from Paolo. Meda's lab and René Holtackers from Patrick Meraldi's lab for their advice on cell lines. We also thank Patrick Meraldi, Roland Dosch and Ueli Schibler for critical reading of the manuscript, the team from bioimaging core facility for help with microscopy and image analyses and the genomics platform of the National Center of Competence in Research Frontiers in Genetics, Switzerland for real-time PCR and the whole CALIPHO team for fruitful discussions.

Author Contributions

Conceived and designed the experiments: FB RJF FC JPB CP MNA LL. Performed the experiments: FB RJF IB FL SC FC. Analyzed the data: FB RJF FL FC CP AB LL. Contributed reagents/materials/analysis tools: FC JPB CP MNA AB. Wrote the paper: FB FL FC CP LL.

References

1. Carvalho-Santos Z, Azimzadeh J, Pereira-Leal JB, Bettencourt-Dias M (2011) Evolution: Tracing the origins of centrioles, cilia, and flagella. J Cell Biol 194: 165–175.
2. Ishikawa H, Marshall WF (2011) Ciliogenesis: building the cell's antenna. Nat Rev Mol Cell Biol 12: 222–234. doi:10.1038/nrm3085.
3. Nonaka S, Tanaka Y, Okada Y, Takeda S, Harada A, et al. (1998) Randomization of left-right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein. Cell 95: 829–837.
4. Essner JJ, Amack JD, Nyholm MK, Harris EB, Yost HJ (2005) Kupffer's vesicle is a ciliated organ of asymmetry in the zebrafish embryo that initiates left-right development of the brain, heart and gut. Development 132: 1247–1260. doi:10.1242/dev.01663.
5. Hong SK, Dawid IB (2009) FGF-dependent left-right asymmetry patterning in zebrafish is mediated by Ier2 and Fibp1. Proc Natl Acad Sci U S A 106: 2230–2235. doi:10.1073/pnas.0812880106.
6. Oishi I, Kawakami Y, Raya A, Callol-Massot C, Izpisúa Belmonte JC (2006) Regulation of primary cilia formation and left-right patterning in zebrafish by a noncanonical Wnt signaling mediator, duboraya. Nat Genet 38: 1316–1322. doi:10.1038/ng1892.
7. Caron A, Xu X, Lin X (2012) Wnt/β-catenin signaling directly regulates Foxj1 expression and ciliogenesis in zebrafish Kupffer's vesicle. Development 139: 514–524. doi:10.1242/dev.071746.
8. Dale RM, Sisson BE, Topczewski J (2009) The emerging role of Wnt/PCP signaling in organ formation. Zebrafish 6: 9–14. doi:10.1089/zeb.2008.0563.
9. Singla V, Reiter JF (2006) The primary cilium as the cell's antenna: signaling at a sensory organelle. Science 313: 629–633. doi:10.1126/science.1124534.

10. Dawe HR, Farr H, Gull K (2007) Centriole/basal body morphogenesis and migration during ciliogenesis in animal cells. J Cell Sci 120: 7–15. doi:10.1242/jcs.03305.
11. Molla-Herman A, Ghossoub R, Blisnick T, Meunier A, Serres C, et al. (2010) The ciliary pocket: an endocytic membrane domain at the base of primary and motile cilia. J Cell Sci 123: 1785–1795. doi:10.1242/jcs.059519.
12. Nogales-Cadenas R, Abascal F, Díez-Pérez J, Carazo JM, Pascual-Montano A (2009) CentrosomeDB: a human centrosomal proteins database. Nucleic Acids Res 37: D175–D180.
13. Inglis PN, Boroevich KA, Leroux MR (2006) Piecing together a ciliome. Trends Genet 22: 491–500.
14. Gherman A, Davis EE, Katsanis N (2006) The ciliary proteome database: an integrated community resource for the genetic and functional dissection of cilia. Nat Genet 38: 961–962.
15. Arnaiz O, Malinowska A, Klotz C, Sperling L, Dadlez M, et al. (2009) Cildb: a knowledgebase for centrosomes and cilia. Database (Oxford) 2009: 14.
16. Fliegauf M, Benzing T, Omran H (2007) When cilia go bad: cilia defects and ciliopathies. Nat Rev Mol Cell Biol 8: 880–893.
17. Kim J, Lee JE, Heynen-Genel S, Suyama E, Ono K, et al. (2010) Functional genomic screen for modulators of ciliogenesis and cilium length. Nature 464: 1048–1051.
18. Evangelista M, Lim TY, Lee J, Parker L, Ashique A, et al. (2008) Kinome siRNA screen identifies regulators of ciliogenesis and hedgehog signal transduction. Sci Signal 1: ra7.
19. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, et al. (2011) The human proteome project: current state and future direction. Mol Cell Proteomics 10: M111.009993.
20. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, et al. (2012) neXtProt: a knowledge platform for human proteins. Nucleic Acids Res 40: D76-83.
21. Lee JE, Silhavy JL, Zaki MS, Schroth J, Bielas SL, et al. (2012) CEP41 is mutated in Joubert syndrome and is required for tubulin glutamylation at the cilium. Nat Genet 44: 193–199. doi:10.1038/ng.1078.
22. Eisen JS, Smith JC (2008) Controlling morpholino experiments: don't stop making antisense. Development 135: 1735–1743. doi:10.1242/dev.001115.
23. Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, et al. (2012) The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol 30: 221–223.
24. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, et al. (2010) Towards a knowledge-based Human Protein Atlas. Nat Biotechnol 28: 1248–1250.
25. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, et al. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res 39: D945–50.
26. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211–22. doi:10.1093/nar/gkp985.
27. Carr DW, Stofko-Hahn RE, Fraser ID, Bishop SM, Acott TS, et al. (1991) Interaction of the regulatory subunit (RII) of cAMP-dependent protein kinase with RII-anchoring proteins occurs through an amphipathic helix binding motif. J Biol Chem 266: 14188–14192.
28. Carr DW, Fujita A, Stentz CL, Liberty GA, Olson GE, et al. (2001) Identification of sperm-specific proteins that interact with A-kinase anchoring proteins in a manner similar to the type II regulatory subunit of PKA. J Biol Chem 276: 17332–17338.
29. Newell AEH, Fiedler SE, Ruan JM, Pan J, Wang PJ, et al. (2008) Protein kinase A RII-like (R2D2) proteins exhibit differential localization and AKAP interaction. Cell Motil Cytoskeleton 65: 539–552.
30. Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, et al. (2010) The Drosophila melanogaster sperm proteome-II (DmSP-II). J Proteomics 73: 2171–2185.
31. Bradford Y, Conlin T, Dunn N, Fashena D, Frazer K, et al. (2011) ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res 39: D822–D829.
32. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF (1995) Stages of embryonic development of the zebrafish. Dev Dyn 203: 253–310. doi:10.1002/aja.1002030302.
33. Kane DA, Kimmel CB (1993) The zebrafish midblastula transition. Development 119: 447–456.
34. Stooke-Vaughan GA, Huang P, Hammond KL, Schier AF, Whitfield TT (2012) The role of hair cells, cilia and ciliary motility in otolith formation in the zebrafish otic vesicle. Development 139: 1777–1787. doi:10.1242/dev.079947.
35. Gerdes JM, Liu Y, Zaghloul NA, Leitch CC, Lawson SS, et al. (2007) Disruption of the basal body compromises proteasomal function and perturbs intracellular Wnt response. Nat Genet 39: 1350–1360. doi:10.1038/ng.2007.12.
36. Nielsen SK, Møllgård K, Clement CA, Veland IR, Awan A, et al. (2008) Characterization of primary cilia and Hedgehog signaling during development of the human pancreas and in human pancreatic duct cancer cell lines. Dev Dyn 237: 2039–2052. doi:10.1002/dvdy.21610.
37. Kim S, Zaghloul NA, Bubenshchikova E, Oh EC, Rankin S, et al. (2011) Nde1-mediated inhibition of ciliogenesis affects cell cycle re-entry. Nat Cell Biol 13: 351–360. doi:10.1038/ncb2183.
38. Alieva IB, Gorgidze LA, Komarova YA, Chernobelskaya OA, Vorobjev IA (1999) Experimental model for studying the primary cilia in tissue culture cells. Membr Cell Biol 12: 895–905.
39. Seeley ES, Nachury MV (2010) The perennial organelle: assembly and disassembly of the primary cilium. J Cell Sci 123: 511–518.
40. Carvalho-Santos Z, Machado P, Branco P, Tavares-Cadete F, Rodrigues-Martins A, et al. (2010) Stepwise evolution of the centriole-assembly pathway. J Cell Sci 123: 1414–1426.
41. Vulprecht J, David A, Tibelius A, Castiel A, Konotop G, et al. (2012) STIL is required for centriole duplication in human cells. J Cell Sci 125: 1353–1362. doi:10.1242/jcs.104109.
42. Ansley SJ, Badano JL, Blacque OE, Hill J, Hoskins BE, et al. (2003) Basal body dysfunction is a likely cause of pleiotropic Bardet-Biedl syndrome. Nature 425: 628–633. doi:10.1038/nature02030.
43. Pathak N, Obara T, Mangos S, Liu Y, Drummond IA (2007) The zebrafish fleer gene encodes an essential regulator of cilia tubulin polyglutamylation. Mol Biol Cell 18: 4353–4364. doi:10.1091/mbc.E07-06-0537.
44. Tran P V, Haycraft CJ, Besschetnova TY, Turbe-Doan A, Stottmann RW, et al. (2008) THM1 negatively modulates mouse sonic hedgehog signal transduction and affects retrograde intraflagellar transport in cilia. Nat Genet 40: 403–410. doi:10.1038/ng.105.
45. Schrøder JM, Larsen J, Komarova Y, Akhmanova A, Thorsteinsson RI, et al. (2011) EB1 and EB3 promote cilia biogenesis by several centrosome-related mechanisms. J Cell Sci 124: 2539–2551. doi:10.1242/jcs.085852.
46. Taniguchi H, Nakamura Y, Tomonaga T, Kondo T, Ichikawa H, et al. (2012) Proteomic-based identification of the APC-binding protein EB1 as a candidate of novel tissue biomarker and therapeutic target for colorectal cancer. J Proteomics 75: 5342–5355. doi:10.1016/j.jprot.2012.06.013.
47. Pitaval A, Tseng Q, Bornens M, Théry M (2010) Cell shape and contractility regulate ciliogenesis in cell cycle-arrested cells. J Cell Biol 191: 303–312.
48. Yan X, Zhu X (2013) Branched F-actin as a negative regulator of cilia formation. Exp Cell Res 319:147-51. doi:10.1016/j.yexcr.2012.08.009.
49. Robu ME, Larson JD, Nasevicius A, Beiraghi S, Brenner C, et al. (2007) p53 activation by knockdown technologies. PLoS Genet 3: e78. doi:10.1371/journal.pgen.0030078.
50. Bakkers J (2011) Zebrafish as a model to study cardiac development and human cardiac disease. Cardiovasc Res 91: 279–288. doi:10.1093/cvr/cvr098.
51. Hawkins TA, Cavodeassi F, Erdélyi F, Szabó G, Lele Z (2008) The small molecule Mek1/2 inhibitor U0126 disrupts the chordamesoderm to notochord transition in zebrafish. BMC Dev Biol 8: 42. doi:10.1186/1471-213X-8-42.
52. Ravanelli AM, Klingensmith J (2011) The actin nucleator Cordon-bleu is required for development of motile cilia in zebrafish. Dev Biol 350: 101–111. doi:10.1016/j.ydbio.2010.11.023.
53. Slanchev K, Pütz M, Schmitt A, Kramer-Zucker A, Walz G (2011) Nephrocystin-4 is required for pronephric duct-dependent cloaca formation in zebrafish. Hum Mol Genet 20: 3119–3128. doi:10.1093/hmg/ddr214.
54. Adams M, Simms RJ, Abdelhamed Z, Dawe HR, Szymanska K, et al. (2012) A meckelin-filamin A interaction mediates ciliogenesis. Hum Mol Genet 21: 1272–1286. doi:10.1093/hmg/ddr557.
55. Li Y, Rubin CS (1995) Mutagenesis of the regulatory subunit (RII beta) of cAMP-dependent protein kinase II beta reveals hydrophobic amino acids that are essential for RII beta dimerization and/or anchoring RII beta to the cytoskeleton. J Biol Chem 270: 1935–1944.
56. Spektor A, Tsang WY, Khoo D, Dynlacht BD (2007) Cep97 and CP110 suppress a cilia assembly program. Cell 130: 678–690. doi:10.1016/j.cell.2007.06.027.
57. Pelletier L, O'Toole E, Schwager A, Hyman AA, Müller-Reichert T (2006) Centriole assembly in Caenorhabditis elegans. Nature 444: 619–623. doi:10.1038/nature05318.
58. Hoyer-Fender S (2012) Centrosomes in fertilization, early embryonic development, stem cell division, and cancer. Atlas Genet Cytogenet Oncol Haematol 16 (4): 306-319. doi:10.4267/2042/47311
59. Stubbs JL, Vladar EK, Kintner C, Axelrod JD (2012) Multicilin promotes centriole assembly and ciliogenesis during multiciliate cell differentiation. Nat Cell Biol 14: 140–147. doi:10.1038/ncb2406.
60. Chen J, Knowles HJ, Hebert JL, Hackett BP (1998) Mutation of the mouse hepatocyte nuclear factor/forkhead homologue 4 gene results in an absence of cilia and random left-right asymmetry. J Clin Invest 102: 1077–1082.
61. Stubbs JL, Oishi I, Izpisúa Belmonte JC, Kintner C (2008) The forkhead protein Foxj1 specifies node-like cilia in Xenopus and zebrafish embryos. Nat Genet 40: 1454–1460. doi:10.1038/ng.267.
62. Klos Dehring DA, Vladar EK, Werner ME, Mitchell JW, Hwang P, et al. (2013) Deuterosome-mediated centriole biogenesis. Dev Cell: 1–10.
63. Cheng JC, Miller AL, Webb SE (2004) Organization and function of microfilaments during late epiboly in zebrafish embryos. Dev Dyn 231: 313–323.
64. Köppen M, Fernández BG, Carvalho L, Jacinto A, Heisenberg C-P (2006) Coordinated cell-shape changes control epithelial movement in zebrafish and Drosophila. Development 133: 2671–2681. doi:10.1242/dev.02439.
65. Solnica-Krezel L (2006) Gastrulation in zebrafish -- all just about adhesion? Curr Opin Genet Dev 16: 433–441.
66. Popgeorgiev N, Bonneau B, Ferri KF, Prudent J, Thibaut J, et al. (2011) The apoptotic regulator Nrz controls cytoskeletal dynamics via the regulation of Ca2+ trafficking in the zebrafish blastula. Dev Cell 20: 663–676. doi:10.1016/j.devcel.2011.03.016.

67. Kinderman FS, Kim C, Von Daake S, Ma Y, Pham BQ, et al. (2006) A dynamic mechanism for AKAP binding to RII isoforms of cAMP-dependent protein kinase. Mol Cell 24: 397–408.
68. Diviani D, Scott JD (2001) AKAP signaling complexes at the cytoskeleton. J Cell Sci 114: 1431–1437.
69. Vijayaraghavan S, Goueli SA, Davey MP, Carr DW (1997) Protein kinase A-anchoring inhibitor peptides arrest mammalian sperm motility. J Biol Chem 272: 4747–4752.
70. Fiedler SE, Sisson JH, Wyatt TA, Pavlik JA, Gambling TM, et al. (2012) Loss of ASP but not ROPN1 reduces mammalian ciliary motility. Cytoskeleton (Hoboken) 69: 22–32.
71. Fiedler SE, Dudiki T, Vijayaraghavan S, Carr DW (2013) Loss of R2D2 proteins ROPN1 and ROPN1L causes defects in murine sperm motility, phosphorylation, and fibrous sheath integrity. Biol Reprod 88: 41.
72. Yuan S, Sun Z (2013) Expanding horizons: Ciliary proteins reach beyond cilia. Annu Rev Genet. 47: 353-76.
73. White RM, Sessa A, Burke C, Bowman T, LeBlanc J, et al. (2008) Transparent adult zebrafish as a tool for in vivo transplantation analysis. Cell Stem Cell 2: 183–189. doi:10.1016/j.stem.2007.11.002.
74. Klebig C, Korinth D, Meraldi P (2009) Bub1 regulates chromosome segregation in a kinetochore-independent manner. J Cell Biol 185: 841–858. doi:10.1083/jcb.200902128.
75. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
76. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7: 562–578. doi:10.1038/nprot.2012.016.
77. Tang R, Dodd A, Lai D, McNabb WC, Love DR (2007) Validation of zebrafish (Danio rerio) reference genes for quantitative real-time RT-PCR normalization. Acta Biochim Biophys Sin (Shanghai) 39: 384–390.
78. Fisher S, Grice EA, Vinton RM, Bessling SL, Urasaki A, et al. (2006) Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nat Protoc 1: 1297–1305. doi:10.1038/nprot.2006.230.
79. Thisse C, Thisse B (2008) High-resolution in situ hybridization to whole-mount zebrafish embryos. Nat Protoc 3: 59–69. doi:10.1038/nprot.2007.514.
80. Com E, Rolland AD, Guerrois M, Aubry F, Jégou B, et al. (2006) Identification, molecular cloning, and cellular distribution of the rat homolog of minichromosome maintenance protein 7 (MCM7) in the rat testis. Mol Reprod Dev 73: 866–877.
81. Kardash E, Bandemer J, Raz E (2011) Imaging protein activity in live embryos using fluorescence resonance energy transfer biosensors. Nat Protoc 6: 1835–1846. doi:10.1038/nprot.2011.395.
82. Walhout AJ, Vidal M (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. Methods 24: 297–306. doi:10.1006/meth.2001.1190.
83. Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40: D130–5. doi:10.1093/nar/gkr1079.
84. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. doi:10.1093/nar/gkh340.
85. Kent WJ (2002) BLAT—the BLAST-like alignment tool. Genome Res 12: 656–664.

Systems Biology in Reproductive Medicine, 2012, 58: 179–190
Copyright © 2012 Informa Healthcare USA, Inc.
ISSN 1939-6368 print/1939-6376 online
DOI: 10.3109/19396368.2012.691943

informa
healthcare

Special Issue: SBiRM: Focus on Proteomics and Reproduction

INVITED REVIEW

Spermatogenesis in mammals: proteomic insights

Sophie Chocu, Pierre Calvel, Antoine D. Rolland and Charles Pineau*

Inserm, U1085, IRSET, University of Rennes I, Campus de Beaulieu, Rennes, France

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

Spermatogenesis is a highly sophisticated process involved in the transmission of genetic heritage. It includes halving ploidy, repackaging of the chromatin for transport, and the equipment of developing spermatids and eventually spermatozoa with the advanced apparatus (e.g., tightly packed mitochondrial sheat in the mid piece, elongating of the tail, reduction of cytoplasmic volume) to elicit motility once they reach the epididymis. Mammalian spermatogenesis is divided into three phases. In the first the primitive germ cells or spermatogonia undergo a series of mitotic divisions. In the second the spermatocytes undergo two consecutive divisions in meiosis to produce haploid spermatids. In the third the spermatids differentiate into spermatozoa in a process called spermiogenesis. Paracrine, autocrine, juxtacrine, and endocrine pathways all contribute to the regulation of the process. The array of structural elements and chemical factors modulating somatic and germ cell activity is such that the network linking the various cellular activities during spermatogenesis is unimaginably complex. Over the past two decades, advances in genomics have greatly improved our knowledge of spermatogenesis, by identifying numerous genes essential for the development of functional male gametes. Large-scale analyses of testicular function have deepened our insight into normal and pathological spermatogenesis. Progress in genome sequencing and microarray technology have been exploited for genome-wide expression studies, leading to the identification of hundreds of genes differentially expressed within the testis. However, although proteomics has now come of age, the proteomics-based investigation of spermatogenesis remains in its infancy. Here, we review the state-of-the-art of large-scale proteomic analyses of spermatogenesis, from germ cell development during sex determination to spermatogenesis in the adult. Indeed, a few laboratories have undertaken differential protein profiling expression studies and/or systematic analyses of testicular proteomes in entire organs or isolated cells from various species. We consider the pros and cons of proteomics for studying the testicular germ cell gene expression program. Finally, we address the use of protein datasets, through integrative genomics (i.e., combining genomics, transcriptomics, and proteomics), bioinformatics, and modelling.

