Deciphering the Dark Proteome Use of the Testis and Characterization of Two Dark Proteins

Nathalie Melaine, Emmanuelle Com, Pascale Bellaud, Laëtitia Guillot, Mélanie Lagarrigue, Nick A Morrice, Blandine Guével, Regis Lavigne, Juan-Felipe Velez de La Calle, Jörg Dojahn, et al.

Deciphering the dark proteome: use of the testis and characterization of two dark proteins

Nathalie Melaine [1, 2], *Emmanuelle Com* [1, 2], *Pascale Bellaud* [4], *Laetitia Guillot* [1, 2], *Mélanie Lagarrigue* [1, 2], *Nick A. Morrice* [4], *Blandine Guével* [1, 2], *Régis Lavigne* [1, 2], *Juan-Felipe Velez de la Calle* [6], *Jörg Dojahn* [5] *and Charles Pineau** [1, 2]

1 Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-35042 Rennes cedex, France

2 Protim, Univ Rennes, F-35042 Rennes cedex, France

3 H2P2 Core Facility, UMS BioSit, Univ Rennes, Rennes F-35040, France

4 Sciex, Phoenix House Lakeside Drive Centre Park Warrington WA1 1RX, UK

5 Sciex, Landwehrstr. 54, 64293 Darmstadt, Germany

6 Unité FIV, Clinique Pasteur, 34, Rue du Moulin à Poudre, F-29000 Brest, France

ABSTRACT

For the C-HPP consortium, dark proteins include not only uPE1, but also missing proteins (MPs, PE 2-4), smORFs, proteins from lncRNAs, and products from uncharacterized transcripts. Here, we investigated the expression of dark proteins in the human testis by combining public mRNA and protein expression data for several tissues and performing LC−MS/MS analysis of testis protein extracts. Most uncharacterized proteins are highly expressed in the testis. Thirty could be identified in our dataset, of which two were selected for further analyses: 1) A0AOU1RQG5, a putative cancer/testis antigen specifically expressed in the testis, where it accumulates in the cytoplasm of elongated spermatids; and 2) PNMA6E, which is enriched in the testis, where it is found in the germ cell nuclei during most stages of spermatogenesis. Both proteins are coded on Chromosome X. Finally, we studied the expression of other dark proteins, uPE1 and MPs, in a series of human tissues. Most were highly expressed in the testis at both the mRNA and protein levels. The testis appears to be a relevant organ to study the dark proteome, which may have a function related to spermatogenesis and germ cell differentiation. The mass spectrometry proteomics data have been deposited with the ProteomeXchange Consortium under the data set identifier PXD009598.

Keywords: human proteome project, testis, missing proteins, uPE1, dark proteins, immunohistochemistry, data mining

The mass spectrometry data have been deposited to the PRIDE Archive (http://www.ebi.ac.uk/pride/archive/) via the PRIDE partner repository with the data set identifier PXD009598 and 10.6019/PXD009598.
Username: reviewer96528@ebi.ac.uk
Password: PrYUSRqp

INTRODUCTION

The primary function of the testis is the production of male gametes, also known as spermatozoa. In mammals, this organ can be divided in two compartments: the seminiferous tubules and the interstitium. The interstitium is a rich connective tissue matrix that mostly hosts Leydig cells, of which the primary function is the production of testosterone in the adult testis. The seminiferous tubules, which account for 60 to 80% of the total testis volume in mammals [1], are the site of spermatogenesis, a complex, intricate, tightly controlled, and specialized process [2,3]. Correct operation of the communication network between nurturing somatic cells and germ cells depends on the continuous production of nearly 1,000 gametes/s in an adult man [4]. The effectiveness of this communication network is such that dysfunction of any one of the cellular types which contribute to it has an inevitable cascading effect on other cell types. Thus, the structural and biochemical organization of the human testis makes it one of the most complex organs in the body.

Phase 1 of the Chromosome-Centric Human Proteome Project (C-HPP) aims to catalogue proteins as gene products encoded by the human genome, in a chromosome-centric manner [5], and confirm their existence at the protein level by mass spectrometry. Since the C-HPP was launched in 2012, with the help of coordinated efforts worldwide, neXtProt, the reference protein knowledge-base for C-HPP[6], now contains 19,656 entries (release 2018-01-17). The number of experimentally validated proteins (PE1) has reached 17,470 (89%), whereas the actual count of missing proteins (MPs) stands at 2,186. Nevertheless, 1,260 of those “PE1” have unknown functions and are referred to as “uPE1” (neXtProt 2018-01-17 release). As part of the next phase of the project, the C-HPP consortium has launched the “neXt-CP50 Challenge” (C-HPP

newsletter n°7, may 2018), which aims to functionally characterize the PE1 proteins with completely unknown functions. An experimental workflow has been proposed to fulfill this goal [7], which includes immunological detection, transcriptomic analysis and *in vitro* and *in vivo* functional studies. Additionally, the C-HPP also focuses on "dark proteins", which include not only uPE1 but also missing proteins (PE 2-4), smORFs and proteins from lncRNAs or any uncharacterized transcripts.

It was suggested that the production of proteins that have been systematically overlooked since the launch of the C-HPP may be restricted to unusual organs or cell types and, particularly, the testis.[8] It was shown over a decade ago that the testis expresses the highest number of tissue-specific genes.[9] More recently, two very elegant large-scale studies by Uhlén and collaborators[10, 11] confirmed that the testis is probably the most promising organ to search for missing proteins. Interestingly, the selective pressure on most genes involved in spermatogenesis implies a very high degree of germ-cell specific expression. Thus, approximately half of testis-specific genes are expressed in meiotic and post-meiotic germ cells, whereas a small number was identified in the premeiotic germ cell lineage or somatic testicular cells (*i.e.,* Sertoli and Leydig cells).[12] Whole testis extracts have been largely used to search for elements of the missing proteome[13, 14], whereas the Franco-Swiss contribution to the C-HPP initiative successfully combined the search for missing proteins that lack conclusive mass spectrometric evidence with an extensive examination of the sperm proteome, unambiguously validating over 300 MPs.[15, 16, 17]

The study presented in this paper originates from members of the French C-HPP initiative and aims to characterize dark proteins in the human testis. First, we analyzed the mRNA expression of 520 selected "uncharacterized" proteins extracted from the Human Protein Atlas RNA database and demonstrated that most display higher expression in the testis than in the other human tissues analyzed. We then performed trans-chromosome-based data analysis to catalogue all uncharacterized proteins in a total protein extract from human testis in which cytoplasmic and membrane proteins were preferentially extracted and peptides separated by pH fractionation. A high-quality mass spectrometry dataset was collected on a high-resolution 6600 TripleTOF® system (Sciex). A total of 5,578 proteins were unambiguously identified with at least two peptides, 30 of which were annotated as "uncharacterized". Two uncharacterized proteins were further selected based on their testis specificity. Combining publicly available expression data and immunohistochemistry, we showed that A0AOU1RQG5 is a putative cancer/testis antigen, specifically expressed in the testis, where it accumulates in the cytoplasm of elongated spermatids. We also demonstrate that PNMA6E, which has been referenced as a uPE1 since the beginning of our study, is enriched in the testis, where it is found in the germ cell nuclei during spermatogenesis. Finally, we looked for the gene and protein expression of other dark proteins, uPE1 and missing proteins (PE2-4), in a series of human tissues. Most of the uPE1 and missing proteins are highly expressed in the testis at both transcript and protein levels. Thus, the testis may be a relevant organ to study the dark proteome, as the function of most dark proteins may be related to spermatogenesis and germ cell differentiation.

MATERIALS AND METHODS

Ethics, donor consent, and sample collection and preparation

Human testes were obtained from a patient with a significant family history of cancer and diagnosed with seminoma. Voluntary ablation of the gonads, followed by testosterone supplementation, was performed at the IVF unit of the Pasteur Clinic (Brest, France) and samples were considered as biological waste. Written informed consent was obtained from the donor for their use for research. The testes were washed in PBS, frozen in liquid nitrogen, and stored at -80°C until protein extraction.

Protein extraction and digestion

A piece of frozen human testis was resuspended in extraction buffer (20 mM Hepes pH 7.4) supplemented with a protease inhibitor mix (1 mM EDTA, 0.5 mM dithiothreitol (DTT), 1 mM 4-(2-Aminoethyl)benzenesulfonyl fluoride hydrochloride (AEBSF), and 10 µM L-trans-Epoxysuccinyl-leucylamido(4-guanidino)butane (E64)). The tissue suspension was sonicated on ice. The resulting lysate was then centrifuged 30 min at 15,000 × g at 4°C. The supernatant, containing the soluble proteins, was kept on ice and the pellets retrieved in 100 mM Na_2CO_3 and sonicated. This suspension was centrifuged at 105,000 × g at 4°C for 45 min and the supernatant, enriched for membrane proteins, pooled with the first supernatant containing the soluble proteins. The pellet, containing the membrane fraction, was re-suspended in 8 M urea, 30 mM Tris, and 4% CHAPS supplemented with the same protease inhibitor mix as above, sonicated, frozen and thawed, and finally centrifuged at 105,000 x g at 4°C for 1 h. The supernatant, enriched for membrane proteins, was pooled with the first two supernatants. The protein concentration was determined using a Bradford colorimetric assay and the protein extract stored at −80°C until use.