Keywords germ cell, integrative genomics, proteomics, spermatogenesis, testis, translational regulation

Abbreviations PGCs: primordial germ cells; 2DE: two-dimensional gel electrophoresis; MS: mass spectrometry; hnRPA1: heterogeneous nuclear ribonucleoprotein A1; TRA1: tumor rejection antigen; HSC71: heat shock cognate 71kDa protein; AMH: anti Müllerian hormone; SSCs: spermatogonial stem cells; As: Asingle spermatogonia; Apr: Apaired spermatogonia; Aal: Aaligned spermatogonia; maGSCs: multipotent adult germline stem cells; ESCs: embryonic stem cells; TCTP: translationally controlled tumor protein; MCM7: minichromosome maintenance protein 7; MS: mass spectrometry; dpp: days post-partum; 2D-DIGE: two-dimensional difference in-gel electrophoresis; CLPH: casein-like phosphoprotein; TPs: transition proteins; Prms: protamines; MALDI: matrix-assisted laser desorption/ionization; IMS: imaging-mass spectrometry; MudPIT: multidimensional protein identification technology; AMEN: Annotation, Mapping, Expression and Networks.

Introduction

Mammalian spermatogenesis takes place in the seminiferous tubules of the testis and is classically divided into three phases. In the first (proliferative or mitotic) phase, primitive germ cells (spermatogonia) undergo a series of mitotic divisions. In the second or meiotic phase, spermatocytes undergo two consecutive divisions to produce the haploid spermatids. In the third (spermiogenesis) phase, spermatids differentiate into spermatozoa. An intriguing feature of spermatogenesis is that the developing germ cells form associations of fixed composition, or stages, constituting the seminiferous epithelium cycle. The organization and integrity of the seminiferous epithelium are ensured by the somatic Sertoli cells. In rats, Leblond and Clermont [1952]

Received 25 November 2011; accepted 03 March 2012.

*Address correspondence to Charles Pineau, Inserm U1085, IRSET, Proteomics Core Facility Biogenouest, Campus de Beaulieu, 35042 Rennes cedex, France. E-mail: charles.pineau@univ-rennes1.fr

RIGHTSLINK

divided the seminiferous epithelium cycle into 14 stages (I to XIV), with spermiogenesis, defined as the morphological transformation of spermatids into spermatozoa, broken down further into 19 differentiation steps (from 1 to 19). This last stage of germ cell development provides a striking and unique example of cell differentiation involving acrosome formation, nuclear condensation, and flagellum biogenesis. Another intriguing feature of spermatogenesis is the distinct ordering of cell associations along the length of the seminiferous tubules (segments), often referred to as the 'wave of the seminiferous epithelium' [Perey et al. 1961]. A wave encompasses all 14 segments in rat, 12 in mouse, and 6 in human consisting of various cell associations (for review see [De Kretser and Kerr 1988]). A segment is defined as a longitudinal portion of seminiferous tubule corresponding to a single cell association or stage (for review see [Parvinen 1982; Hess and Renato de Franca 2008]).

This unique differentiation process involves control by autocrine, juxtacrine, paracrine, and endocrine factors. It involves the successive activation and/or repression of thousands of genes and proteins, resulting in a testis that is one of the most complex tissues in the body. The use of cell culture approaches to unravel this complexity has proved problematic, due to difficulties culturing the more highly differentiated types of male germ cells. The resulting lack of a natural 'testing ground' has made high-quality genomic analysis particularly important as a prelude to the *in vivo* testing of hypotheses. Advances in molecular biology and genomics have improved our knowledge of spermatogenesis, by allowing the identification of a large number of genes essential for the development of functional male gametes [de Rooij 2001; Matzuk and Lamb 2002].

Significant progress has been made in the large-scale analysis of testicular function, providing greater insight into normal and pathological spermatogenesis. Rapid progress in genome sequencing and microarray development, carrying out genome-wide expression studies have lead to the identification of hundreds of genes spatially and temporally regulated during the ontogenesis of the testis [Wrobel and Primig 2005]. However, although proteomics has now come of age, the investigation of spermatogenesis by proteomics-based strategies remains in its infancy.

In this review, we will present the state-of-the-art large-scale proteomic analyses of spermatogenesis, from germ cell development during sex determination to spermatogenesis in the adult. We will consider the pros and cons of proteomics for studies on the testicular germ cell gene expression program. A summary of studies in the field is presented in Figure 1. Finally, we will discuss the use of protein datasets, through integrative genomics (i.e., combining genomics, transcriptomics, and proteomics), bioinformatics, and modeling.

Expression profiling of embryonic germ cells

In mammals, the embryonic gonads arise as undifferentiated and bipotential structures in the intermediate mesoderm. The expression of the *Sry* (sex-determining region, chromosome Y) gene in the somatic supporting cells of the genital ridges triggers the sexual differentiation of the gonads into testes. The primordial germ cells (PGCs) develop into prospermatogonia, the precursor cells of the male germline [Wilhelm et al. 2007]. Proteomics-based approaches have been less widely used than microarrays for studies of the sexual determination of somatic cells in the embryonic gonads, but have nevertheless been fruitful [Ewen et al. 2009; Sato et al. 2009; Wilhelm et al. 2006].

Using a combination of two-dimensional gel electrophoresis (2DE) and mass spectrometry (MS), Wilhelm and coworkers [Wilhelm et al. 2006] compared the protein profiles of fetal mouse testes and ovaries. Expression of three proteins, namely the heterogeneous nuclear ribonucleoprotein A1 (hnRPA1), the polymorphic tumor rejection antigen (TRA1), and the heat shock cognate 71kDa protein (HSC71), was found to be increased in the male compared to the female gonads. The male specific expression of the *Tra1* gene as well as higher expression levels of Hsc71 and hnRpa1 genes during gonadal development were further confirmed by quantitative real time PCR. Furthermore, HSC71 was phosphorylated to a greater extent in male than in female gonads, highlighting the importance of proteomic approaches for the detection of posttranslational modifications. These three proteins were not known to be associated with gonadal development or sex differentiation so far, and their roles in such events remain to be defined. HnRPA1 belongs to the class of heterogeneous ribonucleoproteins (hnRNPs) that bind heterogeneous nuclear RNAs (hnRNAs). It modulates the expression of specific genes by binding RNA thanks to three RNA binding domains, thus altering their stability. Interestingly, the gene encoding the testosterone-15-alpha-hydroxylase *(Cyp2A5)* was identified as a target of hnRPA1 [Raffalli-Mathieu et al. 2002]. Whereas the enzyme Cyp2A5 is expressed in the embryonic male gonad, its role in this context is unknown. TRA1, also named Hsp90b1 or Glucose-regulated protein 94 (Grp94) is a glycoprotein of the endoplasmic reticulum [Mazzarella and Green 1987]. Grp94 present in the lumen of ER, chaperons the membrane and secreted proteins and impacts their folding. It also possesses particular functions such as calcium binding needed by the conditions in the ER [Marzec et al. 2011]. The HSC71 protein belongs to the HSC70 protein family. Members of this protein family can shield hydrophobic domains of cytoplasmic proteins and function as molecular chaperones to allow an efficient folding [Agashe and Hartl 2000]. Wilhelm and coworkers [2006] found that HSC71, more abundantly expressed in male than in female gonads, is post-translationally phosphorylated. Why would phosphorylation of HSC71 be important for sex differentiation? HSC70 as well as HSC71 proteins were shown to interact with SOX9 in testicular cell lines [Marshall and Harley 2001] and it is well known that *Amh* is a target of SOX9 during testis differentiation. The anti Müllerian hormone (AMH) causes the regression of the Müllerian ducts in males. The transcriptional activity

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

RIGHTSLINK

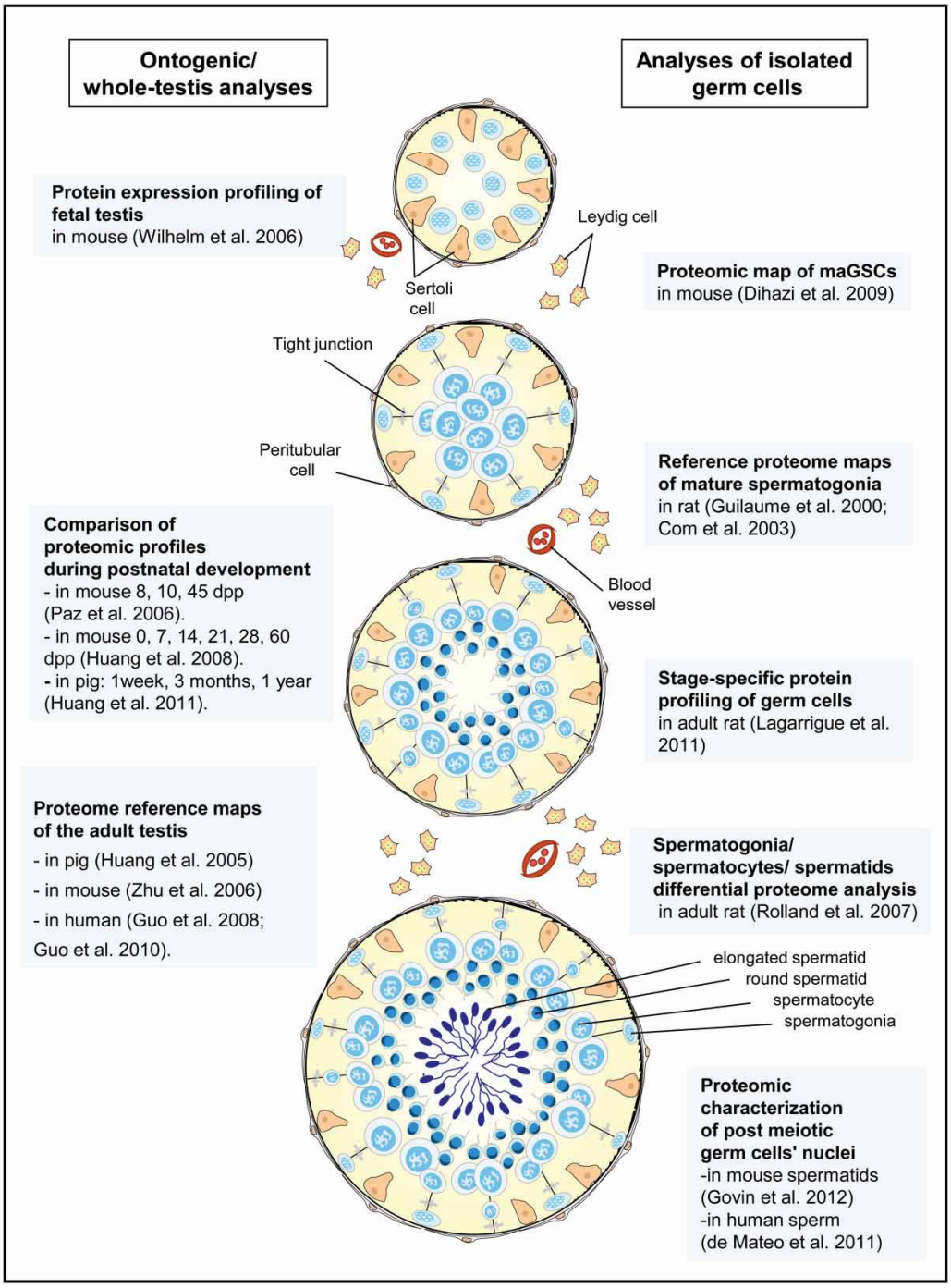

Figure 1. Investigation of spermatogenesis in mammals using proteomic studies. Sections of a seminiferous tubule at different stages of development are represented. The major proteomic studies that have contributed to the knowledge of spermatogenesis are mentioned. Mammalian spermatogenesis takes place in the seminiferous tubules of the testis and is divided into three phases. In the first (proliferative or mitotic) phase, primitive germ cells (spermatogonia) undergo a series of mitotic divisions. In the second (meiotic) phase, spermatocytes undergo two consecutive divisions to produce the haploid spermatids. In the third (spermiogenesis) phase, spermatids differentiate into spermatozoa. maGSCs: multipotent adult germline stem cells.

is enhanced by the additional binding of SF1, GATA4, and WT1 transcription factors to the *Amh* promoter [De Santa Barbara et al. 1998; Hossain and Saunders 2003; Viger et al. 1998].

Moreover, HSC70 was shown to be associated with WT1 [Maheswaran et al. 1998]. According to the authors, it is possible that the closely related proteins HSC70 and HSC71 would stabilize the formation of a SOX9-SF1-WT1 protein complex, crucial for testis differentiation. In this context, it appears that the male specific HSC71 phosphorylation could be an important post-translational regulatory mechanism.

Copyright © 2012 Informa Healthcare USA, Inc.

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

Very little is currently known about the biology of PGCs, due to the small number of these cells in the embryo. A proteomic analysis of cultured PGCs from chicken identified 50 proteins, including growth factors and developmentally regulated proteins [Han et al. 2005]. Most of the proteins found to be expressed in the chicken PGCs such as vimentin, beta-actin, tropomyosin 1 alpha isoforms, tubulin beta-1, desmin, heat shock protein 70, and cathepsin B precursors have homologs in mammals. However, according to the UniProt KB database, the CEphA6 fragment and the ovotransferrin identified in PGCs have no homologs in mammals. Of note no such proteomic study of PGCs has been carried out in mammals, although genome-wide expression analyses have provided consistent information about the genetic programs controlling PGC development (for a review, see [Rolland et al. 2008; Calvel et al. 2010]).

The spermatogonial stem cell niche

Spermatogonial stem cells (SSCs) derived from PGCs, are part of a subset of male germ cells called undifferentiated spermatogonia including Asingle (As) spermatogonia and their progeny cells Apaired (Apr) and Aaligned (Aal) spermatogonia [de Rooij and Russell 2000]. Normal spermatogenesis is maintained by the ability of SSCs to supply continuously the seminiferous tubules of adult testis in differentiating spermatogonia which are the source of all differentiating male germ cells throughout the life of the male. It appears that only a subpopulation of As spermatogonia have stem cell activity [Nakagawa et al. 2007]. These authors estimated that around 10% of As spermatogonia are true SSCs in mice. There are actually two populations of SSCs in the mouse testis. The first is an 'actual (stable) stem cell' population. The actual stem cells have the ability to self-renew and are indeed self-renewing. A second population of 'potential stem cells' also exists; these cells are capable of self-renewing which does not occur in the normal situation. There is a rapid turnover of the potential stem cells in normal testes suggesting their belonging to the transit-amplifying, rather than the dormant population to provide the differentiating spermatogonia [Nakagawa et al. 2007]. To keep their capacity to self-renew, SSCs need to reside in a microenvironment or niche that supplies all the factors and interactions crucial for their survival and development. Thus, the microenvironment of the SSCs in the seminiferous tubule appears to be critical for the production of progeny spermatogonia (for review, see [Caires et al. 2010]). Despite the fundamental role played by spermatogonia in spermatogenesis, little is known about the molecular mechanisms underlying the maintenance of their diverse functions. Cutting-edge technologies for global analyses of gene expression and protein profiles have given scientists a better understanding of the molecular signature and microenvironment of these cells, the SSC niche (for a review, see [Rolland et al. 2008; Caires et al. 2010]).

With only a limited percentage of As spermatogonia being true SSCs, as discussed above, it seems rather difficult to isolate enough SSCs and extract the minimal protein amount for performing proteomic studies and access in depth the SSCs proteome. Bearing in mind these technical limitations, Dihazi and coworkers [Dihazi et al. 2009] build on the fact that under specific culture conditions SSCs acquire pluripotency as they can differentiate into somatic cells of the three germ layers. This is the reason why in their pioneer work Guan and collaborators [Guan et al. 2006] named these cells multipotent adult germline stem cells (maGSCs). Thus, Dihazi and coworkers used a 2-DIGE approach for comparing the reference map proteomes of maGSCs and ESCs mouse cell lines and further identify cell-specific marker by mass spectrometry [Dihazi et al. 2009]. As many as 409 proteins could be identified among which were 166 non-redundant stem cell-associated proteins from maGSCs and ESCs. Of these, only a small subset of 18 proteins was differentially expressed between maGSCs and ESCs, indicating that these two cell types may have rather similar proteomes. Among these, eucaryotic translation initiation factor 5A-1, galectin-1, and lactoylglutathione lyase, involved in cell differentiation were expressed with greater levels in maGSCs compared to ESCs. Interestingly, 27 out of the 166 proteins identified from maGSCs were found to be spermatogonia-associated proteins in the rat [Com et al. 2003]. Finally, Dihazi and coworkers also found that the heat shock cognate 71 kDa protein (Hspa8) was highly expressed and present in different forms in maGSCs and ESCs [Dihazi et al. 2009]. Mass spectrometry further identified Hspa8 on the surface of human ESCs where its expression appeared to be downregulated through differentiation, suggesting that Hspa8 could be a putative marker for undifferentiated human ESCs [Son et al. 2005].

Proteomic analysis, although less widely used than transcriptomic approaches for investigation of the SSC niche, has made a major contribution to our knowledge about developing spermatogonia. Our laboratory has established a reference proteome map of mature spermatogonia freshly isolated from nine-day-old rat testes [Guillaume et al. 2000], leading to the identification of a first set of 53 cytosolic soluble proteins. We compared this small protein subset to the rat spermatogenesis transcriptome dataset performed by Chalmel and coworkers [Chalmel et al. 2007] using the AMEN suite of tools [Chalmel and Primig 2008]. Fifty-one out of the 53 corresponding Gene IDs were expressed in rat spermatogonia and 5 Gene IDs were mitotic and differentially expressed in the testis. This shows that Guillaume and collaborators in this pioneering work mainly identified ubiquitous abundant proteins and only 5 proteins specific to spermatogonia compared to other germ cells (i.e., tropomyosin 4; carbonyl reductase 1; ribosomal protein SA; spermidine synthase, and isocitrate dehydrogenase 2 (NADP+) mitochondrial). Yet, the most interesting proteins expressed by the germline are the so-called 'low copy number proteins' that correspond to discrete proteins in a cell, generally under tight regulation (e.g., transcription factors) and thus potentially crucial for germ cell differentiation. To deepen the analysis of the spermatogonia proteome and detect discrete

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

RIGHTSLINK

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

proteins, overlapping narrow-pH range gels were used to fractionate the cell extracts [Com et al. 2003]. These 'zoom' gels greatly enhance protein separation, providing a much higher resolution than classical broad-pH range gels. Some low-copy number proteins were identified that are involved in regulatory processes (cell defense and detoxification, regulation of gene expression, signal transduction, calcium binding, trafficking, and DNA replication). These two studies provided the first significant repertoire of proteins expressed by rat spermatogonia. We subsequently investigated the roles of some of these proteins in spermatogenesis in more detail, because of their possible implication in proliferation and differentiation [Com et al. 2006; Guillaume et al. 2001a; Guillaume et al. 2001b]. The spatio-temporal pattern of protein and mRNA production in rat testis was investigated for stathmin, which is known to be associated with microtubule dynamics [Guillaume et al. 2001a]. This protein forms aggregates in the cytoplasm of late spermatids, before being eliminated at the time of spermiation. Its production in the germ line may therefore reflect the reorganization of cell structure in germ cells during spermatogenesis. A similar approach was used to analyze the translationally controlled tumor protein (TCTP) in adult human testis, and in neonatal or adult rat testes [Guillaume et al. 2001b]. The TCTP protein is produced in various amounts, in all populations of isolated testicular cells, but its high abundance in spermatogonia suggests a significant role in spermatogenesis. Com et al. [2006] analyzed minichromosome maintenance protein 7 (MCM7), which had previously been identified in rat spermatogonia. They described the spatial distribution of the MCM7 protein and its mRNA within the cells of the rat testis, and their temporal distribution during testis development. They reported differences in the distribution of mRNA and protein for this gene. However, no conclusion was drawn concerning the role of MCM7 in spermatogonia and the germ cell lineage. More recently, a 2-D gel/MS-based differential study was carried out on the testis of the mature dogfish *Scyliorhinus canicula L.*, with the aim of identifying proteins specific to the spermatogonial stem cell compartment [Loppion et al. 2010]. The authors compared protein profiles in the germinative zone with those in the spermatocyte zone. *De novo* sequences were obtained for 33 of the 169 proteins selected for identification by mass spectrometry, but only 16 proteins were identified. One of these proteins, the proteasome alpha-6 subunit was found specifically in the germinative zone of the testes, which contains large isolated spermatogonia. Several of the other proteins identified, (e.g., stathmin) had also been identified in previous proteomic analyses of spermatogonia and other germ cells in rodents [Guillaume et al. 2000; Com et al. 2003; Rolland et al. 2007]. Although preliminary, this study demonstrated the utility of dogfish as a model for proteomic analysis of the spermatogonial stem cell niche. Studies of this type should shed light on the intricate protein networks governing the biology of spermatogonia, provided that genomic data for elasmobranchs are made available.

Postnatal development of male germ cells

The production of spermatozoa by spermatogenesis remains the most widely studied domain of male reproductive biology principally through large-scale experiments. Over the last 10 years, genomics and postgenomics have successfully been applied to the identification of many genes and proteins essential for the development of functional male gametes [Rolland et al. 2008]. The use of transcriptomics and proteomics to study spermatogenesis is entirely logical, because the development of spermatozoa proceeds via a succession of sophisticated and tightly regulated events. These techniques can thus generate very precise snapshots of the molecular networks sequentially involved in spermatogenesis.