The protein extract (407 µg in 25 µL) was diluted in 275 µL 8 M Urea, 0.1 M ammonium bicarbonate, pH 7.8. The proteins were reduced and alkylated by adding 5.3 µL 700 mM DTT,

incubating at 37°C for 30 min, and then adding 17.7 µL 135 mM iodoacetamide and incubating an additional 30 min in the dark at room temperature. Finally, 10.2 µL of a 0.4 µg/µL trypsin solution and 677 µL 100 mM ammonium bicarbonate were added to the sample, to adjust the final volume to 1 mL, and incubated overnight at 37°C. The tryptic digested sample was then desalted using 100 mg Sep-Pak tC18 reverse phase cartridges. After elution, the tryptic peptides were stored in 50-100 µL 0.1% formic acid at -80°C.

High pH fractionation of the testis tryptic digest

A human testis peptide sample (407 µg at 8 mg/mL) was injected onto a 250 x 4.6 mm Durashell RP 5 µm column (Bonna-Agela) connected to a Shimadzu Nexera HPLC system fitted with a UV detector. The column was equilibrated at 1 mL/min with buffer A (2 mM ammonium hydroxide /2% acetonitrile in water) and the peptides eluted using a discontinuous gradient of buffer A and buffer B (2 mM ammonium hydroxide /90% acetonitrile /10% water) as follows: 0 min (4% B), 6 min (4% B), 42 min (28% B), 50 min (50% B), 50.1 min (80% B), and 55 min (80% B). Fractions were collected every 2 min from 6 to 52 min and dried under vacuum. Fractions were reconstituted in 40 µL 5% acetonitrile/0.1% FA in water plus 1 µL 10x HRM peptides (Biognosys) and analyzed by microflow LC-MS as described below.

Microflow LC-MS analysis

LC-MS analysis was performed on the 6600 TripleTOF system (Sciex) connected to an NanoLC^{TM} 425 system (Sciex). The 6600 TripleTOF was operated with a Duospray^{TM} ion source fitted with a 50-mm microflow electrode in positive polarity at 5,500 V, with the GS1 at 10 and curtain gas at 25. The nanoLC ^{TM} 425 system was run with a low microflow module (1-10 mL/min) at 5 mL/min with buffer A (0.1% formic acid in water) and buffer B (0.1% formic

acid in acetonitrile). The LC system was run in the trap/elute mode with a 5 x 0.5 mm Triart C_{18} 3 mm trap column (YMC) and a 150 x 0.3 mm Triart C_{18} 3 mm trap column. Samples were injected onto the trap column at 10 mL/min with the loading pump and then eluted with a linear gradient of acetonitrile at 5 mL/min.

Ten microliters of each fraction from the high pH fractionation were analyzed in data-dependent acquisition using a linear acetonitrile gradient (0-45 min (2-40% B), 50 min (80% B), and 55 min (80% B) with the following acquisition parameters: survey scan m/z 400-1250 (250 ms) followed by Top 30 precursors (2-5+) with the intensity >150 cps. The TOF MS/MS was acquired for 50 ms in high sensitivity mode from m/z 100-1500 using rolling collision energy (CE spread 5) and precursors were excluded for 15 s after one occurrence.

MS/MS data analysis/protein identification and validation

ProteinPilot ™ 5.0 (Sciex) was used to perform automatic mass recalibration of MS and MS/MS spectra and export the recalibrated peak lists as mgf files. The Mascot server v2.2.07 (http://www.matrixscience.com) database search engine was used for peptide and protein identification, using its automatic decoy database search to calculate the false discovery rate (FDR). MS/MS spectra were compared to the UniProt Homo Sapiens reference proteome database (UP000005640, release 2016-07, 92,578 protein sequences which contains canonical and variant isoforms). Peptide and MS/MS tolerances were set at 0.05 and 0.1 Da, respectively. The enzyme selectivity was set to full trypsin with one mis-cleavage allowed. Protein modifications were fixed carbamidomethylation of cysteines and variable oxidation of methionine. Mascot dat files obtained from each query were then imported into Proline Studio 1.3 software, used for the validation of identification results (http://proline.profiproteomics.fr/).

Each search result was thus validated with a peptide rank of 1, a minimal peptide length of 9 amino acids, a FDR of 1% on mascot e-value at the peptide spectrum match (PSM) level, and a FDR of 1% at the protein level using a FDR calculation methodology previously described.[16] The number of target and decoy values at PSM, peptide and protein levels together with respective FDR values were reported in supplementary Table 3. All validated Mascot search results obtained from each analyzed fraction were merged and a protein inference list created. Proteins identified with the exact same set of peptides or with a subset of the same peptides were grouped in a protein set. This protein set was represented by a typical protein corresponding to the protein identified with the best score or, for the same protein set, with the SwissProt accession number, if applicable.

The identified peptides from proteins of interest were analyzed with the neXtProt peptide uniqueness checker[18] to verify that they were unique to the inferred protein.

RNA-Seq analysis

The RNA-Seq data of isolated testicular cells (Leydig cells, peritubular cells, Sertoli cells, spermatocytes, and round spermatids) and total testis, were obtained from the SRA (https://www.ncbi.nlm.nih.gov/sra) with identification SRX1426397 to SRX1426408.[19] The fastq format collection of paired reads was obtained using a Galaxy (https://galaxyproject.org/) tool wrapper based on the fastq-dump utility of the SRA Toolkit.[20] The transcript abundance of ENSG00000277535 and ENSG00000214897 was estimated using Kallisto tool wrapper on the Galaxy interface[21] by alignment on the human genome (Ensembl release 92, GRCh38) and

expressed in "Transcripts Per Million" (TPM). The cut-off was set to 1 TPM. The TPM average is given as a histogram for each cell duplicate.

Datamining

Protein evidence (PE1-4) was retrieved from the neXtProt database (Release 2018-01-17). The uPE1 list was retrieved using the SPARLQ query of NXQ-00022 selecting only the validated PE1 proteins. Ensembl and UniProt IDs were retrieved using Retrieve/ID mapping tools on the UniProt website (http://www.uniprot.org/).

Our data set was compared to the neXtProt uPE1 list using a Venn diagram (J Ven V tool[22], http://jvenn.toulouse.inra.fr/app/index.html).

mRNA expression data in human tissues for 19,613 genes was obtained from the Human Protein Atlas database (HPA RNA gene data, Human Protein Atlas version 18 and Ensembl version 88.38).

Data on the levels of 13,205 proteins in human tissues, based on immunohistochemistry using tissue microarrays, were obtained from the HPA (Normal tissue data). The data are based on The Human Protein Atlas version 18 and Ensembl version 88.38. For each organ, the highest staining found in a cell type was assigned to the whole organ and global staining was then re-calculated per organ and expressed as 0 for not detected, 1 for low, 2 for medium, and 3 for high.

Heatmaps were generated using the Heatmapper tool[23] (http://www1.heatmapper.ca/). Row clustering was performed using the "complete linkage" method and the Euclidean method was used for distance measurement.

Polyclonal antibody production

Two specific peptides (Figs. 3 and 5) were designed for A0A0U1RQG5 and A0A0J9XYQ4. Protein sequences were first blasted against the human protein database to determine the homology region. Peptides were then chosen based on several criteria: no sequence homology, no post-translational modification, and low hydrophobicity. Peptide synthesis and production of rabbit polyclonal antibodies were performed by Biotem (Apprieu, France) using the "42 days" protocol. The antibodies were purified by Protein-A affinity chromatography.

Immunohistochemistry

The testicular expression of the two proteins of interest was confirmed by immunohistochemistry on human testes fixed in 4% paraformaldehyde and embedded in paraffin, as previously described.[15] Briefly, paraffin-embedded tissues were cut into 4 μm thick slices, mounted on slides, and dried at 58°C for 60 min. Immunohistochemical staining was performed on a Discovery Automated IHC stainer using the Ventana DABMap and OMNIMap detection kit (Ventana Medical Systems, Tucson, USA). Antigen retrieval was performed using the proprietary Ventana Tris-based buffer solution, CC1, at 95 to 100°C for 48 min. Tissue sections were then saturated for 1 h with 5% BSA in TBS and endogenous peroxidase was blocked with Inhibitor-D, 3% H_2O_2 (Ventana) for 8 min at 37°C. After rinsing in TBS, slides were incubated at 37°C for 60 min with polyclonal rabbit antibodies specific for the selected peptide diluted in TBS containing 0.2% Tween 20 (v/v) and 3% BSA (TBST-BSA). Nonimmune rabbit serum (1:1,000) was used as a negative control. After several washes in TBS, sections were incubated for 16 min with a biotinylated goat anti-rabbit antibody (Roche) at a final dilution of 1:500 in TBST-BSA. The signal was enhanced using the Ventana DABMap or OMNIMap kit. Sections were then counterstained for 16 min with hematoxylin (commercial