The strategies initially adopted were based on the systematic characterization of the proteins present either in isolated germ cells at a particular time point during development or throughout the entire testis. A set of 132 abundant spermatogenic proteins emerged from a systematic identification of germ cell chromatin-associated proteins in the nematode *Caenorhabditis elegans* [Chu et al. 2006]. Functional analysis of these proteins led to the identification of conserved spermatogenesis-specific proteins crucial for DNA compaction, chromosome segregation, and fertility. By analyzing the acid-soluble proteins present in condensing mouse spermatids, Govin and coworkers identified HSPA2 as the first transition-protein chaperone [Govin et al. 2006]. HSPA2 controls the 'histone-to-transition-protein' transition, thereby contributing to the spermatid-specific genome-wide reorganization. A combination of 2-D gel fractionation and MS has been used to establish several proteome reference maps of the testis for *Drosophila* [Takemori and Yamamoto 2009], pigs [Huang et al. 2005], mice [Zhu et al. 2006], and humans [Guo et al. 2010]. Thirty-nine testis-specific proteins potentially important for testis function were identified by 1-D SDS-PAGE combined with RP-LC-MS/MS together with bioinformatic analysis of the human testis proteome [Guo et al. 2008]. Guo and coworkers subsequently highlighted differences in the mass and pI of the proteins produced in the human testis, due to alternative splicing and various forms of post-translational modification, mostly involving phosphorylation [Guo et al. 2010].

These studies generated large sets of proteins thought to be important for testicular function, but no information about the spatio-temporal regulation of these proteins during the course of spermatogenesis was provided. With the objective of obtaining a more accurate view of spermatogenesis, several groups have since used more sophisticated strategies, mostly based on differential expression analyses during the course of spermatogenesis. Comparative proteome profiles for mouse testis have been established for specific time points in the first wave of spermatogenesis, by several groups [Huang et al. 2008; Paz et al. 2006]. Paz and coworkers [2006] compared the proteomic profiles of the soluble proteins present in the testes of mice 8, 18, and 45 days post-partum (dpp). They identified 44 proteins or

Copyright © 2012 Informa Healthcare USA, Inc.

RIGHTSLINK

variant forms displaying differential expression during the course of development. Similarly, Huang and coworkers [2008] compared the testis proteomes of mice 0, 7, 14, 21, 28 and 60 dpp. They identified 257 proteins displaying differential expression, potentially involved in the initiation of mouse spermatogenesis. These proteins included AOP1A and GSTM2, which appeared to be down regulated, and PGK2 and PRDX4, which appeared to be up regulated during testis development [Huang et al. 2008]. Differential protein expression during postnatal development has also been investigated in pig testes [Huang et al. 2011]. Huang and coworkers [2011] studied testes from four pigs each at the ages of one week, three months, and one year. They found that 108 proteins were differentially expressed; 90 of these proteins were identified by mass spectrometry and sorted on the basis of differences in abundance at different developmental stages.

Another widely used strategy involves comparing the levels of particular proteins between different categories of purified germ cells. There are two main reasons for the choice of such an approach rather than the use of total testis samples at different stages of the first wave of spermatogenesis. First, the use of isolated populations of germ cells rather than a whole organ is more efficient for the identification of low-copy number proteins. Second, this approach makes it possible to establish the cellular origin of the proteins identified, although it must be borne in mind that isolated germ cell populations are never 100% pure and that a specific protein could be expressed by one of the contaminating cell types.

Our group carried out a differential proteome analysis of rat spermatogenesis based on two-dimensional difference in-gel electrophoresis (2D-DIGE) [Rolland et al. 2007]. This study provided the first description of the use of 2D-DIGE for identifying a large set of proteins with relative abundances differing significantly between rat spermatogonia, pachytene spermatocytes, and post-meiotic spermatids. Crude cytosolic protein extracts from spermatogonia, spermatocytes, and spermatids were initially analyzed, with limited success in terms of the differentially expressed proteins identified (35 proteins). Germ cell proteomes were then investigated in more detail, by subjecting cytosolic extracts to chromatographic fractionation into four subproteomes before 2D-DIGE. Independent 2D-DIGE analyses of each subproteome fraction led to the identification of as many as 977 protein spots displaying differential expression, many of which corresponded to single proteins. Based on the hypothesis that proteins displaying highly differential patterns of expression are likely to play a key role in differentiation, we focused on protein spots with mean ratios of at least 2.5 between two cell types. This approach led to the identification of 123 unique proteins with reproducible differential patterns of expression during spermatogenesis [Rolland et al. 2007]. Some of the proteins identified in the male germ line for the first time in this study were then characterized further, in targeted studies. For example, the spermatid-specific casein-like phosphoprotein (CLPH) was subsequently shown to be a calcium-binding disordered protein. CLPH is phosphorylated by casein kinase 2, which plays a major role in sperm head development [Calvel et al. 2009]. Other proteins identified by Rolland and co-workers [Rolland et al. 2007] involved in various biological processes may also play key roles in spermatogenesis and worthy of further investigation. For example, the proapoptotic protein Smac/Diablo is present in large amounts in mouse spermatocytes and spermatids [Tikoo et al. 2002]. Its expression pattern is similar in mouse and rat testes [Rolland et al. 2007; Tikoo et al. 2002] suggesting functional conservation and involvement in spermatogenesis [Vera et al. 2004]. In a recent review, Huang and Sha [2011] announced that they are currently using a shotgun proteomics strategy to study the specific protein expression profiles of tetraploid and haploid germ cells purified by flow cytometry from adult mouse testis. In addition to constructing large-scale tetraploid and haploid germ cell proteomes, these authors aim to characterize in more detail the proteins involved in specific cellular events. For example, as many as 3,507 proteins were identified in tetraploid germ cells and 216 of these proteins had homologs in yeast known to be involved in meiosis. Approaches of this kind are generally powerful and should provide us with valuable information, improving our understanding of the mechanisms governing spermatogenesis.

Very recently, a cutting-edge proteomic strategy has been used to investigate the nuclear proteome of post-meiotic male germ cells (i.e., round, elongating, and condensed spermatids [Govin et al. 2012]. During spermatid elongation and packaging of the haploid genome, nucleosomes are disassembled. Most histone proteins are sequentially replaced by transition proteins (TPs), themselves replaced by protamines (Prms) that are highly basic [Balhorn 2007]. Interestingly, 5-15% of the sperm chromatin remains associated with nucleosomes especially at loci of developmental importance, highlighting the epigenetic role of this differential chromatin distribution [Arpanahi et al. 2009; Hammoud et al. 2009]. To identify and characterize critical actors in the post-meiotic packing of the male genome and in the establishment of its associated epigenome in the mouse, Govin and coworkers [Govin et al. 2012] used a strategy based on the identification of protein pools from male post-meiotic germ cells at two different stages. On the one hand, the authors analyzed acid soluble nuclear extracts constituted of basic proteins including DNA-packaging proteins such as histones, TPs and Prms [Govin et al. 2006], and probably of yet uncharacterized DNA-packaging proteins. On the other hand they searched for acid proteins potentially acting as chaperones for the basic DNA-interacting proteins. For identifying acidic proteins, the authors developed an approach based on TP based-chromatography that consisted of applying nuclear extracts from round spermatids, elongating spermatids, and condensed spermatids to TP-coated CNBr activated beads into a column. Retained proteins were then eluted and analyzed by 1D-SDS-PAGE and MS. This work led to the identification of 70 proteins, 46 identified in the acid soluble extract of elongated condensed spermatids, and 29 by TP chromatography. Five proteins were identified

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

RIGHTSLINK

both in the acid soluble extracts and by TP chromatography. Most of the proteins identified in the acid soluble extract belonged to the histone family or non-histone small basic proteins, and most of the proteins represented in the TP pull down extracts were either chaperone or stress-related factors. As an example, HSPA2, a testis-specific heat shock protein was found in both acid soluble and TP pull down extracts, suggesting that it could act as a chaperone for TPs regardless to its charge.

Govin and coworkers finally confronted their proteomic data with transcriptome datasets of normal mouse tissue or male germ cells available in the Gene Expression Omnibus (GEO) repository, and performed functional analyses using the Gene Ontology. A list of factors involved in the post-meiotic packaging and programming of the male genome was then proposed [Govin et al. 2012]. Interestingly, major constituents of the nuclear proteome of post-meiotic male germ cells include DNA binding proteins (involved in the transmission of the epigenetic information to the embryo) and chaperone factors interacting with these proteins (involved in the incorporation of specific DNA-binding proteins to specific regions of the male genome).

Recent developments in proteomic studies of spermatogenesis

The use of total testis sections may be useful for investigating spermatogenesis. Indeed, the direct monitoring of germ cell protein production within the seminiferous tubules has recently become possible thanks to matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging. This emerging technology is now recognized as a powerful tool for protein detection and identification *in situ,* on thin tissue sections, with no effect on the native distribution or function of the proteins concerned (for a review, see [Lagarrigue et al. 2011]). MALDI imaging-mass spectrometry (IMS) has successfully been used to monitor protein profiles in the seminiferous tubules of adult rat testis [Lagarrigue et al. 2011]. The testis has an anatomy among the most complex of any organ in mammals. Indeed, it was for this reason that the testis was selected by the authors as a relevant model for validating and optimizing their technological developments. Lagarrigue and coworkers [2011] were able to visualize, at a resolution of 20 μm, various stages of germ cell development in the testicular seminiferous tubules and to provide a molecular correlate for the well-established stage-specific classification. Proteins of interest were identified by a top-down mass spectrometry approach, and included the full-length thymosin β-10 and β-4 proteins. The authors were then able to superimpose molecular and immunohistochemistry images, as immunohistochemical analysis confirmed that these two proteins were produced in germ cells in a stage-dependent manner. Thymosin β-10 had already been implicated in spermatid development, but this work provided the first evidence for thymosin β-4 production in the testis. Based on the promising results obtained, we foresee that MALDI imaging-mass spectrometry will have a major impact on studies of spermatogenesis and, more widely, in reproductive research generally, by contributing to our understanding of molecular mechanisms and the diagnosis of reproductive diseases.

Transcriptome versus proteome for the study of spermatogenesis

Proteomics has advanced considerably in recent years and may now be considered to have come of age, but it remains unclear whether it can match transcriptomics technologies in terms of ease of use and consistency of the results obtained. Obviously, transcriptomics still has a greater throughput capacity than proteomics but protein diversity cannot be fully characterized by gene expression analyses alone. Moreover, the known complexity of multilayered gene expression mechanisms in mammals is in part responsible for the discrepancies frequently reported between mRNA and protein abundances. It is therefore obvious that one should not choose between these technologies as a combination of these two approaches is likely to provide the best solution in studies of spermatogenesis, increasing the amount of information obtained concerning various aspects.

In addition to establishing proteome reference maps and identifying new proteins of interest, proteomics-based studies have highlighted discrepancies between whole testis and germ cell transcriptomes and their respective proteomes. Several studies have investigated the correlations between transcriptomic and proteomic data, trying to take into account the mechanisms regulating translation. Using multidimensional protein identification technology (MudPIT), Cagney and coworkers carried out a tissue-profiling experiment on nuclear protein-enriched extracts from eight human tissues [Cagney et al. 2005]. They compared 683 non-redundant protein expression profiles with those obtained in microarray experiments. They found that the testis had the lowest coefficient of correlation between its transcriptome and proteome of any of the organs studied (i.e., 0.138 versus 0.432 for the liver, which displayed the strongest correlation). This very weak correlation may reflect original features of gene/transcript regulation during spermatogenesis.

The communication network linking cellular activities during spermatogenesis is known to be highly complex and is thus less well documented than other biological processes. The mining of lists of proteins generated by systematic or differential proteomics approaches with transcriptomics datasets has the potential to provide valuable insight into complex biological processes, such as spermatogenesis. We are thus currently carrying out an integrative genomics project to decipher the seminiferous fluid proteome by shotgun proteomics. Our objective is to investigate the germ cell (and Sertoli cell) secretome, by mining the list of proteins obtained with either Sertoli cell or germ cell transcriptome datasets. Indeed, it has been known for decades that germ cells modulate somatic Sertoli cell function via diffusible proteins (for a review, see [Jégou 1993]). However, up to now, the impossibility of maintaining germ cells *in vitro*

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

Copyright © 2012 Informa Healthcare USA, Inc.

RIGHTSLINK

makes it difficult to study their secretome. For this reason, the role of germ cells in controlling spermatogenesis has essentially been left on the back burner since it first rose to prominence in the early 1990s (for a review see [Jégou et al. 1999]). This new project is based on the rationale that the seminiferous fluid should contain secreted proteins originating from germ cells and/or Sertoli cells. It should lead to the characterization of new proteins involved in crosstalk between Sertoli cells and germ cells. Interestingly, according to Sato and coworkers it is now possible to obtain differentiated germ cells and functional sperm from neonatal mouse tissue testis under certain organ culture methods and medium conditions [Sato et al. 2011]. No doubt this is a major milestone in the field, but before it becomes routine, alternative strategies will still be needed to study germ cell secretomes.

The further exploitation of proteomics datasets: towards a holistic biology of spermatogenesis

The studies reviewed here have contributed to identify proteins likely to be of importance for particular steps in testicular development or required for male fertility (Table 1). However, most of these studies have failed to obtain useful and meaningful results from the large amounts of data generated, to enhance our overall understanding of the cellular events involved in spermatogenesis. Several strategies can be used focusing on a small number of relevant proteins (genes). These strategies include tissue-profiling experiments for the detection of testis-specific genes and cross-species comparisons to detect genes conserved during evolution. The specific expression patterns and conserved expression profiles of these genes may be associated with essential functions. These approaches have proved effective for the identification of key factors from huge lists of anonymous candidates, but they cannot bridge the gap between the identification of thousands of co-expressed genes or proteins and an understanding of the connections between them.

What should we do next with the protein datasets generated by systematic and differential analyses? Important efforts have been made in recent years to integrate data from large-scale experiments and to develop tools enabling researchers not only to describe a group of genes or proteins with similar expression profiles, but also to propose and develop new hypotheses. One way of analyzing such experiments is to use the gene descriptions of the Gene Ontology (GO) Consortium for functional data mining [Ashburner et al. 2000]. The enrichment of a set of genes/proteins in particular types of GO terms can therefore be assessed and used to demonstrate objectively that specific functions are significantly associated with a given process. This approach facilitates the identification of unexpectedly important pathways and predicts functions for uncharacterized genes. It can provide significant additional insight into the transcriptional profiles observed and/or facilitate the identification of genes or proteins belonging to the same complex from sets of co-expressed genes or proteins.

Another useful bridge for workers in this field will be the rational compilation of large- and small-scale omics data into a reproduction-oriented repository system, such as the GermOnline database (http://www.germonline.org) [Primig et al. 2003; Lardenois et al. 2010]. This knowledgebase compiles studies relevant to the cell cycle, gametogenesis, and fertility. It contains a unique combination of information and incorporates a cross-species systems browser to provide annotations concerning DNA sequence, evolutionary relationships, and gene expression and function. The database, based on the 'Ensemble' genome browser, covers several model organisms and *H. sapiens*. Current efforts in our group to integrate proteome datasets into GermOnline aim at improving this tool further to allow decision support and hypothesis generation.

Facilitation of the interpretation of the large multifaceted datasets generated by high-throughput genome experiments will also require flexible and user-friendly analysis tools. In this context, the AMEN (Annotation, Mapping, Expression and Networks) suite of tools for molecular systems biology [Chalmel and Primig 2008] enables life scientists to manage and explore genome annotation, chromosomal mapping, protein–protein interaction, expression profiling, and proteomics data, without the need for advanced computational skills. We strongly encourage scientists in the reproduction field to adopt the AMEN software that can be downloaded freely at http://sourceforge.net/projects/amen/. The current version provides modules for uploading data from microarray expression profiling experiments, detecting groups of significantly co-expressed genes, and searching for an enrichment of these groups in certain functional annotations. In addition, the software allows the simultaneous visualization of several types of data, such as protein–protein interaction networks, expression profiles, and cellular colocalization. Efforts are currently being made to also integrate this information with the next-generation datasets, such as the testis metabolome and microRNAome.

Conclusion

Proteomics has already led to major breakthroughs in protein research and is continually evolving. Resolution of the wide dynamic range of protein expression within cells remains a major challenge, as there is currently no available technology for the amplification of low-copy number proteins from a biological sample. However, instrument sensitivity is rapidly improving and extensive proteome coverage should soon be attainable in large proteome experiments [de Godoy et al. 2006; de Godoy et al. 2008]. Similarly, the absolute quantification of proteins in complex mixtures is becoming feasible, thanks to strategies, such as the protein standard absolute quantification (PSAQ) technique [Brun et al. 2007]. Tools for studying the plethora of posttranslational modifications, are also rapidly expanding [Hilger et al. 2009]. The SILAC (stable isotope labeling by amino acids in cell culture) approach is a versatile tool for the quantitative comparison of organ and cell proteomes from different mouse strains, including knockout mouse

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

RIGHTSLINK

Copyright © 2012 Informa Healthcare USA, Inc.
Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

Table 1. Key proteins identified for the first time using proteomics in the male gonad, or as differentially expressed in male germ cells.

Protein name	Species	Reference	Localization (a)	Differential expression in GCs	Proteomic approach (b)	UniProt accession number	Gene Ontology term processes
Vimentin	chicken	Han et al. 2005	PGCs	–	2-DE /MS	P09654	
HSPA5, Heat Shock 70-kDa protein 5	chicken	Han et al. 2005	PGCs	–	2-DE /MS	Q90593	negative regulation of apoptotic process - cellular response to glucose starvation - negative regulation of Transforming Growth Factor beta receptor pathway
FGF8, Fibroblast Growth Factor-8 precursor	chicken	Han et al. 2005	PGCs	–	2-DE /MS	Q90722	cell differentiation - activation of Wint signaling pathway - positive regulation of cell division
hnRPA1, heterogeneous nuclear ribonucleoprotein A1	mouse	Wilhelm et al. 2006	FT	–	2-DE /MS	P49312	mRNA alternative splicing via spliceosome - mRNA processing - - nuclear export
TRA1, polymorphic tumor rejection antigen	mouse	Wilhelm et al. 2006	FT	–	2-DE /MS	P08113	anti apoptosis - response to stress/hypoxia - protein folding - ER associated catabolic process - actin rod assembly - regulation of phosphatase activity
HSC71, heat shock cognate 71kDa protein	mouse	Wilhelm et al. 2006	FT	–	2-DE /MS	P63017	response to stress - protein folding - negative regulation of transcription, DNA dependant - regulation of cell cycle
		Dihazi et al. 2009	maGSCs	–	2-DE /MS		
eIF-5A, Eucaryotic translation Initiation factor 5A-1	mouse	Dihazi et al. 2009	maGSCs	–	2-DE /MS	P63242	translation - peptidyl-lysine modification to hypusine - negative regulation of apoptotic process - protein transport
Lactoylglutathione lyase	mouse	Dihazi et al. 2009	maGSCs	–	2-DE /MS	Q9CPU0	glutathione and carbohydrate metabolic processes - anti apoptosis - regulation of transcription - methyglyoxal metabolic process
Galectin-1	mouse	Dihazi et al. 2009	maGSCs	–	2-DE /MS	P16045	cellular response to drug / organo cyclic compound - response to glucose stimulus - positive regulation of kappa B kinase/NFkB cascade - negative regulation of cell/substrate adhesion
HSPA2, Heat shock-related 70 kDa protein 2	mouse	Govin et al. 2012	eSPT, cSPT	–	TPI colomn/MS-acidic extraction /MS	P17156	positive regulation of protein phosphorylation - response to stress - male meiosis (I) - spermatid development - synaptonemal complex disassembly
Nucleoplasmin-3	mouse	Govin et al. 2012	eSPT, cSPT	–	TPI colomn/MS-acidic extraction /MS	Q9CPP0	rRNA processing - rRNA transcription
Fau protein	mouse	Govin et al. 2012	eSPT, cSPT	–	TPI colomn/MS-acidic extraction /MS	Q91V99	translation
EF1G, Elongation factor 1-gamma	mouse	Govin et al. 2012	eSPT, cSPT	–	TPI colomn/MS-acidic extraction /MS	Q9D8N0	translation
Calm I	mouse	Govin et al. 2012	eSPT, cSPT	–	acidic extraction/MS	Q9D6G4	calcium binding - protein binding
CFLI, Cofilin	mouse	Govin et al. 2012	eSPT, cSPT	–	TPI colomn/MS	COF1	actin filament binding - cytokinesis