solution, Microm) and 4 min with bluing reagent (commercial solution, Microm), before rinsing with Milli-Q water. After removal from the instrument, slides were manually dehydrated and mounted in Eukitt (Labnord, Villeneuve d'Ascq, France). Finally, immunohistology images were obtained using NDP.Scan acquisition software (v2.5, Hamamatsu) and visualized with NDP.View2 software (Hamamatsu). Representative images are shown in figures 3 and 5.

Data Availability

The mass spectrometry proteomics data, including raw files and identification files, form a complete submission with the ProteomeXchange Consortium.[24] Data were submitted via the PRIDE partner repository under the dataset identifiers PXD009598 and 10.6019/PXD009598.

RESULTS AND DISCUSSION

Identification of "uncharacterized" dark proteins in the human testis

The aim of the present work was to study dark proteins in the testis with priority given to uncharacterized proteins. We assessed whether the testis is a relevant organ for identifying uncharacterized proteins by first analyzing the tissue expression of 520 mRNAs, corresponding to referenced uncharacterized proteins. Data were extracted for the 19,613 expressed genes in the HPA RNA database (Fig.1). Most of these uncharacterized genes/proteins were more highly expressed in the testis than in the 36 other human tissues analyzed. Fallopian tubes and, to a lesser extent, the cerebral cortex, also appear to be relevant tissues for studying uncharacterized proteins (Fig. 1). This observation strengthened our choice to look for uncharacterized proteins in the human testis.

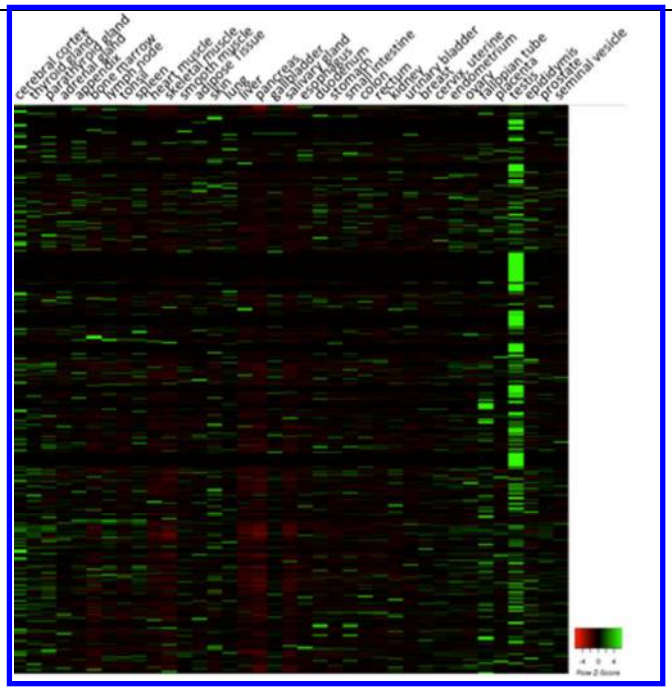


Figure 1: Tissue expression of uncharacterized genes in 37 human tissues. mRNA expression of 19,613 genes were obtained from the HPA database. Gene were filtered by names reviewed as uncharacterized. This heat map represents the clustering of gene expression of uncharacterized genes in 37 different human tissues. Interestingly, most of these uncharacterized genes are testis-enriched.

We generated a high-quality and high-resolution mass spectrometry dataset from a total human testis and were able to identify 6,897 protein groups with a 1% FDR threshold at the PSM and protein levels and at least one peptide with a minimal length of nine amino acids. The dataset was filtered by the number of peptides and UniProt description (Table 1 and Supplemental Table S1). Among the 5,578 protein groups we identified with at least two peptides, 30 were annotated as uncharacterized based on information from the 2016-07 UniProt release. Examination of the mRNA expression profile corresponding to these 30 uncharacterized dark proteins in the HPA database (Table 1) showed 15 to have a testis-enriched profile. Two candidates were selected for