Continued

RIGHTSLINK

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

Table 1. *Continued*

Protein name	Species	Reference	Localization (a)	Differential expression in GCs	Proteomic approach (b)	UniProt accession number	Gene Ontology term processes
Stathmin	rat	Com et al. 2003; Guillaume et al. 2001a	GCs, abundant in cSPT	yes	2-DE /MS	P13668	nervous system development - brain development - microtubule dynamics - cell differentiation
TCTP, translationally controlled tumor protein	rat	Com et al. 2003; Guillaume et al. 2001b	T, abundant in SPG	yes	2-DE /MS	P63029	spermatogenesis - cell proliferation
DNA replication licensing factor MCM7	human	Com et al. 2003; Com et al. 2006; Rolland et al. 2007	GCs, abundant in SPG	yes	2-DE /MS	P33993	DNA replication - cell proliferation -response to DNA damage - regulation of phosphorylation
CLPH, Casein-like phosphoprotein, Calcium-binding and spermatid-specific protein 1	rat	Rolland et al. 2007; Calvel et al. 2009	SPT	yes	2D-DIGE/MS	Q68FX6	spermatogenesis
Smac/Diablo	mouse	Rolland et al. 2007; Tikoo et al. 2002	SPC, SPT	yes	2D-DIGE/MS	Q542V8	induction of apoptosis
GADPH, Glyceraldehyde-3-phosphate dehydrogenase	rat	Rolland et al. 2007	S, SPG	yes	2D-DIGE/MS	P04797	glycocysis - apoptotic process - oxydation reduction process
Polypyrimidine tract-binding protein 2	rat	Rolland et al. 2007	SPC, SPT	yes	2D-DIGE/MS	Q66H20	mRNA processing - RNA splicing
Grp58, Protein disulfide-isomerase A3	rat	Rolland et al. 2007	SPG, SPC, abundant in SPT(acrosome)	yes	2D-DIGE/MS	P11598	cell redox homeostasis - positive regulation of apoptotic process
AOP1A/PRDX3, Thioredoxin-dependent peroxide reductase, mitochondrial	mouse	Huang et al. 2008	T, abundant in SPG	yes	2-DE/MS	Q8K4K8	mitochondrion organisation - antioxidant activity - positive regulation of cell proliferation - negative regulation of apoptotic process - negative regulation of kinase activity
GSTM2, Glutathione S-transferase Mu 2	human	Huang et al. 2008	T, abundant in primary SPC	yes	2-DE/MS	P28161	glutathione metabolic process - xenobiotic metabolic process
PGK2, Phosphoglycerate kinase 2	human	Huang et al. 2008	T, abundant in SPZ	yes	2-DE/MS	P07205	glycolysis - phosphorylation
PRDX4, Peroxiredoxin-4	mouse	Huang et al. 2008	T, abundant in SPZ	yes	2-DE/MS	O08807	oxydation - reduction process
Thymosin β-10	rat	Lagarrigue et al. 2011	SPC, abundant in SPT, eSPT, RBs	yes	MALDI-IMS	P63312	spermatid development - actin cytoskeleton organization - sequestering of actin monomers
Thymosin β-4	rat	Lagarrigue et al. 2011	abundant in SPC, abundant in rSPT; eSPT (head), RBs	yes	MALDI-IMS	P62329	actin cytoskeleton organization - sequestering of actin monomers

(a) T: testis; FT: fetal testis; PGCs: primordial germ cells; maGSCs: multipotent adult germline stem cells; SPG: spermatogonia; SPC: spermatocyte; SPT: spermatid; rSPT: round spermatid; eSPT: elongating spermatid; cSPT: condensed spermatid; RBs: residual body; SPZ: spermatozoa; S: Sertoli cells; GCs: germ cells.
(b) proteomic approaches for protein purification followed by mass spectrometry (MS) identification. 2D-DIGE: two-dimensional fluorescence difference gel electrophoresis; 2-DE: two-dimensional gel electrophoresis; MALDI: matrix associated laser desorption/ionization; TPI column: transition proteins chromatography column.

RIGHTSLINK

models, under *in vivo* conditions [Krüger et al. 2008]. The advances in MALDI imaging described above should also attract the interest of biologists and clinicians, as the benefits of this technology become more widely known.

The studies reviewed here have increased our knowledge on spermatogenesis. Nevertheless, unprecedented efforts are still required to improve our understanding of the highly complex and sophisticated communication networks linking cellular activities during this process. It should now be possible to obtain much clearer snapshots of the molecular mechanisms operating at each step in spermatogenesis, through a relevant use of mature proteomics techniques and exploitation of the data produced through integrative genomics strategies. Interestingly, top-down proteomics and peptidomics, which are emerging should also be considered as valuable new tools for addressing very precise questions relating to normal and pathological spermatogenesis. No doubt that striking biological discoveries in this field will stem from the new breed of comparative Omics studies.

Acknowledgments

The authors wish to thank Dr Frédéric Chalmel for stimulating discussion and help with datamining using the AMEN Software.

Declaration of interest: This work was supported in part by Biogenouest and by Infrastructures en Biologie Santé et Agronomie, Fonds Européen de Développement Régional, and Conseil Régional de Bretagne grants. The authors declare that there are no conflicts of interest.

References

Agashe, V.R. and Hartl, F.U. (2000) Roles of molecular chaperones in cytoplasmic protein folding. Semin Cell Dev Biol **11**:15–25.

Arpanahi, A., Brinkworth, M., Iles, D., Krawetz, S.A., Paradowska, A., Platts, A.E., et al. (2009) Endonuclease-sensitive regions of human spermatozoal chromatin are highly enriched in promoter and CTCF binding sequences. Genome Res **19**:1338–1349.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25**:25–29.

Balhorn, R. (2007) The protamine family of sperm nuclear proteins. Genome Biol **8**:227–227.

Brun, V., Dupuis, A., Adrait, A., Marcellin, M., Thomas, D., Court, M., et al. (2007) Isotope-labeled protein standards: toward absolute quantitative proteomics. Mol Cell Proteomics **6**:2139–2149.

Cagney, G., Park, S., Chung, C., Tong, B., O'Dushlaine, C., Shields, D. C., et al. (2005) Human tissue profiling with multidimensional protein identification technology. J Proteome Res **4**:1757–1767.

Caires, K., Broady, J. and McLean, D. (2010) Maintaining the male germline: regulation of spermatogonial stem cells. J Endocrinol **205**:133–145.

Calvel, P., Kervarrec, C., Lavigne, R., Vallet-Erdtmann, V., Guerrois, M., Rolland, A.D., et al. (2009) CLPH, a novel casein kinase 2-phosphorylated disordered protein, is specifically associated with postmeiotic germ cells in rat spermatogenesis. J Proteome Res **8**:2953–2965.

Calvel, P., Rolland, A.D., Jégou, B. and Pineau, C. (2010) Testicular postgenomics: targeting the regulation of spermatogenesis. Philos Trans R Soc Lond B Biol Sci **365**:1481–1500.

Chalmel, F. and Primig, M. (2008) The Annotation, Mapping, Expression and Network (AMEN) suite of tools for molecular systems biology. BMC Bioinformatics **9**:86–86.

Chalmel, F., Rolland, A.D., Niederhauser-Wiederkehr, C., Chung, S.S. W., Demougin, P., Gattiker, A., et al. (2007) The conserved transcriptome in human and rodent male gametogenesis. Proc Natl Acad Sci USA **104**:8346–8351.

Chu, D.S., Liu, H., Nix, P., Wu, T.F., Ralston, E.J., Yates, J.R., 3rd, et al. (2006) Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. Nature **443**:101–105.

Com, E., Evrard, B., Roepstorff, P., Aubry, F. and Pineau, C. (2003) New Insights into the Rat Spermatogonial Proteome. Mol Cell Proteomics **2**:248–261.

Com, E., Rolland, A.D., Guerrois, M., Aubry, F., Jégou, B., Vallet-Erdtmann, V., et al. (2006) Identification, molecular cloning, and cellular distribution of the rat homolog of minichromosome maintenance protein 7 (MCM7) in the rat testis. Mol Reprod Dev **73**:866–877.

de Godoy, L.M.F., Olsen, J.V., Cox, J., Nielsen, M.L., Hubner, N.C., Fröhlich, F., et al. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature **455**:1251–1254.

de Godoy, L.M.F., Olsen, J.V., de Souza, G.A., Li, G., Mortensen, P. and Mann, M. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol **7**:R50–R50.

De Kretser, D. and Kerr, J. (1988) The cytology of the testis. The physiology of reproduction. Knobil E and Neill JD (eds) Raven Press: New York 837–932.

de Mateo, S., Castillo, J., Estanyol, J.M., Ballescà, J.L. and Oliva, R. (2011) Proteomic characterization of the human sperm nucleus. Proteomics **11**:2714–2726.

de Rooij, D.G. (2001) Proliferation and differentiation of spermatogonial stem cells. Reproduction (Cambridge, England) **121**:347–354.

de Rooij, D.G. and Russell, L.D. (2000) All you wanted to know about spermatogonia but were afraid to ask. J Androl **21**:776–776.

De Santa Barbara, P., Bonneaud, N., Boizet, B., Desclozeaux, M., Moniot, B., Sudbeck, P., et al. (1998) Direct interaction of SRY-related protein SOX9 and steroidogenic factor 1 regulates transcription of the human anti-Müllerian hormone gene. Mol Cell Biol **18**:6653–6665.

Dihazi, H., Dihazi, G.H., Nolte, J., Meyer, S., Jahn, O., Muüller, G.A., et al. (2009) Multipotent Adult Germline Stem Cells and Embryonic Stem Cells: Comparative Proteomic Approach. J. Proteome Res. **8**:5497–5510.

Ewen, K., Baker, M., Wilhelm, D., Aitken, R.J. and Koopman, P. (2009) Global survey of protein expression during gonadal sex determination in mice. Mol Cell Proteomics **8**:2624–2641.

Govin, J., Caron, C., Escoffier, E., Ferro, M., Kuhn, L., Rousseaux, S., et al. (2006) Post-meiotic shifts in HSPA2/HSP70.2 chaperone activity during mouse spermatogenesis. J Biol Chem **281**:37888–37892.

Govin, J., Gaucher, J., Ferro, M., Debernardi, A., Garin, J., Khochbin, S., et al. (2012) Proteomic strategy for the identification of critical actors in reorganization of the post-meiotic male genome. Mol Hum Reprod **18**:1–13.

Guan, K., Nayernia, K., Maier, L.S., Wagner, S., Dressel, R., Lee, J.H., et al. (2006) Pluripotency of spermatogonial stem cells from adult mouse testis. Nature **440**:1199–1203.

Guillaume, E., Dupaix, A., Moertz, E., Courtens, J.L., Jégou, B. and Pineau, C. (2000) Proteome analysis of spermatogonia : identification of a first set of 53 spermatogonial proteins. Proteome DOI 10.1007/s102160000003

Guillaume, E., Evrard, B., Com, E., Moertz, E., Jégou, B. and Pineau, C. (2001a) Proteome analysis of rat spermatogonia: reinvestigation of stathmin spatio-temporal expression within the testis. Mol Reprod Dev **60**:439–445.

Guillaume, E., Pineau, C., Evrard, B., Dupaix, A., Moertz, E., Sanchez, J.C., et al. (2001b) Cellular distribution of translationally controlled tumor protein in rat and human testes. Proteomics **1**:880–889.

Guo, X., Zhang, P., Huo, R., Zhou, Z. and Sha, J. (2008) Analysis of the human testis proteome by mass spectrometry and bioinformatics. Proteomics Clin Appl **2**:1651–1657.

Guo, X., Zhao, C., Wang, F., Zhu, Y., Cui, Y., Zhou, Z., et al. (2010) Investigation of human testis protein heterogeneity using 2-dimensional electrophoresis. J Androl **31**:419–429.

Copyright © 2012 Informa Healthcare USA, Inc.

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

RIGHTSLINK

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T. and Cairns, B.R. (2009) Distinctive Chromatin in Human Sperm Packages Genes for Embryo Development. Nature **460**:473–478.

Han, B.K., Kim, J.N., Shin, J.H., Kim, J.-K., Jo, D.H., Kim, H., et al. (2005) Proteome analysis of chicken embryonic gonads: identification of major proteins from cultured gonadal primordial germ cells. Mol Reprod Dev **72**:521–529.

Hess, R.A. and Renato de Franca, L. (2008) Spermatogenesis and cycle of the seminiferous epithelium. Advances in Experimental Medicine and Biology **636**:1–15.

Hilger, M., Bonaldi, T., Gnad, F. and Mann, M. (2009) Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. Mol Cell Proteomics **8**:1908–1920.

Hossain, A. and Saunders, G.F. (2003) Role of Wilms tumor 1 (WT1) in the transcriptional regulation of the Mullerian-inhibiting substance promoter. Biol Reprod **69**:1808–1814.

Huang, S.-Y., Lin, J.-H., Chen, Y.-H., Chuang, C.-k., Lin, E.-C., Huang, M.-C., et al. (2005) A reference map and identification of porcine testis proteins using 2-DE and MS. Proteomics **5**:4205–4212.

Huang, S.-Y., Lin, J.-H., Teng, S.-H., Sun, H.S., Chen, Y.-H., Chen, H.-H., et al. (2011) Differential expression of porcine testis proteins during postnatal development. Anim Reprod Sci **123**:221–233.

Huang, X.-Y., Guo, X.-J., Shen, J., Wang, Y.-F., Chen, L., Xie, J., et al. (2008) Construction of a proteome profile and functional analysis of the proteins involved in the initiation of mouse spermatogenesis. J Proteome Res **7**:3435–3446.

Huang, X.-Y. and Sha, J.-H. (2011) Proteomics of spermatogenesis: from protein lists to understanding the regulation of male fertility and infertility. Asian J Androl **13**:18–23.

Jégou, B. (1993) The Sertoli-germ cell communication network in mammals. International Review of Cytology **147**:25–96.

Jégou, B., Pineau, C. and Dupaix, A. (1999) Paracrine control of testis function. Male Reproductive Function Wang C. Ed. Endocrine Update Series Kluwer Academic Berlin:41–64.

Krüger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C.A., Forner, F., et al. (2008) SILAC Mouse for Quantitative Proteomics Uncovers Kindlin-3 as an Essential Factor for Red Blood Cell Function. Cell **134**:353–364.

Lagarrigue, M., Becker, M., Lavigne, R., Deininger, S.-O., Walch, A., Aubry, F., et al. (2011) Revisiting rat spermatogenesis with MALDI imaging at 20-microm resolution. Mol Cell Proteomics **10**:M110.005991–M005110.005991.

Lardenois, A., Gattiker, A., Collin, O., Chalmel, F. and Primig, M. (2010) GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. Database **2010**:baq030-baq030-baq030-baq030.

Leblond, C.P. and Clermont, Y. (1952) Definition of the stages of the cycle of the seminiferous epithelium in the rat. Ann N Y Acad Sci **55**:548–573.

Loppion, G., Lavigne, R., Pineau, C., Auvray, P. and Sourdaine, P. (2010) Proteomic analysis of the spermatogonial stem cell compartment in dogfish Scyliorhinus canicula L. Comp Biochem Physiol. Part D, Genomics & Proteomics **5**:157–164.

Maheswaran, S., Englert, C., Zheng, G., Lee, S.B., Wong, J., Harkin, D. P., et al. (1998) Inhibition of cellular proliferation by the Wilms tumor suppressor WT1 requires association with the inducible chaperone Hsp70. Genes Dev **12**:1108–1120.

Marshall, O.J. and Harley, V.R. (2001) Identification of an interaction between SOX9 and HSP70. FEBS Letters **496**:75–80.

Marzec, M., Eletto, D. and Argon, Y. (2011) GRP94: An HSP90-like protein specialized for protein folding and quality control in the endoplasmic reticulum. Biochim Biophys Acta Epub 2011 Nov 3.

Matzuk, M.M. and Lamb, D.J. (2002) Genetic dissection of mammalian fertility pathways. Nat Cell Biol. **4**:s41–49.

Mazzarella, R.A. and Green, M. (1987) ERp99, an abundant, conserved glycoprotein of the endoplasmic reticulum, is homologous to the 90-kDa heat shock protein (hsp90) and the 94-kDa glucose regulated protein (GRP94). J Biol Chem **262**:8875–8883.

Nakagawa, T., Nabeshima, Y.-I. and Yoshida, S. (2007) Functional identification of the actual and potential stem cell compartments in mouse spermatogenesis. Dev Cell **12**:195–206.

Parvinen, M. (1982) Regulation of the seminiferous epithelium. Endocr Rev **3**:404–417.

Paz, M., Morín, M. and Del Mazo, J. (2006) Proteome profile changes during mouse testis development. Comp Biochem Physiol Part D, Genomics & Proteomics **1**:404–415.

Perey, B., Clermont, Y. and Leblond, C.P. (1961) The wave of the seminiferous epithelium in the rat. Am J Anat **108**:47–77.

Primig, M., Wiederkehr, C., Basavaraj, R., Sarrauste de Menthière, C., Hermida, L., Koch, R., et al. (2003) GermOnline, a new cross-species community annotation database on germ-line development and gametogenesis. Nat Genet **35**:291–292.

Raffalli-Mathieu, F., Glisovic, T., Ben-David, Y. and Lang, M.A. (2002) Heterogeneous nuclear ribonucleoprotein A1 and regulation of the xenobiotic-inducible gene Cyp2a5. Mol Pharmacol **61**:795–799.

Rolland, A.D., Evrard, B., Guitton, N., Lavigne, R., Calvel, P., Couvet, M., et al. (2007) Two-dimensional fluorescence difference gel electrophoresis analysis of spermatogenesis in the rat. J Proteome Res **6**:683–697.

Rolland, A.D., Jégou, B. and Pineau, C. (2008) Testicular development and spermatogenesis: harvesting the postgenomics bounty. Adv Exp Med Biol **636**:16–41.

Sato, T., Katagiri, K., Gohbara, A., Inoue, K., Ogonuki, N., Ogura, A., et al. (2011) In vitro production of functional sperm in cultured neonatal mouse testes. Nature **471**:504–507.

Sato, Y., Shinka, T., Chen, G., Yan, H.-T., Sakamoto, K., Ewis, A.A., et al. (2009) Proteomics and transcriptome approaches to investigate the mechanism of human sex determination. Cell Biol Int **33**:839–847.

Son, Y.S., Park, J.H., Kang, Y.K., Park, J.-S., Choi, H.S., Lim, J.Y., et al. (2005) Heat shock 70-kDa protein 8 isoform 1 is expressed on the surface of human embryonic stem cells and downregulated upon differentiation. Stem Cells (Dayton, Ohio) **23**:1502–1513.

Takemori, N. and Yamamoto, M.T. (2009) Proteome mapping of the Drosophila melanogaster male reproductive system. Proteomics **9**:2484–2493.

Tikoo, A., O'Reilly, L., Day, C.L., Verhagen, A.M., Pakusch, M. and Vaux, D.L. (2002) Tissue distribution of Diablo/Smac revealed by monoclonal antibodies. Cell Death Differ **9**:710–716.

Vera, Y., Diaz-Romero, M., Rodriguez, S., Lue, Y., Wang, C., Swerdloff, R.S., et al. (2004) Mitochondria-Dependent Pathway Is Involved in Heat-Induced Male Germ Cell Death: Lessons from Mutant Mice. Biol Reprod **70**:1534–1540.

Viger, R.S., Mertineit, C., Trasler, J.M. and Nemer, M. (1998) Transcription factor GATA-4 is expressed in a sexually dimorphic pattern during mouse gonadal development and is a potent activator of the Müllerian inhibiting substance promoter. Development (Cambridge, England) **125**:2665–2675.

Wilhelm, D., Huang, E., Svingen, T., Stanfield, S., Dinnis, D. and Koopman, P. (2006) Comparative proteomic analysis to study molecular events during gonad development in mice. Genesis **44**:168–176.

Wilhelm, D., Palmer, S. and Koopman, P. (2007) Sex determination and gonadal development in mammals. Physiol Rev **87**:1–28.

Wrobel, G. and Primig, M. (2005) Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. Reproduction **129**:1–7.

Zhu, Y.-F., Cui, Y.-G., Guo, X.-J., Wang, L., Bi, Y., Hu, Y.-Q., et al. (2006) Proteomic analysis of effect of hyperthermia on spermatogenesis in adult male mice. J Proteome Res **5**:2217–2225.

Syst Biol Reprod Med Downloaded from informahealthcare.com by Joaquin Ibanez Esteb on 07/20/12
For personal use only.

RIGHTSLINK

mt Médecine de la Reproduction, Gynécologie Endocrinologie 2012 ; 14 (4) : 272-86

La protéomique, un outil puissant pour comprendre la spermatogenèse normale et pathologique

Proteomics, a powerful technology to help understanding normal and pathological spermatogenesis

Sophie Chocu
Pierre Calvel
Antoine D. Rolland
Charles Pineau

Inserm,
U1085,
Irset,
université de Rennes I,
campus de Beaulieu,
35042 Rennes,
France
<charles.pineau@inserm.fr>

Résumé. La spermatogenèse est un processus très sophistiqué impliqué dans la transmission du patrimoine génétique. Elle implique une succession d'événements cellulaires sophistiqués et pour certains uniques. Chez les mammifères, la spermatogenèse est classiquement divisée en trois phases. Dans la première — ou phase de prolifération mitotique — les cellules germinales primitives ou spermatogonies subissent une série de divisions mitotiques. Dans la deuxième — la phase méiotique — les spermatocytes subissent deux cycles de division consécutifs pour produire des spermatides haploïdes. Dans la troisième — spermiogenèse — les spermatides se différencient en spermatozoïdes. Des mécanismes de régulation paracrine, autocrine, juxtacrine et endocriniens contribuent à la régulation du processus spermatogénétique. L'ensemble des éléments de structure et des facteurs chimiques modulant l'activité des cellules somatiques et germinales est tel que le réseau reliant les différentes activités cellulaires au cours de la spermatogenèse est incroyablement complexe. Au cours des 20 dernières années, les progrès de la génomique ont contribué à améliorer notre connaissance de la spermatogenèse, en identifiant de nombreux gènes essentiels pour le développement des gamètes mâles fonctionnels. Des analyses à grande échelle de la fonction testiculaire ont approfondi notre compréhension de la spermatogenèse normale et pathologique. Les progrès dans le séquençage des génomes et dans la technologie des puces ont été exploités pour des études d'expression de génomes entiers, conduisant à l'identification de centaines de gènes différentiellement exprimés dans le testicule. Cependant, bien que la protéomique soit désormais arrivée à maturité, l'analyse protéomique de la spermatogenèse en est encore à ses balbutiements. Dans ce chapitre, nous passerons en revue l'état de l'art des analyses protéomiques à grande échelle de la spermatogenèse, du développement des cellules germinales au cours de la détermination du sexe, à la spermatogenèse chez l'adulte. Quelques laboratoires ont entrepris l'étude des profils d'expression des protéines et/ou l'analyse systématique de protéomes testiculaires à partir d'organes entiers ou de cellules isolées chez différentes espèces. Nous considérons dans cette revue les avantages et les inconvénients de la protéomique pour étudier le programme d'expression génique des cellules germinales testiculaires. Enfin, nous considérerons l'utilisation de jeux de données de protéines, dans des approches de génomique intégrative (c'est-à-dire combinant génomique, transcriptomique et protéomique), de bioinformatique et de modélisation.

Mots clés : testicule, spermatogenèse, cellule germinale, protéomique, génomique intégrative, régulation traductionnelle

Abstract. Spermatogenesis is a highly sophisticated process involved in transmission of genetic heritage. It includes halving ploidy, repackaging of the chromatin for transport, and equipment of developing spermatids and eventually spermatozoa with the advanced apparatus (e.g. tightly packed mitochondrial sheat in the mid piece, elongating of the tail, reduction of cytoplasmic volume) to elicit motility once they reach the epididymis. Mammalian spermatogenesis is classically divided into three phases. In the first — the proliferative or mitotic phase — primitive germ cells or spermatogonia undergo a series of mitotic divisions. In the second — the meiotic phase — the spermatocytes undergo two consecutive divisions to produce haploid spermatids. In the third — spermiogenesis — spermatids differentiate into spermatozoa. Paracrine, autocrine, juxtacrine and endocrine pathways all contribute to regulation of the process; the array of structural elements and chemical factors modulating somatic and germ cell activity is such that the network linking the various cellular activities during spermatogenesis is unimaginably complex. Over the past two decades, advances in genomics have greatly improved our knowledge of spermatogenesis, by identifying numerous genes essential for the development of functional male gametes. Large-scale analyses of testicular function have deepened our insight into normal and pathological spermatogenesis. Progresses in genome sequencing and microarray technology have been exploited for genome-wide expression studies, leading to the identification of hundreds of genes differentially expressed within the testis. However, although proteomics has now come of age, the proteomics-based investigation of spermatogenesis remains in its infancy. Here, we review the state-of-the-art of large-scale proteomic analyses of spermatogenesis, from germ cell development during sex determination to spermatogenesis in the adult. Indeed, a few laboratories have undertaken differential protein profiling expression studies and/or systematic analyses of testicular proteomes in entire organs or isolated cells from various species. We consider the pros and cons of proteomics for studying the testicular germ cell gene expression program. Finally, we address the use of protein datasets, through integrative genomics (*i.e.*, combining genomics, transcriptomics and proteomics), bioinformatics and modelling.

Key words: testis, spermatogenesis, germ cells, proteomics, integrative genomics, translational regulation

médecine thérapeutique
mt Médecine de la Reproduction Gynécologie Endocrinologie

Tirés à part : C. Pineau

doi:10.1684/mte.2012.0430

Pour citer cet article : Chocu S, Calvel P, Rolland AD, Pineau C. La protéomique, un outil puissant pour comprendre la spermatogenèse normale et pathologique. *mt Médecine de la Reproduction, Gynécologie Endocrinologie* 2012 ; 14 (4) : 272-86 doi:10.1684/mte.2012.0430

© John Libbey Eurotext, 2012

Tiré à part auteur

Chez les mammifères, les tubes séminifères sont le siège de la spermatogenèse qui est classiquement divisée en trois phases. Dans la première phase proliférative (ou phase mitotique), les cellules germinales primitives (spermatogonies) subissent une série de divisions mitotiques. Dans la deuxième phase (ou phase méiotique), les spermatocytes subissent deux cycles consécutifs pour produire les spermatides haploïdes. Dans la troisième phase (spermiogenèse), les spermatides se différencient en spermatozoïdes. L'une des caractéristiques intéressantes de la spermatogenèse est que les cellules germinales en développement forment des associations de composition fixe, ou étapes, constituant le cycle de l'épithélium séminifère. L'organisation mais également l'intégrité de l'épithélium séminifère sont assurées par les cellules somatiques de Sertoli. Chez le rat [1], le cycle de l'épithélium séminifère est divisé en 14 stades (1 à 14) et la spermiogenèse, définie comme la transformation morphologique des spermatides en spermatozoïdes, est quant à elle subdivisée en 19 étapes de différenciation (de 1 à 19). Cette dernière étape du développement des cellules germinales fournit un exemple frappant et unique de la différenciation cellulaire impliquant tout à la fois la formation de l'acrosome, une condensation nucléaire et la biogenèse du flagelle. Une seconde caractéristique de la spermatogenèse est l'ordonnancement d'associations distinctes des cellules le long des tubes séminifères (segments), souvent désigné comme la « vague de l'épithélium séminifère » [2]. Une vague englobe l'ensemble des 14 segments chez le rat, de 12 chez la souris et de six chez l'homme, chacun composé d'associations cellulaires différentes [3]. Un segment est défini comme une portion longitudinale du tube séminifère correspondant à une association unique de cellules, et appelé stade [4, 5].

Ce processus de différenciation unique implique des facteurs de régulation autocrine, juxtacrine, paracrine et endocrine. Il est conditionné par l'activation successive et/ou la répression de milliers de gènes et de protéines, parmi lesquelles de nombreuses isoformes testicule-spécifiques. Toutes ces caractéristiques font du testicule l'un des tissus les plus complexes. Démêler cette complexité par des approches de culture cellulaire s'est avéré problématique en raison de la difficulté à cultiver les cellules les plus hautement différenciées de la lignée germinale mâle. De ce fait, l'absence d'un « terrain d'essai » naturel a rendu les analyses génomiques particulièrement importantes comme prélude aux tests d'hypothèses *in vivo*. Les progrès de la biologie moléculaire et de la génomique ont permis d'améliorer notre connaissance de la spermatogenèse, en permettant l'identification d'un grand nombre de gènes essentiels pour le développement des gamètes mâles fonctionnels [6, 7]. En effet, des progrès significatifs ont été accomplis dans l'analyse à grande échelle de la fonction testiculaire, permettant de fait d'approfondir notre connaissance de la spermatogenèse normale et pathologique. Plusieurs laboratoires ont capitalisé sur de rapides progrès dans le séquençage du génome et le développement de puces à ADN, menant des études d'expression du génome entier qui ont conduit à l'identification de centaines de gènes spatialement et temporellement régulés au cours de l'ontogenèse du testicule [8]. Cependant, bien que la protéomique soit parvenue à maturité, l'étude de la spermatogenèse par des stratégies de protéomique en est encore à ses balbutiements.

Dans cette revue, nous présenterons l'état de l'art des analyses protéomiques à grande échelle de la spermatogenèse, depuis le développement des cellules germinales au cours de la détermination du sexe, jusqu'à la spermatogenèse chez l'adulte. Nous examinerons les avantages et les inconvénients de la protéomique pour des études ciblant le programme d'expression des gènes des cellules germinales au sein du testicule. Un résumé des études réalisées à ce jour dans le domaine est présenté dans la *figure 1*. Enfin, nous discuterons de l'utilisation de jeux de données protéomiques *via* des approches de génomique intégrative (à savoir génomique, transcriptomique et protéomique), de bioinformatique et de modélisation.

Le profil d'expression de cellules germinales embryonnaires

Chez les mammifères, les gonades embryonnaires proviennent de structures indifférenciées et bipotentielles dans le mésoderme intermédiaire. L'expression transitoire d'un gène unique, *sex-determining region*, chromosome Y (Sry), dans les cellules somatiques de soutien des crêtes génitales, déclenche la différenciation sexuelle des gonades en testicules. Sous l'influence des cellules pré-Sertoli, les cellules germinales primordiales (PGC) se développent en prospermatogonies, qui sont les cellules précurseurs de la lignée germinale mâle [9]. Les études basées sur des approches protéomiques ont été moins utilisées que les études de puces pour étudier la détermination sexuelle de cellules somatiques dans les gonades embryonnaires, mais ont néanmoins été fructueuses [10-12]. En combinant électrophorèse bidimensionnelle sur gel (2DE) et spectrométrie de masse (MS), Wilhelm *et al.* [12] ont comparé les profils d'expression protéique des testicules et des ovaires fœtaux de souris. L'expression de trois protéines, la protéine *heterogeneous nuclear ribo Spermatogenesis nucleoprotein* A1 (hnRPA1), la protéine *polymorphic tumor rejection antigen* (TRA1) et la protéine *heat shock cognate 71kDa protein* (HSC71), s'est révélée augmentée dans les gonades mâles par rapport aux gonades femelles. L'expression spécifiquement mâle du gène *TRA1* ainsi que les niveaux d'expression plus élevés des gènes *Hsc71* et *hnRpa1* au cours du développement des gonades ont été confirmés par PCR quantitative en temps réel. En outre, HSC71 a été trouvée phosphory-

© John Libbey Eurotext, 2012

Tiré à part auteur

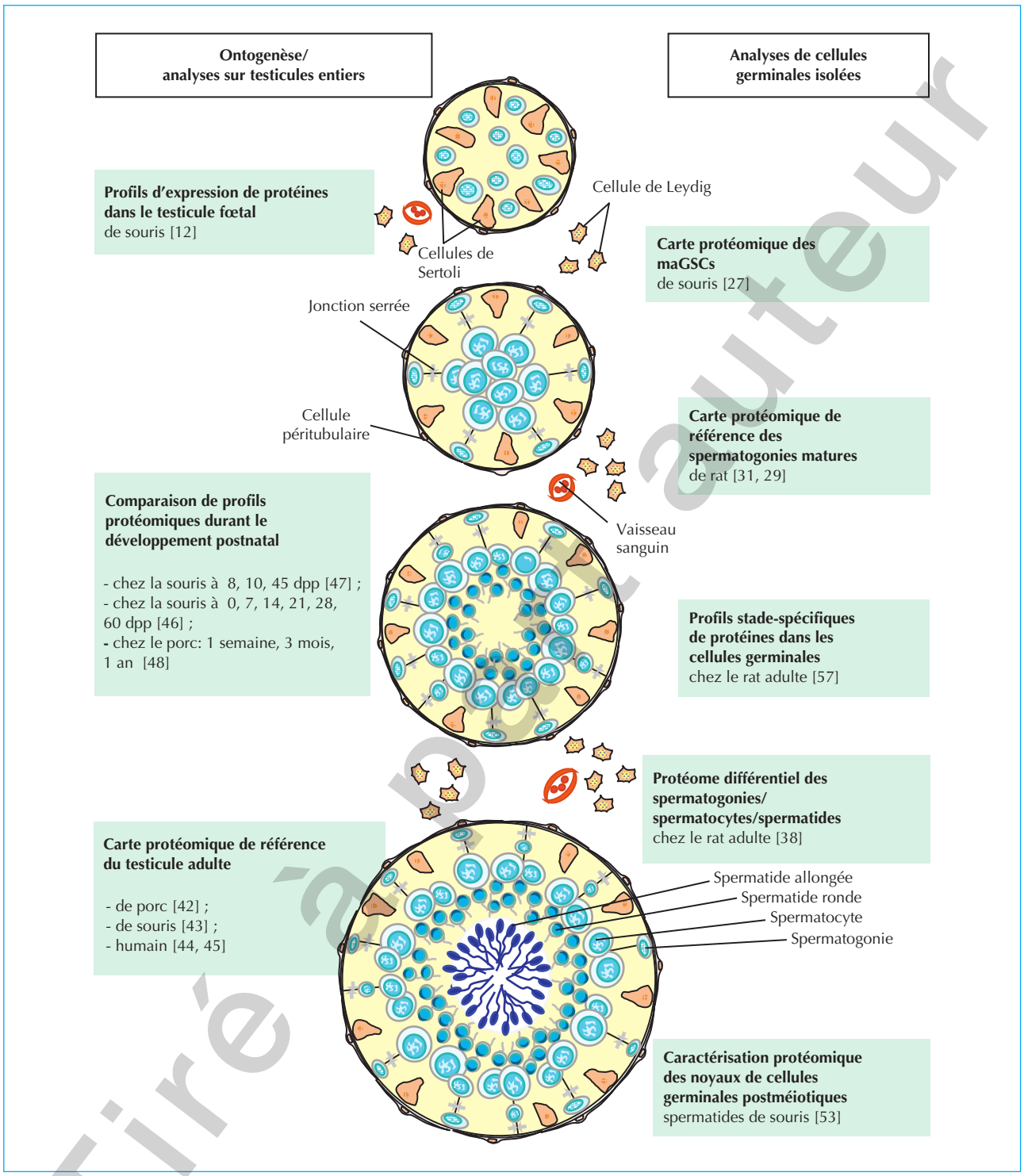

Figure 1. Étude de la spermatogenèse chez les mammifères par des études protéomiques. Les sections d'un tube séminifère à des stades de développement différents sont représentées ici. Les principales études protéomiques, qui ont contribué à une meilleure connaissance de la spermatogenèse sont mentionnées. La spermatogenèse chez les mammifères a lieu dans les tubes séminifères et est divisée en trois phases. Dans la première phase (proliférative ou mitotique), les cellules germinales primitives (spermatogonies) subissent une série de divisions mitotiques. Pendant la deuxième phase (méiose), les spermatocytes subissent deux cycles consécutifs pour produire les spermatides haploïdes. Pendant la troisième phase (spermiogenèse), les spermatides se différencient en spermatozoïdes.
maGSCs : cellules germinales souches multipotentes adultes.

Tiré à part auteur

© John Libbey Eurotext, 2012

lée de manière plus importante dans les gonades mâles que dans les gonades femelles, soulignant l'importance des approches protéomiques pour la détection des modifications post-traductionnelles. Jusqu'à présent, ces trois protéines ne sont pas connues pour être associées avec le développement des gonades ou la différenciation sexuelle, et leur rôle dans ces événements reste à définir. HnRPA1 appartient à la classe de ribonucléoprotéines hétérogènes (hnRNP) qui se lient aux ARNs nucléaires hétérogènes (hnRNAs). Elle module l'expression de gènes spécifiques par liaison à l'ARN grâce à trois domaines de liaison à l'ARN, altérant alors leur stabilité. Fait intéressant, le gène codant pour la testostérone-15-alpha-hydroxylase (*Cyp2A5*) a été identifié comme une cible de hnRPA1 [13]. Alors que l'enzyme Cyp2A5 est exprimée dans la gonade embryonnaire mâle, son rôle dans ce contexte est inconnu. TRA1, également nommée Hsp90b1 ou *glucose-regulated protein 94* (Grp94) est une glycoprotéine du réticulum endoplasmique [14]. Grp94, présente dans la lumière du RE, est une protéine chaperonne des protéines sécrétées et membranaires et impacte leur repliement. Cette protéine possède également des propriétés particulières telles que la capacité de liaison au calcium, requise par les conditions intraréticulum endoplasmique [15]. La protéine HSC71 appartient à la famille des protéines HSC70. Les membres de cette famille de protéines peuvent protéger les domaines hydrophobes des protéines cytoplasmiques et fonctionnent comme des chaperonnes pour permettre un repliement efficace [16]. Wilhelm *et al*. ont trouvé également que HSC71, plus abondamment exprimée dans les gonades mâles que dans les gonades femelles, est phosphorylée après la traduction [12]. Pourquoi la phosphorylation de HSC71 serait-elle importante pour la différenciation sexuelle ? Il a été montré que les protéines HSC70 et HSC71 interagissent avec SOX9 dans les lignées cellulaires testiculaires [17]. Or, il est bien connu que l'hormone anti-mullérienne (AMH) qui entraîne la régression des canaux de Müller chez le mâle, est une cible de SOX9 au cours de la différenciation testiculaire. L'activité de transcription au promoteur Amh est augmentée par la liaison de ce dernier aux facteurs de transcription SF1, GATA4 et WT1 [18-20]. En outre, la protéine HSC70 a été montrée comme étant associée à WT1 [21]. Selon les auteurs, il est possible que les protéines étroitement apparentées Hsc70 et HSC71 permettent de stabiliser la formation d'un complexe protéique SOX9-SF1-WT1 essentiel pour la différenciation testiculaire. Dans un tel contexte, il apparaît que la phosphorylation de HSC71, spécifique chez le mâle, pourrait être un mécanisme de régulation post-traductionnelle important.

Très peu de choses sont connues actuellement sur la biologie des PGC en raison du faible nombre de ces cellules chez l'embryon. Une analyse protéomique sur une culture de PGC de poulet a permis d'identifier 50 protéines, y compris des facteurs de croissance et des protéines régulées au cours du développement [22]. La plupart des protéines exprimées dans les PGC de poulet telles que la vimentine, la béta-actine, les isoformes de la tropomyosine alpha-1, tubuline béta-1, desmine, HSP70, ou les précurseurs de la cathepsine B, ont des homologues chez les mammifères. Toutefois, selon la base de données UniProt, le fragment CEphA6 et l'ovotransferrine identifiés dans les PGC n'ont pas d'homologue chez les mammifères. Il est à noter qu'aucune étude protéomique sur les PGC n'a été réalisée chez les mammifères, bien que les analyses d'expression du génome entier aient fourni des informations conséquentes quant aux programmes génétiques qui contrôlent le développement de ces cellules [23, 24].

La niche des cellules souches spermatogoniales

Les cellules souches spermatogoniales (*spermatogonial stem cells* ou SSCs) provenant des PGCs font partie d'un sous-ensemble de cellules germinales mâles appelées spermatogonies indifférenciées, qui incluent les spermatogonies A *single* (As) et leurs cellules filles A *paired* (Ap) et A *aligned* (Aa) [6]. La spermatogenèse normale est maintenue par la capacité des SSC à alimenter en continu les tubules séminifères des testicules adultes, en spermatogonies différenciées, sources de toutes les cellules germinales tout au long de la vie du mâle. Il semble qu'une sous-population seulement des spermatogonies ait une réelle activité de cellule souche [25]. Ces auteurs ont en effet estimé que chez la souris, environ 10 % seulement des spermatogonies As sont des vraies SSCs. Il existe en fait deux populations de SSCs dans les testicules de souris. La première est une population de « cellules souches réelles (stables) ». Les cellules souches réelles ont la capacité de s'auto-renouveler et le font vraiment. Une seconde population de « cellules souches potentielles » existe aussi, ces cellules sont capables de s'auto-renouveler mais cela ne se produit pas dans la situation normale. Une rotation rapide des cellules souches potentielles a lieu dans les testicules normaux suggérant l'appartenance de ces cellules au transit d'amplification plutôt qu'à la population en sommeil, pour fournir les spermatogonies différenciées [25]. Pour maintenir leur capacité d'auto-renouvellement, les SSC ont besoin de résider dans un micro-environnement ou niche qui fournit tous les facteurs et les interactions cruciales pour leur survie et leur développement. Ainsi, le microenvironnement de la SSC dans les tubules séminifères semble être critique pour la production de la descendance des spermatogonies [26]. Malgré le rôle fondamental que jouent les spermatogonies dans la spermatogenèse, peu de choses sont connues quant aux mécanismes moléculaires qui sous-tendent le maintien de leurs diverses fonctions. Des technologies de

© John Libbey Eurotext, 2012

Tirés à part auteur

pointe permettant des analyses globales de l'expression des gènes et des profils protéiques ont permis aux scientifiques de mieux comprendre la signature moléculaire des SSCs et de leur micro-environnement, la niche des SSCs.

Avec seulement un pourcentage limité de spermatogonies As étant de vraies SSCs, comme indiqué précédemment, il semble assez difficile d'isoler et d'extraire suffisamment de SSCs et de protéines en quantité suffisante pour réaliser des études protéomiques et accéder en profondeur à leur protéome. Compte tenu de ces limitations techniques, Dihazi *et al*. s'appuient sur le fait que sous des conditions de culture spécifiques, les SSCs acquièrent leur pluripotence, car elles peuvent se différencier en cellules somatiques des trois feuillets embryonnaires [27]. C'est la raison pour laquelle dans un travail précurseur, Guan *et al*. ont nommé ces cellules multipotentes *multipotent adult germline stem cells* (maGSCs) [28]. Dihazi *et al*. ont alors utilisé une approche 2-DIGE pour comparer les protéomes de référence des lignées maGSCs et ESCs de souris et identifier par spectrométrie de masse des marqueurs spécifiques de ces deux lignées [27]. Ainsi 409 protéines ont-elles pu être identifiés, dont 166 non redondantes sont associées aux cellules souches maGSCs et ESCs (*embryonic stem cells* ou cellules souches embryonnaires). Parmi celles-ci, seulement un petit sous-ensemble de 18 protéines se révèle différentiellement exprimé entre les ESCs et maGSCs, ce qui indique que ces deux types de cellules peuvent avoir des protéomes assez similaires. Parmi ces protéines, le facteur d'initiation de traduction eucaryote 5A-1, la galectine-1, la lactoylglutathione lyase, impliquées dans la différenciation cellulaire, ont un niveau d'expression plus élevé dans les maGSCs que dans les ESCs. Fait intéressant, 27 des 166 protéines identifiées dans les maGSCs ont été montrées comme étant associées aux spermatogonies chez le rat [29]. Enfin, Dihazi *et al*. ont également constaté que la protéine HSPA8 de 71kDa, apparentée aux protéines du choc thermique, était fortement exprimée et présente sous différentes formes dans les maGSCs et les ESCs [27]. La spectrométrie de masse a en outre permis d'identifier HSPA8 sur la surface des ESCs humaines où son expression semblerait être sous-régulée pendant la différenciation, ce qui suggère que HSPA8 puisse être un marqueur potentiel pour les ESC indifférenciées humaines [30].

Les analyses protéomiques, bien que moins utilisées que les approches transcriptomiques pour l'investigation de la niche des SSC, ont contribué de façon majeure à notre connaissance des spermatogonies en développement. Notre laboratoire a ainsi mis en place la première carte du protéome de référence des spermatogonies matures, fraîchement isolées à partir de testicules de rats de neuf jours, et identifié un premier ensemble de 53 protéines cytosoliques solubles [31]. Nous avons comparé ce modeste jeu de protéines à l'ensemble des données du transcriptome de la spermatogenèse chez le rat réalisé par Chalmel *et al*. [32] à l'aide de la suite d'outils logiciels AMEN [33]. Cinquante et un des 53 identifiants de gènes (GeneID) correspondant aux protéines identifiées, sont exprimés dans les spermatogonies de rat, et parmi eux, cinq GeneID sont mitotiques et différentiellement exprimés dans le testicule. Cela montre que Guillaume *et al*. ont identifié dans ce travail précurseur une majorité de protéines majoritaires et ubiquitistes et seulement cinq protéines spécifiques des spermatogonies, non exprimées dans les autres types de cellules germinales (à savoir la tropomyosine 4, la carbonyl réductase 1, la protéine ribosomique SA, la spermidine synthase et l'isocitrate déhydrogénase 2 (NADP+) mitochondriale). Pourtant, les protéines les plus intéressantes susceptibles d'être exprimées par la lignée germinale sont les protéines dites « à faible nombre de copies » qui correspondent à des protéines cellulaires discrètes, généralement sous-régulées (par exemple, des facteurs de transcription) et donc potentiellement cruciales pour la différenciation des cellules germinales. Pour approfondir l'analyse du protéome des spermatogonies et détecter les protéines discrètes, des gels à gamme étroite de pH chevauchants ont alors été utilisés pour fractionner des extraits cellulaires [29]. Ces gels « zoom » améliorent grandement la séparation des protéines, permettant une résolution beaucoup plus élevée que les gels classiques à large gamme de pH. Certaines protéines minoritaires qui ont pu être identifiées sont impliquées dans les processus de régulation (défense cellulaire, détoxification, régulation de l'expression des gènes, transduction du signal, fixation du calcium, trafic et réplication de l'ADN). Ces deux études ont fourni, à l'époque, le premier répertoire important de protéines exprimées par les spermatogonies de rat.

Nous avons ensuite étudié plus en détail le rôle de certaines de ces protéines dans la spermatogenèse, en raison de leur possible implication dans les événements de prolifération et de différenciation [34-36]. Le profil d'expression spatio-temporel des protéines et la production d'ARN messagers dans les testicules de rat ont été étudiés pour la stathmine, une protéine connue pour être associée à la dynamique des microtubules [35]. Cette protéine forme des agrégats dans le cytoplasme des spermatides matures avant son élimination au moment de la spermiation. Sa production dans la lignée germinale peut donc refléter la réorganisation de la structure cellulaire dans les cellules germinales au cours de la spermatogenèse. Une approche similaire a été utilisée pour analyser la protéine *translationally controlled tumor protein* (TCTP) dans les testicules de l'homme adulte, et dans les testicules de rats nouveau-nés ou adultes [36]. La protéine TCTP est produite, en quantités diverses, dans toutes les populations de cellules testiculaires isolées, mais son abondance dans les spermatogonies suggère qu'elle puisse jouer un rôle

© John Libbey Eurotext, 2012

Tiré à part auteur

important dans la spermatogenèse. Com *et al*. [34] ont analysé la protéine *minichromosome maintenance protein* 7 (MCM7) préalablement identifiée dans les spermatogonies de rat. Ils ont décrit la distribution spatiale de la protéine MCM7 et de son ARNm dans les cellules testiculaires chez le rat, ainsi que leur expression temporelle au cours du développement testiculaire. Ces auteurs ont rapporté une divergence d'expression entre la distribution des ARNm et de la protéine. Cependant, aucune conclusion n'a été tirée concernant le rôle de MCM7 dans les spermatogonies ou la lignée germinale. Plus récemment, une étude différentielle sur gel 2-D/MS a été réalisée sur les testicules de la roussette mature *Scyliorhinus L. canicula*, dans le but d'identifier des protéines spécifiques du compartiment des cellules souches spermatogoniales [37]. Les auteurs ont comparé les profils protéiques de la zone germinative avec ceux de la zone des spermatocytes. Des séquences *de novo* ont été obtenues pour 33 des 169 protéines sélectionnées pour l'identification par spectrométrie de masse, mais seulement 16 protéines ont été identifiées. L'une de ces protéines, la sous-unité alpha-6 du protéasome, a été spécifiquement mise en évidence dans la zone germinative des testicules qui contient des spermatogonies isolées de grande taille. Plusieurs autres protéines identifiées (par exemple, la stathmine) avaient déjà été mises en évidence dans de précédentes analyses protéomiques de spermatogonies et de cellules germinales chez les rongeurs [29, 31, 38]. Bien que préliminaire, cette étude a démontré l'utilité de la Roussette commune comme un modèle pertinent pour l'analyse protéomique de la niche des cellules souches spermatogoniales. Des études de ce type devraient faire la lumière sur les réseaux protéiques complexes qui régissent la biologie des spermatogonies, à condition que soient disponibles des données génomiques pour les élasmobranches.

Le développement postnatal des cellules germinales mâles

La production de spermatozoïdes *via* la spermatogenèse reste le phénomène le plus étudié dans le domaine de la reproduction mâle, principalement par le biais d'expériences à grande échelle. Au cours des dix dernières années, la génomique et la postgénomique ont été utilisées avec succès pour l'identification de nombreux gènes et protéines essentiels pour le développement de gamètes mâles fonctionnels [23]. L'utilisation de la transcriptomique et de la protéomique pour l'étude de la spermatogenèse est parfaitement justifiée, car la production de spermatozoïdes est assurée par une succession d'événements complexes et étroitement régulés. Ces techniques permettent de générer des clichés très précis des réseaux moléculaires impliqués séquentiellement pendant la spermatogenèse.

Les stratégies adoptées initialement étaient basées sur la caractérisation systématique des protéines présentes soit dans les cellules germinales isolées à un instant donné du développement, soit dans le testicule entier. Un ensemble de 132 protéines spermatogéniques abondantes a débouché de l'identification systématique des protéines associées à la chromatine dans les cellules germinales chez le nématode *Caenorhabditis elegans* [39]. L'analyse fonctionnelle de ces protéines a conduit à l'identification de protéines conservées spécifiques de la spermatogenèse, cruciales pour la compaction de l'ADN, la ségrégation des chromosomes et la fécondité. En analysant les protéines acido-solubles présentes dans les spermatides de souris dont la chromatine est en cours de condensation, Govin *et al*. ont identifié HSPA2, la première protéine chaperonne des protéines de transition [40]. HSPA2 contrôle la transition « histone-protéine de transition » contribuant ainsi à la réorganisation entière du génome spécifique des spermatides. Une approche gel 2D/MS a été utilisée afin d'établir plusieurs cartes de référence du protéome du testicule chez la drosophile [41], le porc [42], la souris [43] et l'humain [44]. Trente-neuf protéines spécifiques du testicule et potentiellement importantes pour la fonction testiculaire ont été identifiées par fractionnement SDS-PAGE combiné à une RP-LC-MS/MS et à l'analyse bioinformatique du protéome du testicule humain [45]. Guo *et al*. ont par la suite mis en évidence des différences dans la masse et le point isoélectrique des protéines produites dans le testicule humain, en raison d'évènements d'épissage alternatif et de différentes formes de modifications post-traductionnelles, la plupart étant des phosphorylations [44].

Ces études ont généré de grands jeux de données de protéines considérées comme importantes pour la fonction testiculaire, mais aucune information quant à la régulation de la spermatogenèse. Avec l'objectif de mieux appréhender ce processus, différentes équipes ont depuis utilisé des stratégies plus sophistiquées, pour la plupart fondées sur des analyses d'expression différentielle au cours de la spermatogenèse. Des études comparatives des profils d'expression protéique dans les testicules de souris à des moments spécifiques de la première vague de la spermatogenèse ont été mises en place par plusieurs groupes [46, 47]. Paz *et al*. ont comparé les profils d'expression des protéines solubles présentes dans les testicules de souris à 8, 18, et 45 jours *postpartum* (DPP). Ils ont identifié 44 protéines ou formes variantes présentant une expression différentielle au cours du développement [47]. De la même façon, Huang *et al*. ont comparé les protéomes des testicules de souris à 0, 7, 14, 21, 28 et 60 dpp. Ils ont identifié 257 protéines présentant une expression différentielle, et potentiellement impliquées dans l'initiation de la spermatogenèse. Ces protéines incluent AOP1A et GSTM2, qui semblent être sous-exprimées, ainsi que PGK2 et PRDX4, qui semblent être sur-exprimées au cours du dévelop-

© John Libbey Eurotext, 2012

Tiré à part auteur

pement testiculaire [46]. L'expression différentielle des protéines au cours du développement postnatal a également été étudiée dans les testicules de porc [48]. Huang *et al*. ont étudié les testicules de quatre porcs aux âges d'une semaine, trois mois et un an. Ils ont constaté que 108 protéines étaient différentiellement exprimées ; 90 d'entre elles ont été identifiées par spectrométrie de masse et triées en fonction de leur différence d'abondance à ces stades de développement.

Une autre stratégie couramment utilisée consiste à comparer les niveaux d'expression de protéines particulières entre les différents types de cellules germinales purifiées. Il y a deux raisons principales pour le choix d'une telle approche plutôt que l'utilisation d'échantillons de testicules entiers à différents stades de la première vague spermatogénétique. Tout d'abord, l'utilisation de populations isolées de cellules germinales et non d'un organe entier est plus efficace pour l'identification de protéines à faible nombre de copies par cellule. Par ailleurs, cette approche permet d'établir l'origine cellulaire des protéines identifiées, mais il faut garder à l'esprit que les populations isolées des cellules germinales ne sont jamais pures à 100 % et qu'une protéine spécifique peut être exprimée par l'un des types cellulaires contaminants.

Notre groupe a réalisé une analyse différentielle du protéome de la spermatogenèse chez le rat reposant sur la technologie 2D-DIGE [38]. Cette étude a été la première à décrire l'utilisation de la 2D-DIGE pour identifier un grand nombre de protéines avec des abondances relatives différant considérablement entre les spermatogonies, les spermatocytes au stade pachytène et les spermatides postméiotiques chez le rat. Des extraits bruts de protéines cytosoliques de spermatogonies, de spermatocytes et de spermatides ont d'abord été analysés, avec un succès limité en termes de protéines différentiellement exprimées (35 protéines). Les protéomes des cellules germinales ont été ensuite examinés plus en détail après fractionnement chromatographique des extraits cytosoliques en quatre sous-protéomes préalablement à l'analyse 2D-DIGE. L'ensemble des analyses réalisées a permis d'identifier pas moins de 977 spots protéiques présentant une expression différentielle, dont la plupart correspondaient à des protéines individuelles. En se basant sur l'hypothèse selon laquelle les protéines ayant un profil d'expression hautement différentiel sont susceptibles de jouer un rôle clé dans la différenciation, nous nous sommes concentrés sur les spots protéiques différentiels avec des ratios moyens d'au moins 2,5 entre deux types cellulaires. Cette approche a conduit à l'identification de 123 protéines uniques aux profils différentiels (et reproductibles) d'expression au cours de la spermatogenèse [38].

Parmi les protéines identifiées dans cette étude, celles qui étaient identifiées pour la première fois dans la lignée germinale mâle ont ensuite été caractérisées par des études ciblées. Ainsi, la caséine-phosphoprotéine (CLPH) spécifique des spermatides a été démontrée comme étant une protéine désordonnée et ayant la capacité de lier le calcium. La CLPH est phosphorylée par la caséine kinase 2, qui joue un rôle majeur dans le développement de la tête des spermatozoïdes [49]. D'autres protéines identifiées par Rolland *et al*. [38] impliquées dans divers processus biologiques susceptibles de jouer un rôle clé dans la spermatogenèse pourraient faire l'objet d'une étude approfondie. Par exemple, la protéine pro-apoptotique Smac/Diablo qui est présente en grande quantité dans les spermatocytes et les spermatides de souris [50]. Son mode d'expression est similaire dans les testicules de souris et de rat [38, 50], ce qui suggère une conservation fonctionnelle et son implication dans la spermatogenèse [51]. Dans une étude récente, Huang et Sha [52] ont présenté l'utilisation d'une stratégie de protéomique *shotgun* pour étudier les profils d'expression de protéines spécifiques des cellules germinales haploïdes et tétraploïdes, purifiées par cytométrie en flux à partir de testicules de souris adultes. Plus que de construire les protéomes à grande échelle des cellules germinales haploïdes et tétraploïdes, ces auteurs visent à caractériser plus en détail les protéines impliquées dans des événements cellulaires spécifiques. Ainsi, pas moins de 3 507 protéines ont été identifiées dans les cellules germinales tétraploïdes et il s'avère que 216 de ces protéines ont des homologues chez la levure connus pour être impliqués dans la méiose. Ce type d'approche est généralement puissant et devrait fournir de précieuses informations pour améliorer notre compréhension des mécanismes qui sous-tendent la spermatogenèse.

Très récemment, une stratégie de pointe en protéomique a été utilisée pour étudier le protéome nucléaire de cellules germinales mâles postméiotiques, à savoir les spermatides allongées et condensées [53]. Pendant l'allongement des spermatides et la condensation du génome haploïde, les nucléosomes sont désassemblés. La plupart des histones sont successivement remplacées par des protéines de transition (TPs), elles-mêmes remplacées ensuite par les protamines (Prms) qui sont très basiques [54]. Fait intéressant, 5 à 15 % de la chromatine des spermatozoïdes reste associée aux nucléosomes, en particulier à des loci importants pour le développement, mettant ainsi en évidence le rôle épigénétique de cette distribution différentielle [55, 56]. Afin d'identifier et de caractériser des acteurs essentiels dans la compaction postméiotique du génome mâle et la mise en place de son épigénome associé chez la souris, Govin *et al*. [53] ont utilisé une stratégie basée sur l'identification de protéines obtenues à partir de cellules germinales mâles postméiotiques à deux stades différents. D'une part, les auteurs ont analysé les extraits nucléaires constitués de protéines basiques acido-solubles, y compris les protéines d'emballage de l'ADN telles que les histones, les TPs et les Prms [40], ainsi que des protéines non encore

© John Libbey Eurotext, 2012

Tiré à part auteur

caractérisées. D'autre part, ils ont recherché les protéines acides agissant potentiellement comme des chaperonnes pour les protéines basiques interagissant avec l'ADN. Pour identifier les protéines acides, les auteurs ont développé une approche basée sur l'interaction avec les protéines de transition (*transition proteins*/TPs) sur le principe de la chromatographie. Cette approche a consisté à déposer des extraits nucléaires de spermatides rondes, de spermatides allongées et de spermatides condensées sur une colonne contenant des billes de CNBr activées revêtues de TPs. Les protéines retenues ont ensuite été éluées et analysées par SDS-PAGE et spectrométrie de masse. Ce travail a conduit à l'identification de 70 protéines, dont 46 dans l'extrait acido-soluble des spermatides allongées condensées, et 29 par chromatographie TP. Cinq protéines ont été identifiées à la fois dans les extraits acido-solubles et par chromatographie TP. La plupart des protéines identifiées dans l'extrait acido-soluble appartenaient à la famille des histones ou la famille des petites protéines non basiques, et non histones. La plupart des protéines représentées dans l'éluat de la chromatographie TP étaient soit des chaperonnes, soit des facteurs liés au stress. À titre d'exemple, HSPA2, une protéine de choc thermique spécifique des testicules, a été trouvée à la fois dans l'éluat de la chromatographie TP et dans l'extrait acido-soluble, ce qui suggère qu'HSPA2 puisse agir comme une chaperonne pour les TPs, indépendamment de sa charge.

Govin *et al*. ont finalement comparé leurs données protéomiques avec les jeux de données transcriptomiques d'expression des gènes dans les tissus normaux de souris ou dans les cellules germinales mâles disponibles dans *gene expression omnibus repository* (GEO). Ils ont réalisé des analyses fonctionnelles en utilisant la GeneOntology (GO). Une liste des facteurs impliqués dans le compactage postméiotique et la programmation du génome mâle a pu être proposée [53]. Fait intéressant, les principaux constituants du protéome nucléaire des cellules germinales mâles postméiotiques sont des protéines liant l'ADN (impliquées dans la transmission de l'information épigénétique à l'embryon) et des facteurs chaperons pouvant interagir avec ces protéines (impliqués dans l'incorporation de certaines protéines de liaison à l'ADN à des régions spécifiques du génome mâle).

Les développements récents des études protéomiques de la spermatogenèse

L'utilisation de sections de testicules totaux peut être utile pour étudier la spermatogenèse. En effet, le suivi de l'expression de protéines germinales dans les tubules séminifères est récemment devenu possible grâce à l'imagerie par spectrométrie de masse *matrix-assisted laser desorption/ionization* (MALDI). Cette nouvelle technologie est désormais reconnue comme un outil puissant pour la détection et l'identification de protéines *in situ*, sur des coupes minces de tissus, sans effet sur la distribution native ou la fonction des protéines concernées [57]. L'imagerie par spectrométrie de masse MALDI a été utilisée avec succès pour étudier les profils protéiques au sein des tubules séminifères chez le rat adulte [57].

Chez les mammifères, le testicule est de tous les organes celui qui présente l'anatomie la plus complexe. C'est pour cette raison qu'il a été choisi par ces auteurs comme un modèle pertinent pour valider et optimiser leurs développements technologiques. Lagarrigue *et al*. ont pu visualiser, avec une résolution latérale de 20 µm, différents stades de développement des cellules germinales dans les tubules séminifères et corréler les images moléculaires obtenues aux stades parfaitement établis de la classification de Leblond et Clermont. Les protéines d'intérêt ont été identifiées par spectrométrie de masse dite *top-down*, et comprennent les thymosine β-10 et β-4. Les auteurs ont pu également superposer les images moléculaires avec celles d'immunohistochimie et confirmé que ces deux protéines sont produites dans les cellules germinales de manière stade-dépendante. La thymosine β-10 était déjà connue pour être impliquée dans le développement des spermatides, mais ce travail constitue la première preuve de la production de thymosine β-4 dans le testicule. Sur la base de ces résultats prometteurs, nous prévoyons que l'imagerie par spectrométrie de masse MALDI devrait impacter de façon importante l'étude de la spermatogenèse et, plus largement, la recherche dans le domaine de la reproduction, en contribuant à améliorer notre compréhension des mécanismes moléculaires et au diagnostic des pathologies de la reproduction.

Transcriptome *versus* protéome pour l'étude de la spermatogenèse

La protéomique a considérablement progressé au cours des dernières années et peut maintenant être considérée comme parvenue à maturité, mais il n'est toujours pas évident qu'elle corresponde aux technologies de la transcriptomique en termes de facilité d'utilisation et de cohérence des résultats obtenus. De toute évidence, la transcriptomique présente une plus grande capacité de débit que la protéomique, mais la diversité des protéines ne peut pas être entièrement caractérisée par des analyses d'expression des gènes seules. En outre, la complexité connue des différents mécanismes d'expression de gènes chez les mammifères est en partie responsable des écarts fréquemment rapportés entre les niveaux d'ARNm et l'abondance des protéines. Il apparaît donc évident que l'on ne doit pas choisir entre ces technologies et qu'une combinaison de ces deux approches apparaît comme une solution intéressante, en augmentant la quantité d'informations produites à différents niveaux.

© John Libbey Eurotext, 2012

Tiré à part auteur

En plus d'établir des cartes de référence du protéome et d'identifier de nouvelles protéines d'intérêt, les études protéomiques ont mis en évidence des divergences entre les transcriptomes des testicules entiers et des cellules germinales et leurs protéomes respectifs. Plusieurs études ont étudié les corrélations existant entre les données transcriptomiques et protéomiques, en essayant de prendre en compte les mécanismes de régulation de la traduction. À l'aide d'une technologie d'identification multidimensionnelle des protéines (MudPIT), Cagney *et al*. ont étudié des extraits enrichis en protéines nucléaires issus de huit tissus humains [58]. Ils ont comparé les profils d'expression de 683 protéines non redondantes avec ceux obtenus dans des expériences de puces à ADN. Les auteurs ont constaté que de tous les organes étudiés le testicule avait le plus faible coefficient de corrélation entre le transcriptome et le protéome (à savoir 0,138 contre 0,432 pour le foie, présentant la plus forte corrélation). Cette très faible corrélation peut refléter des caractéristiques originales de la régulation du gène et de la transcription au cours de la spermatogenèse.

Le réseau de communication reliant les activités cellulaires au cours de la spermatogenèse est connu pour être très complexe et est donc moins bien documenté que nombre d'autres processus biologiques. L'extraction de listes de protéines générées par des approches protéomiques systématiques ou différentielles couplées avec des ensembles de données transcriptomiques a le potentiel de fournir des informations précieuses sur des processus biologiques complexes, tels que la spermatogenèse. Nous menons actuellement un projet de génomique intégrative visant à caractériser le protéome du fluide séminifère par une approche protéomique *shotgun*. Notre objectif est d'étudier le sécrétome des cellules germinales (et des cellules de Sertoli) en recoupant la liste des protéines obtenues avec les ensembles de données du transcriptome testiculaire (cellules isolées). Nous savons en effet, depuis plus de 20 ans, que les cellules germinales modulent la fonction des cellules somatiques de Sertoli *via* des protéines diffusibles [59]. Cependant, jusqu'à présent, l'impossibilité de maintenir les cellules germinales *in vitro* sur une période suffisante rend difficile l'étude de leur sécrétome. Pour cette raison, le rôle des cellules germinales dans le contrôle de la spermatogenèse, connu dès le début des années 1990, a été complètement laissé de côté [60]. Ce nouveau projet repose sur l'hypothèse selon laquelle le fluide séminifère doit contenir des protéines sécrétées par les cellules germinales et/ou les cellules de Sertoli. Il devrait dès lors conduire à la caractérisation de nouvelles protéines impliquées dans le dialogue entre les cellules de Sertoli et les cellules germinales. Fait intéressant, selon Sato *et al*., il est maintenant possible d'obtenir des cellules germinales différenciées et des spermatozoïdes fonctionnels à partir de tissu testiculaire néonatal de souris, avec certaines méthodes de culture d'organes et sous certaines conditions de milieu [61]. Il s'agit sans aucun doute d'une étape majeure dans ce domaine, mais avant que cela devienne courant, d'autres stratégies sont encore nécessaires pour étudier le secrétome des cellules germinales.

L'exploitation des données de protéomique : vers une biologie globale de la spermatogenèse

Les études décrites dans cette revue ont contribué à identifier les protéines susceptibles de jouer un rôle majeur dans certaines étapes du développement testiculaire (*tableau 1*). Cependant, la plupart de ces études n'ont pas permis d'obtenir des résultats utiles et significatifs à partir des grandes quantités de données générées, qui permettraient d'améliorer notre compréhension des événements cellulaires qui sous-tendent la spermatogenèse. Plusieurs stratégies, se concentrant sur un groupe restreint de protéines (de gènes) d'intérêt susceptible d'être étudiés plus en détail, peuvent être utilisées. Ces stratégies comprennent des expériences de profilage des tissus, et des comparaisons inter-espèces pour la détection de gènes testicule-spécifiques ou de gènes conservés au cours de l'évolution. Les profils d'expression spécifiques et les profils d'expression de ces gènes conservés peuvent être associés à des fonctions essentielles. Ces processus de filtrage se sont révélés efficaces pour l'identification des facteurs d'intérêt à partir de grands jeux de données, mais ils ne peuvent pas combler le fossé existant entre l'identification de milliers de gènes ou des protéines co-exprimés et le décryptage des liens existant entre eux et susceptible de fournir une explication satisfaisante du bon déroulement d'un processus biologique.

Que devons-nous alors faire des listes de protéines générées par les analyses systématiques et différentielles ? Des efforts considérables ont été réalisés ces dernières années pour intégrer les données provenant d'expériences à grande échelle mais également pour mettre au point des outils permettant aux chercheurs non seulement de décrire un groupe de gènes ou de protéines avec des profils d'expression similaires, mais aussi de proposer et de développer de nouvelles hypothèses à partir de leurs analyses. Une méthode répandue pour analyser ces expériences repose sur l'utilisation des descriptions de gènes (annotations) du consortium de la GO pour l'intégration de données fonctionnelles : *data mining* [62]. L'enrichissement d'un ensemble de gènes ou de protéines en certaines catégories de termes GO permet ainsi de démontrer objectivement que certaines fonctions spécifiques sont significativement associées à un processus donné. Cette approche reste essentiellement descriptive mais elle facilite à la fois l'identification de voies qui se

© John Libbey Eurotext, 2012

Tiré à part auteur

Tableau I. Protéines clés identifiées pour la première fois par protéomique dans la gonade mâle ou pour la première fois décrites comme différentiellement exprimées dans les cellules germinales, citées dans cette revue.

Protéine	Espèce	Références	Localisation[a]	Expression différentielle dans les CGs	Approche protéomique choisie[b]	Numéro d'accession UniProt	Termes de la *GeneOntology* (GO)
AOP1A/PRDX3, *Thioredoxin-dependent peroxide reductase, mitochondrial*	Souris	[46]	T, abondante dans SPG	Oui	2-DE/MS	Q8K4K8	Organisation mitochondriale - activité antioxidante - régulation positive de la prolifération cellulaire - régulation négative des processus apoptotiques - régulation négative de l'activité des kinases
Calm I	Souris	[53]	eSPT, cSPT	-	Extraction acide/MS	Q9D6G4	Liaison au calcium - liaison de protéines
Cofilin (CFLI)	Souris	[53]	eSPT, cSPT	-	Colonne TPI/MS	COF1	Liaison aux filaments d'actine - cytokinèse
Casein-like phosphoprotein, Calcium-binding and spermatid-specific protein 1 (CLPH)	Rat	[38, 49]	SPT	Oui	2D-DIGE/MS	Q68FX6	Spermatogenèse
DNA replication licensing factor MCM7	Homme	[29, 34, 38]	CGs, abondante dans SPG	Oui	2-DEF/MS	P33993	Réplication de l'ADN - prolifération cellulaire - réponse aux dommages à l'ADN - régulation de la phosphorylation
Elongation factor 1-gamma (EF1G)	Souris	[53]	eSPT, cSPT	-	Colonne TPI/MS-extraction acide/MS	Q9D8N0	Traduction
Eucaryotic translation initiation factor 5A-1 (eIF-5A)	Souris	[27]	maGSCs	-	2-DE/MS	P63242	Traduction - modification des peptidyl-lysine en hypusine - régulation négative des processus apoptotiques - transport de protéines
Fau protein	Souris	[53]	eSPT, cSPT	-	Colonne TPI/MS-extraction acide/MS	Q91V99	Traduction
Fibroblast growth factor-8 precursor (FGF8)	Poulet	[22]	PGCs	-	2-DE/MS	Q90722	Différenciation cellulaire - activation de la voie de signalisation Wint - régulation positive des divisions cellulaires
Glyceraldehyde-3-phosphate dehydrogenase (GADPH)	Rat	[38]	S, SPG	Oui	2D-DIGE/MS	P04797	Glycolyse - processus apoptotique - oxydation réduction

Tiré à part auteur

© John Libbey Eurotext, 2012

Tableau I. (*Suite*)

Protéine	Espèce	Références	Localisation[a]	Expression différentielle dans les CGs	Approche protéomique choisie[b]	Numéro d'accession UniProt	Termes de la *GeneOntology* (GO)
Galectin-1	Souris	[27]	maGSCs	-	2-DE/MS	P16045	Réponse cellulaire aux drogues / composés organo-cycliques - réponse aux stimulus glucose - régulation positive de la cascade kappa B kinase/NFkB - régulation négative de l'adhésion cellule/substrat
Grp58, *Protein disulfide-isomerase* A3	Rat	[38]	SPG, SPC, abondante dans SPT (acrosome)	Oui	2D-DIGE/MS	P11598	Homéostasie cellulaire redox - régulation positive des processus apoptotiques
Glutathione S-transferase Mu 2 (GSTM2)	Homme	[46]	T, abondante dans SPC primaires	Oui	2-DE/MS	P28161	Processus métabolique du glutathion - processus métabolique des xénobiotiques
Heterogeneous nuclear ribonucleoprotein A1 (hnRPA1)	Souris	[12]	TF	-	2-DE/MS	P49312	Splicing alternatif des ARNm *via* les spliceosomes - processing des ARNm - export nucléaire
Heat shock cognate 71kDa protein (HSC71)	Souris	[12, 27]	TF	-	2-DE/MS	P63017	Réponse au stress - repliement de protéines - régulation négative de la transcription, dépendante de l'ADN - régulation du cycle cellulaire
		[27]	maGSCs	-	2-DE/MS		
Heat shock-related 70 kDa protein 2 (HSPA2)	Souris	[53]	eSPT, cSPT	-	TPI colomn/MS-acidic extraction/MS	P17156	Régulation positive de la phosphorylation des protéines - réponse au stress - méiose mâle (I) - développement des spermatides - désassemblage des complexes synaptonémaux
Heat shock 70-kDa protein 5 (HSPA5)	Poulet	[22]	PGCs	-	2-DE/MS	Q90593	Régulation négative des processus apoptotiques - réponse cellulaire à la privation de glucose - régulation négative de la voie du récepteur au *transforming growth factor beta* (TGF-bêta)
Lactoylglutathione lyase	Souris	[27]	maGSCs	-	2-DE/MS	Q9CPU0	Processus métabolique du glutathion et des carbohydrates - anti-apoptose - régulation de la transcription - processus métabolique du méthyglyoxal

Tiré à part auteur

© John Libbey Eurotext, 2012

Tableau I. (*Suite*)

Protéine	Espèce	Références	Localisation[a]	Expression différentielle dans les CGs	Approche protéomique choisie[b]	Numéro d'accession UniProt	Termes de la *GeneOntology* (GO)
Nucleoplasmin-3	Souris	[53]	eSPT, cSPT	-	Colonne TPI/MS-extraction acide/MS	Q9CPP0	Processing des ARNr - transcription des ARNr
Phosphoglycerate kinase 2 (PGK2)	Homme	[46]	T, abondante dans SPZ	Oui	2-DE/MS	P07205	Glycolyse - phosphorylation
Polypyrimidine tract-binding protein 2	Rat	[38]	SPC, SPT	Oui	2D-DIGE/MS	Q66H20	Processing des ARNm - ARN splicing
PRDX4, Peroxiredoxin-4	Souris	[46]	T, abondante dans SPZ	Oui	2-DE/MS	O08807	Oxydation - réduction
Smac/Diablo	Souris	[38, 50]	SPC-SPT	Oui	2D-DIGE/MS	Q542V8	Induction de l'apoptose
Stathmin	Rat	[29, 35]	GCs abondante dans cSPT	Oui	2-DE/MS	P13668	Développement du système nerveux - développement du cerveau - dynamique des microtubules - différenciation cellulaire
Translationally controlled tumor protein (TCTP)	Rat	[29, 36]	T, abondante dans SPG	Oui	2-DE/MS	P63029	Spermatogenèse - prolifération cellulaire
Thymosin β-10	Rat	[57]	SPC, abondante dans SPT, eSPT, CRs	Oui	MALDI-IMS	P63312	Développement des spermatides - organisation du cytosquelette d'actine - séquestration des monomères d'actine
Thymosin β-4	Rat	[57]	Abondante dans SPC, abondante dans rSPT ; eSPT (tête), CRs	Oui	MALDI-IMS	P62329	Organisation du cytosquelette d'actine - séquestration des monomères d'actine
Polymorphic tumor rejection antigen (TRA1)	Souris	[12]	TF	-	2-DE/MS	P08113	Anti-apoptose - réponse au stress/hypoxie - repliement de protéines - processus cataboliques liés au réticulum endoplasmique - assemblage des filaments d'actine - régulation de l'activité phosphatasique
Vimentin	Poulet	[22]	PGCs	-	2-DE/MS	P09654	

[a] T : testicule ; TF : testicule fœtal ; PGCs : cellules germinales primordiales ; maGSCs : cellules souches multipotentes de la lignée germinale adulte ; SPG : spermatogonies ; SPC : spermatocytes ; SPT : spermatides ; rSPT : spermatides rondes ; eSPT : spermatides allongées ; cSPT : spermatides condensées ; CRs : corps résiduels ; SPZ : spermatozoïdes ; S : cellules de Sertoli ; CGs : cellules germinales.
[b] Approches protéomiques pour la purification de protéines suivie d'une identification par spectrométrie de masse (MS). 2D-DIGE : *two-dimensional fluorescence difference gel electrophoresis* ; 2-DE : électrophorèse bidimensionnelle ; MALDI : *matrix assisted laser desorption/ionization* ; TPI column : colonne chromatographique avec billes de CNBr recouvertes de protéines de transition.

Tiré à part auteur

© John Libbey Eurotext, 2012

révèlent importantes et la prédiction de fonctions de gènes non caractérisées à ce jour. Elle peut aussi apporter de l'information supplémentaire sur des profils transcriptionnels observés et/ou faciliter, à partir d'ensembles de gènes ou de protéines co-exprimées, l'identification des gènes ou des protéines appartenant à un même complexe.

Une autre approche particulièrement utile dans ce domaine repose sur la compilation rationnelle des données omiques à petite et à grande échelle dans un entrepôt de données focalisé sur la reproduction, tel que la base de données *GermOnline* [63, 64]. Cette base de connaissances compile des études pertinentes sur le cycle cellulaire, la gamétogenèse et la fertilité. Elle contient une combinaison unique d'informations et intègre un navigateur dans les systèmes inter-espèces capable de fournir des annotations concernant les séquences d'ADN, les relations évolutives, l'expression des gènes et leur fonction. La base de données, reposant sur le navigateur de génome *Ensembl*, couvre huit organismes modèles et *H. sapiens*, pour lesquels des données complètes d'annotation du génome sont disponibles. Primig *et al.* s'efforcent actuellement d'intégrer des ensembles de données de protéomique dans *GermOnline,* ce qui permettra d'en faire un outil d'aide à la décision et à la formulation d'hypothèses.

Faciliter l'interprétation de grands ensembles de données multiples générés dans des expériences de génomique à haut débit exige des outils d'analyse flexibles et faciles à utiliser. Dans ce contexte, la suite logicielle AMEN (*Annotation, mapping, expression and network*) dédiée à la biologie moléculaire systématique [33] permet aux biologistes de réaliser et d'explorer l'annotation d'un génome, la cartographie chromosomique, les interactions protéine-protéine, les profils d'expression et les données de protéomique, sans avoir besoin de compétences avancées en bioinformatique. Nous encourageons fortement les chercheurs dans le domaine de la reproduction à adopter le logiciel AMEN qui peut être téléchargé librement sur http://sourceforge.net/projects/amen/. La version actuelle fournit des modules pour :

– le téléchargement et le prétraitement des données issues des expériences de profiling d'expression sur puces ;

– la détection des groupes de gènes co-exprimés de manière significative ;

– la recherche d'un enrichissement de ces groupes en certaines annotations fonctionnelles.

En outre, l'interface utilisateur du logiciel est conçue de manière à permettre la visualisation simultanée de plusieurs types de données tels que les réseaux d'interactions protéine-protéine avec les profils d'expression et de co-localisation cellulaire. Des efforts sont actuellement déployés pour rendre possible la prise en charge de la prochaine génération de jeux de données tels que le microRNAome et le métabolome testiculaires.

Conclusion

La protéomique est en constante évolution, et a déjà permis des avancées majeures dans l'étude des protéines. En revanche, couvrir la gamme dynamique de l'expression des protéines dans les cellules reste un défi majeur, car il n'existe actuellement aucune technologie disponible pour amplifier les protéines minoritaires au sein d'un échantillon biologique. Cependant, la sensibilité des instruments s'améliore rapidement, et une couverture étendue du protéome devrait bientôt être atteinte par le biais d'études à grande échelle [65, 66]. De même, la quantification absolue de protéines au sein de mélanges complexes devient possible grâce à des stratégies novatrices, telles que la technologie *protein standard absolute quantification* (PSAQ) [67]. Le développement d'outils permettant de suivre la pléthore de modifications post-traductionnelles existantes, dont la phosphorylation, est également en plein essor [68]. Le marquage isotopique stable de cultures cellulaires par des acides aminés ou *stable isotope labeling by amino acids in cell culture* (SILAC) est un outil polyvalent permettant la comparaison quantitative de protéomes d'organes ou de cellules, chez différentes souches de souris, y compris des modèles de souris *knock-out* sous diverses conditions *in vivo* [69]. Les progrès de l'imagerie MALDI décrite précédemment devraient également susciter l'intérêt des biologistes et des cliniciens, dès lors que les avantages de cette technologie deviennent de plus en plus reconnus.

Les travaux évoqués dans cette revue ont permis d'accroître nos connaissances sur la spermatogenèse. Néanmoins, des efforts importants sont encore nécessaires pour améliorer notre compréhension des réseaux de communication très sophistiqués reliant les activités cellulaires au cours de ce processus. Il devrait maintenant être possible d'avoir une vision beaucoup plus claire des mécanismes moléculaires intervenant à chaque étape de la spermatogenèse, grâce à l'utilisation pertinente de techniques éprouvées en protéomique et à l'exploitation des données produites par des stratégies de génomique intégrative. Par ailleurs, la protéomique *top-down* et la peptidomique, qui sont encore des approches émergentes constituent de précieux outils qui permettront de répondre à des questions précises relatives à la spermatogenèse normale et pathologique. Il ne fait nul doute que des découvertes biologiques marquantes verront le jour dans le domaine grâce à une nouvelle génération d'études omiques comparatives.

Conflits d'intérêts : aucun.

Remerciements
Les auteurs tiennent à remercier le Dr Frédéric Chalmel pour ses réflexions stimulantes et pour l'aide à la fouille de données par le logiciel AMEN. Ce travail a été soutenu en partie par Biogenouest et

© John Libbey Eurotext, 2012

Tiré à part auteur

par les infrastructures en biologie santé et agronomie (IBISA), le Fonds européen de développement régional (FEDER) et par des subventions du Conseil régional de Bretagne.

Références

1. Leblond CP, Clermont Y. Definition of the stages of the cycle of the seminiferous epithelium in the rat. *Ann N Y Acad Sci* 1952 ; 55 : 548-73.

2. Perey B, Clermont Y, Leblond CP. The wave of the seminiferous epithelium in the rat. *Am J Anat* 1961 ; 108 : 47-77.

3. De Kretser D, Kerr J. The cytology of the testis. *The physiology of reproduction*. Eds Knobil E, Neil JD, Ewing LL, Greenwald GS, Markert CL, Pfaff DW. New York, NY, Raven Press, 1988 : 837-932.

4. Parvinen M. Regulation of the seminiferous epithelium. *Endocr Rev* 1982 ; 3 : 404-17.

5. Hess RA, Renato de Franca L. Spermatogenesis and cycle of the seminiferous epithelium. *Adv Exp Med Biol* 2008 ; 636 : 1-15.

6. de Rooij DG, Russell LD. All you wanted to know about spermatogonia but were afraid to ask. *J Androl*. 2000 ; 21 : 776-98.

7. Matzuk MM, Lamb DJ. Genetic dissection of mammalian fertility pathways. *Nat Cell Biol* 2002 ; 4(Suppl) : s41-49.

8. Wrobel G, Primig M. Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. *Reproduction* 2005 ; 129 : 1-7.

9. Wilhelm D, Palmer S, Koopman P. Sex determination and gonadal development in mammals. *Physiol Rev* 2007 ; 87 : 1-28.

10. Ewen K, Baker M, Wilhelm D, Aitken RJ, Koopman P. Global survey of protein expression during gonadal sex determination in mice. *Mol Cell Proteomics* 2009 ; 8 : 2624-41.

11. Sato Y, Shinka T, Chen G, *et al*. Proteomics and transcriptome approaches to investigate the mechanism of human sex determination. *Cell Biol Int* 2009 ; 33 : 839-47.

12. Wilhelm D, Huang E, Svingen T, Stanfield S, Dinnis D, Koopman P. Comparative proteomic analysis to study molecular events during gonad development in mice. *Genesis* 2006 ; 44 : 168-76.

13. Raffalli-Mathieu F, Glisovic T, Ben-David Y, Lang MA. Heterogeneous nuclear ribonucleoprotein A1 and regulation of the xenobiotic-inducible gene Cyp2a5. *Mol Pharmacol* 2002 ; 61 : 795-9.

14. Mazzarella RA, Green M. ERp99, an abundant, conserved glycoprotein of the endoplasmic reticulum, is homologous to the 90-kDa heat shock protein (hsp90) and the 94-kDa glucose regulated protein (GRP94). *J Biol Chem* 1987 ; 262 : 8875-83.

15. Marzec M, Eletto D, Argon Y. GRP94: An HSP90-like protein specialized for protein folding and quality control in the endoplasmic reticulum. *Biochimica Biophysica Acta* 2012 ; 1823 : 774-87.

16. Agashe VR, Hartl FU. Roles of molecular chaperones in cytoplasmic protein folding. *Semin Cell Dev Biol* 2000 ; 11 : 15-25.

17. Marshall OJ, Harley VR. Identification of an interaction between SOX9 and HSP70. *FEBS Lett* 2001 ; 496 : 75-80.

18. De Santa Barbara P, Bonneaud N, Boizet B, *et al*. Direct interaction of SRY-related protein SOX9 and steroidogenic factor 1 regulates transcription of the human anti-Müllerian hormone gene. *Mol Cell Biol* 1998 ; 18 : 6653-65.

19. Hossain A, Saunders GF. Role of Wilms tumor 1 (WT1) in the transcriptional regulation of the Mullerian-inhibiting substance promoter. *Biol Reprod* 2003 ; 69 : 1808-14.

20. Viger RS, Mertineit C, Trasler JM, Nemer M. Transcription factor GATA-4 is expressed in a sexually dimorphic pattern during mouse gonadal development and is a potent activator of the Müllerian inhibiting substance promoter. *Dev Suppl* 1998 ; 125 : 2665-75.

21. Maheswaran S, Englert C, Zheng G, *et al*. Inhibition of cellular proliferation by the Wilms tumor suppressor WT1 requires association with the inducible chaperone Hsp70. *Genes Dev* 1998 ; 12 : 1108-20.

22. Han BK, Kim JN, Shin JH, *et al*. Proteome analysis of chicken embryonic gonads: identification of major proteins from cultured gonadal primordial germ cells. *Mol Reprod Dev* 2005 ; 72 : 521-9.

23. Rolland AD, Jégou B, Pineau C. Testicular development and spermatogenesis: harvesting the postgenomics bounty. *Adv Exp Med Biol* 2008 ; 636 : 16-41.

24. Calvel P, Rolland AD, Jégou B, Pineau C. Testicular postgenomics: targeting the regulation of spermatogenesis. *Philos Trans R Soc Lond B Biol Sci* 2010 ; 365 : 1481-500.

25. Nakagawa T, Nabeshima YI, Yoshida S. Functional identification of the actual and potential stem cell compartments in mouse spermatogenesis. *Dev Cell* 2007 ; 12 : 195-206.

26. Caires K, Broady J, McLean D. Maintaining the male germline: regulation of spermatogonial stem cells. *J Endocrinol* 2010 ; 205 : 133-45.

27. Dihazi H, Dihazi GH, Nolte J, *et al*. Multipotent adult germline stem cells and embryonic stem cells: comparative proteomic approach. *J Proteome Res* 2009 ; 8 : 5497-510.

28. Guan K, Nayernia K, Maier LS, *et al*. Pluripotency of spermatogonial stem cells from adult mouse testis. *Nature* 2006 ; 440 : 1199-203.

29. Com E, Evrard B, Roepstorff P, Aubry F, Pineau C. New insights into the rat spermatogonial proteome. *Mol Cell Proteomics* 2003 ; 2 : 248-61.

30. Son YS, Park JH, Kang YK, *et al*. Heat shock 70-kDa protein 8 isoform 1 is expressed on the surface of human embryonic stem cells and downregulated upon differentiation. *Stem Cells* 2005 ; 23 : 1502-13.

31. Guillaume E, Dupaix A, Moertz E, Courtens JL, Jégou B, Pineau C. Proteome analysis of spermatogonia: identification of a first set of 53 spermatogonial proteins. *Proteome* 2000. doi: 10.1007/s102160000003.

32. Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, *et al*. The conserved transcriptome in human and rodent male gametogenesis. *Proc Natl Acad Sci U S A* 2007 ; 104 : 8346-51.

33. Chalmel F, Primig M. The annotation, mapping, expression and network (AMEN) suite of tools for molecular systems biology. *BMC Bioinformatics*. 2008 ; 9 : 86-97.

34. Com E, Rolland AD, Guerrois M, *et al*. Identification, molecular cloning, and cellular distribution of the rat homolog of minichromosome maintenance protein 7 (MCM7) in the rat testis. *Mol Reprod Dev* 2006 ; 73 : 866-77.

Tiré à part auteur

© John Libbey Eurotext, 2012

35. Guillaume E, Evrard B, Com E, Moertz E, Jégou B, Pineau C. Proteome analysis of rat spermatogonia: reinvestigation of stathmin spatio-temporal expression within the testis. *Mol Reprod Dev* 2001 ; 60 : 439-45.

36. Guillaume E, Pineau C, Evrard B, *et al*. Cellular distribution of translationally controlled tumor protein in rat and human testes. *Proteomics* 2001 ; 1 : 880-9.

37. Loppion G, Lavigne R, Pineau C, Auvray P, Sourdaine P. Proteomic analysis of the spermatogonial stem cell compartment in dogfish Scyliorhinus canicula L. *Comp Biochem Physiol Part D, Genomics & Proteomics*. 2010 ; 5 : 157-64.

38. Rolland AD, Evrard B, Guitton N, *et al*. Two-dimensional fluorescence difference gel electrophoresis analysis of spermatogenesis in the rat. *J Proteome Res* 2007 ; 6 : 683-97.

39. Chu DS, Liu H, Nix P, *et al*. Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. *Nature* 2006 ; 443 : 101-5.

40. Govin J, Caron C, Escoffier E, *et al*. Post-meiotic shifts in HSPA2/HSP70.2 chaperone activity during mouse spermatogenesis. *J Biol Chem* 2006 ; 281 : 37888-92.

41. Takemori N, Yamamoto MT. Proteome mapping of the Drosophila melanogaster male reproductive system. *Proteomics* 2009 ; 9 : 2484-93.

42. Huang SY, Lin JH, Chen YH, *et al*. A reference map and identification of porcine testis proteins using 2-DE and MS. *Proteomics* 2005 ; 5 : 4205-12.

43. Zhu YF, Cui YG, Guo XJ, *et al*. Proteomic analysis of effect of hyperthermia on spermatogenesis in adult male mice. *J Proteome Res* 2006 ; 5 : 2217-25.

44. Guo X, Zhao C, Wang F, *et al*. Investigation of human testis protein heterogeneity using 2-dimensional electrophoresis. *J Androl* 2010 ; 31 : 419-29.

45. Guo X, Zhang P, Huo R, Zhou Z, Sha J. Analysis of the human testis proteome by mass spectrometry and bioinformatics. *Proteomics Clin Appl* 2008 ; 2 : 1651-7.

46. Huang XY, Guo XJ, Shen J, *et al*. Construction of a proteome profile and functional analysis of the proteins involved in the initiation of mouse spermatogenesis. *J Proteome Res* 2008 ; 7 : 3435-46.

47. Paz M, Morín M, Del Mazo J. Proteome profile changes during mouse testis development. *Comp Biochem Physiol Part D, Genomics & Proteomics* 2006 ; 1 : 404-15.

48. Huang SY, Lin JH, Teng SH, *et al*. Differential expression of porcine testis proteins during postnatal development. *Anim Reprod Sci* 2011 ; 123 : 221-33.

49. Calvel P, Kervarrec C, Lavigne R, *et al*. CLPH, a novel casein kinase 2-phosphorylated disordered protein, is specifically associated with postmeiotic germ cells in rat spermatogenesis. *J Proteome Res* 2009 ; 8 : 2953-65.

50. Tikoo A, O'Reilly L, Day CL, Verhagen AM, Pakusch M, Vaux DL. Tissue distribution of Diablo/Smac revealed by monoclonal antibodies. *Cell Death Differ* 2002 ; 9 : 710-6.

51. Vera Y, Diaz-Romero M, Rodriguez S, *et al*. Mitochondria-dependent pathway is involved in heat-induced male germ cell death: lessons from mutant mice. *Biol Reprod* 2004 ; 70 : 1534-40.

52. Huang XY, Sha JH. Proteomics of spermatogenesis: from protein lists to understanding the regulation of male fertility and infertility. *Asian J Androl* 2011 ; 13 : 18-23.

53. Govin J, Gaucher J, Ferro M, *et al*. Proteomic strategy for the identification of critical actors in reorganization of the post-meiotic male genome. *Mol Hum Reprod* 2012 ; 18 : 1-13.

54. Balhorn R. The protamine family of sperm nuclear proteins. *Genome Biology*. 2007 ; 8 : 227-234.

55. Arpanahi A, Brinkworth M, Iles D, *et al*. Endonuclease-sensitive regions of human spermatozoal chromatin are highly enriched in promoter and CTCF binding sequences. *Genome Res* 2009 ; 19 : 1338-49.

56. Hammoud SS, Nix DA, Zhang H, Purwar J, Carrell DT, Cairns BR. Distinctive chromatin in human sperm packages genes for embryo development. *Nature* 2009 ; 460 : 473-8.

57. Lagarrigue M, Becker M, Lavigne R, *et al*. Revisiting rat spermatogenesis with MALDI imaging at 20-microm resolution. *Mol Cell Proteomics* 2011 ; 10 : M110.005991-M110.

58. Cagney G, Park S, Chung C, *et al*. Human tissue profiling with multidimensional protein identification technology. *J Proteome Res* 2005 ; 4 : 1757-67.

59. Jégou B. The Sertoli-germ cell communication network in mammals. *Int Rev Cytol* 1993 ; 147 : 25-96.

60. Jégou B, Pineau C, Dupaix A. *Paracrine control of testis function. Male reproductive function Wang C Ed Endocrine update series*. Berlin : Kluwer Academic, 1999 ; 41-64.

61. Sato T, Katagiri K, Gohbara A, *et al*. *In vitro* production of functional sperm in cultured neonatal mouse testes. *Nature* 2011 ; 471 : 504-7.

62. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000 ; 25 : 25-9.

63. Primig M, Wiederkehr C, Basavaraj R, *et al*. GermOnline, a new cross-species community annotation database on germ-line development and gametogenesis. *Nat Genet* 2003 ; 35 : 291-2.

64. Lardenois A, Gattiker A, Collin O, Chalmel F, Primig M. GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. *Database* 2010 ; 2010 : baq030.

65. de Godoy LMF, Olsen JV, de Souza GA, Li G, Mortensen P, Mann M. Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biol* 2006 ; 7 : R50-R.

66. de Godoy LMF, Olsen JV, Cox J, *et al*. Comprehensive mass-spectrometry-based proteome quantification of haploid *versus* diploid yeast. *Nature* 2008 ; 455 : 1251-4.

67. Brun V, Dupuis A, Adrait A, *et al*. Isotope-labeled protein standards: toward absolute quantitative proteomics. *Mol Cell Proteomics* 2007 ; 6 : 2139-49.

68. Hilger M, Bonaldi T, Gnad F, Mann M. Systems-wide analysis of a phosphatase knock-down by quantitative proteomics and phosphoproteomics. *Mol Cell Proteomics* 2009 ; 8 : 1908-20.

69. Krüger M, Moser M, Ussar S, *et al*. SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* 2008 ; 134 : 353-64.

Tiré à part auteur

© John Libbey Eurotext, 2012

ABSTRACT

Spermatogenesis in mammals is a complex biological function including cellular processes such as proliferation, meiosis and differentiation aiming to the production of male gametes in the testis. If the theatre of this process, the seminiferous epithelium, is well described in terms of organization and cellular morphology of cells that compose it, the molecular mechanisms underlying spermatogenesis are not yet fully decrypted. These processes by which diploid undifferentiated germ cells, spermatogonia, enter meiosis to give haploid cells that undergo many morphological changes including compaction of the genome, formation of the acrosome and flagellum are largely borne by the Sertoli nurse cells, which are in close mechanical and functional interaction with germ cells. In this context, germ cells and Sertoli cells maintain a dialogue necessary to the success of spermatogenesis and spermiogenesis, on which relies the production of 100 million sperm per day in rats. Spermatogenesis is based on the coordinated and sequential expression of genes, including specific products for each stage of germ cell development. These gene products are essential at each key stage of spermatogenesis and its smooth running. Transcriptomics since the 1990s, and proteomics since the 2000s have contributed to the improved understanding of these mechanisms. Two major studies have been conducted in our unit and have been the basis of my thesis work. First, a long term proteomic study aiming at characterizing the testicular proteome with a focus on the germ line, and the other one, a very recent study that characterized and quantified the transcriptome of isolated rat testicular cells at high resolution using a *de novo* sequencing of transcripts (RNA-seq) approach. The latter study showed the accumulation of long non-coding RNAs (lncRNAs) and testicular unannotated transcripts (TUTs) at meiotic and post-meiotic stages of spermatogenesis in the rat. In this context, my thesis work aimed at validating the coding potential of many genes expressed in germ cells using RNA-seq combined with shotgun proteomics, a so-called PIT (Proteomics Informed by transcriptomics) approach. In this approach, the protein sequences translated from the transcripts assembled by RNA-seq in the different testicular cell types are integrated into a custom database of protein sequences used to query mass spectrometry data obtained from protein extracts from meiotic and post-meiotic cells. On the other hand, my work aimed at identifying membrane proteins in germ cells and residual bodies that may be involved in the dialogue between Sertoli cells and germ cells using a ICPL relative quantification approach. In addition, I was able to provide the first proteome of rat Sertoli cells, germ cells and residual bodies by shotgun proteomics. The peptides identified by MS/MS in all these cell types are visually available on the ReproGenomicsViewer (http://rgv.genouest.org/). The PIT approach showed that 69 TUTs or lncRNA (corresponding to 44 loci) code for proteins in meiotic cells and post meiotic cells, and we confirmed experimentally the meiotic and post-meiotic expression for two new transcripts encoding for VAMP9, a protein of the SNARE family, and a new testicular enolase T-ENOL. The post-meiotic expression of T-ENOL protein was confirmed by immunohistochemistry using a polyclonal antibody raised against the recombinant protein. This approach also allowed us to identify new isoforms of known proteins, specific to each stage of spermatogenesis. Differential analysis of membrane proteins of germ cells and residual bodies by ICPL enabled us to establish a list of 166 proteins whose expression is differential between pachytene spermatocytes, round spermatids and residual bodies. Their differential expression suggests that these proteins may play a role in spermiogenesis. Thanks to the Gene Ontology annotations, a list of eight proteins with a putative role in signal transduction, cell recognition or differentiation, thus potentially involved in the dialogue between Sertoli and germ cells was drawn. This study also showed the overexpression in germ cells of some Rab family proteins and TMED family proteins recently recognized as having a regulatory role of innate immunity.

RÉSUMÉ

La spermatogenèse chez les mammifères est une fonction biologique complexe incluant des processus de prolifération cellulaire, de méiose et de différenciation uniques visant à la production des gamètes mâles au sein du testicule. Si l'épithélium séminifère, théâtre de ce processus est bien décrit sur le plan de son organisation et de la morphologie des cellules qui le composent, les mécanismes moléculaires qui sous-tendent la spermatogenèse ne sont pas encore complètement décryptés. Ces processus par lesquels les cellules germinales diploïdes indifférenciées, les spermatogonies, entrent en méiose pour donner ensuite des cellules haploïdes qui subissent de nombreuses transformations morphologiques incluant la compaction du génome, la formation de l'acrosome et du flagelle sont largement supportés par les cellules de Sertoli, cellules nourricières, qui sont en étroite interaction mécanique et fonctionnelle avec les cellules germinales. Dans ce contexte, les cellules germinales et les cellules de Sertoli entretiennent le dialogue nécessaire au bon déroulement de la spermatogenèse et de la spermiogénèse, et duquel dépend la production de 100 millions de spermatozoïdes par jour chez le rat. La spermatogenèse repose sur l'expression coordonnée et séquentielle des gènes, dont les produits spécifiques de chaque stade de développement des cellules germinales sont essentiels à chaque étape clé et à son bon déroulement. La transcriptomique depuis les années 1990, et la protéomique depuis les années 2000 ont contribué à l'amélioration de la connaissance de ces mécanismes. Deux études importantes ont été menées dans notre unité, et qui ont été à la base de mes travaux de thèse. D'une part une étude de protéomique sur le long terme visant à caractériser par des approches systématiques et différentielles le protéome testiculaire avec un focus sur celui de la lignée germinale. D'autre part, une étude très récente qui a permis de caractériser et quantifier le transcriptome des cellules testiculaires isolées de rat, avec une haute résolution, en utilisant une approche de séquençage *de novo* des transcrits (RNA-seq). Cette dernière étude a mis en évidence l'accumulation de longs ARNs non codants (lncRNAs), et de transcrits testiculaires non annotés (TUTs) aux stades méiotique et post-méiotique de la spermatogenèse, chez le rat. Dans ce contexte, mon travail de thèse a consisté à valider le potentiel codant de nombreux gènes exprimés dans les cellules germinales, par une approche dite PIT (Proteomics Informed by Transcriptomics) couplant protéomique Shotgun et RNA-seq. Dans ce type d'approche, les séquences protéiques déduites des transcrits des différents types cellulaires du testicule assemblés par RNA-seq sont intégrées dans une base personnalisée de séquences protéiques utilisée pour interroger les données de spectrométrie de masse obtenues à partir d'extraits de protéines totales de cellules méiotiques et post-méiotiques. Mon travail a consisté d'autre part à identifier des protéines membranaires des cellules germinales et des corps résiduels susceptibles d'intervenir dans le dialogue entre les cellules de Sertoli et les cellules germinales, par une approche de quantification relative ICPL. Par ailleurs, j'ai pu établir par protéomique Shotgun un premier protéome des cellules de Sertoli, des cellules germinales et des corps résiduels de rat. Les peptides identifiés par MS/MS sur l'ensemble de ces types cellulaires sont visuellement accessibles sur le ReproGenomicsViewer (http://rgv.genouest.org/). L'approche PIT a permis de montrer que 69 TUTs ou lncRNA (correspondant à 44 loci) codent pour des protéines dans les cellules méiotiques et post méiotiques chez le rat, et de confirmer expérimentalement l'expression post-méiotique de deux nouveaux transcrits, codant pour la protéine VAMP9, une protéine de la famille SNARE, et pour une nouvelle énolase T-ENOL. L'expression post-méiotique de la nouvelle protéine T-ENOL a été confirmée par immunohistochimie à l'aide d'un anticorps polyclonal produit contre la protéine recombinante. Cette approche nous a également permis d'identifier de nouvelles isoformes de protéines connues, spécifiques de chaque stade de la spermatogenèse. L'analyse différentielle par ICPL des protéines membranaires des cellules germinales et des corps résiduels nous a permis d'établir une liste de 166 protéines dont l'expression est différentielle entre les spermatocytes pachytène, les spermatides rondes et les corps résiduels. Du fait de cette expression différentielle, ces protéines sont susceptibles de jouer un rôle dans la spermiogénèse. Grâce aux annotations de le Gene Ontology, j'ai pu établir une liste de 8 protéines ayant un rôle supposé dans la transduction du signal, la reconnaissance cellulaire ou bien la différenciation, donc potentiellement impliquées dans le dialogue entre les cellules de Sertoli et les cellules germinales. Cette étude a montré par ailleurs la surexpression dans les cellules germinales, de quelques protéines de la famille des Rab et de la famille des TMED récemment reconnues comme ayant un rôle de régulation dans l'immunité innée.