Acc	Uniprot ID	Gene names	Description (UniProt release 2016_07)	Description (Uniprot release 2018-03-28)	PE (Nexprot release 2018-01-17)	Domain/ Family	mRNA espression / testis enriched	mRNA expression / testis specific
Q9ULL0	K1210_HUMAN	KIAA1210	Uncharacterized protein KIAA1210	Acrosomal protein KIAA1210	uPE1	DUF4592	Yes	No
B4E1Z4	B4E1Z4_HUMAN	-	Uncharacterized protein	cDNA FLJ55673, highly similar to Complement factor B (EC 3.4.21.47)	N/A	Peptidase S1	No	No
M0QYT0	M0QYT0_HUMAN	-	Uncharacterized protein (Fragment)	Uncharacterized protein (Fragment)	N/A	RNA recognition motif domain , Crotonase superfamily	No	
A0A0J9YXQ4	**PNMA6E_HUMAN /A0A0J9YXQ4_HUMAN**	**PNMA6E**	**Uncharacterized protein**	**Paraneoplastic antigen Ma6E**	**uPE1**	**PNMA family**	**Yes**	**No (expressed at low level in Ovary and Fallopian tube)**
F8W031	F8W031_HUMAN	-	Uncharacterized protein (Fragment)	Uncharacterized protein (Fragment)	N/A	DUF3456, Citrate_synth-like_lrg_a-sub, Saposin-like	No	
H3BN98	H3BN98_HUMAN	-	Uncharacterized protein (Fragment)	Uncharacterized protein (Fragment)	N/A	Ribosomal_S8	No	
G3V3G9	G3V3G9_HUMAN	-	Uncharacterized protein	Uncharacterized protein	N/A	WD_REPEATS_REGION	No	
E5RI56	E5RI56_HUMAN	-	Uncharacterized protein (Fragment)	Uncharacterized protein (Fragment)	N/A	S-phase kinase-associated protein 1-like	No	
Q8IYS4	CP071_HUMAN	C16orf71	Uncharacterized protein C16orf71	Uncharacterized protein C16orf71	uPE1	DUF4701	Yes	No
A0A0B4J203	A0A0B4J203_HUMAN	-	Uncharacterized protein	Uncharacterized protein	N/A	Protein kinase	No	
Q9BRQ4-2	CK070_HUMAN	C11orf70	Isoform 2 of Uncharacterized protein C11orf70	Uncharacterized protein C11orf70	uPE1	DUF4498	Yes	No
H7C1Q1	H7C1Q1_HUMAN	-	Uncharacterized protein (Fragment)	Uncharacterized protein (Fragment)	N/A	BC/rab GTPase-activating protein	No	
Q6ZU35	K1211_HUMAN	KIAA1211	Uncharacterized protein KIAA1211	Uncharacterized protein KIAA1211	uPE1	DUF4592	Yes	No
A1L188	NDUF8_HUMAN / CQ089_HUMAN	NDUFAF8 C17orf89	Uncharacterized protein C17orf89	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex assembly factor 8	PE1	CHCH	No	
Q9NZ63	TLS1_HUMAN / CI078_HUMAN	C9orf78 HCA59	Uncharacterized protein C9orf78	Telomere length and silencing protein 1 homolog (Hepatocellular carcinoma-associated antigen 59)	PE1	TLS1 family	No	
K7ESF4	K7ESF4_HUMAN	-	Uncharacterized protein (Fragment)	Uncharacterized protein (Fragment)	N/A	G6P_Isomerase	No	
O60268-3	K0513_HUMAN	KIAA0513	Isoform 3 of Uncharacterized protein KIAA0513	Uncharacterized protein KIAA0513	uPE1	-	No	
Q9H0B3-4	IQCN_HUMAN / K1683_HUMAN	IQCN KIAA1683	Isoform 4 of Uncharacterized protein KIAA1683	IQ domain-containing protein N	uPE1	Q domain-containing protein N	Yes	No
H7C0S8	H7C0S8_HUMAN	-	Uncharacterized protein (Fragment)	Uncharacterized protein (Fragment)	N/A	RPOL4c	No	
A0A0U1RQG5	**A0A0U1RQG5_HUMAN**	**-**	**Uncharacterized protein**	**Uncharacterized protein**	N/A	**Cancer/testis antigen 47 family**	**Yes**	**Yes**
Q7Z7L8	CK096_HUMAN	C11orf96 AG2	Uncharacterized protein C11orf96	Uncharacterized protein C11orf96 (Protein Ag2 homolog)	uPE1	DUF4695	Yes	No
Q9BV19	CA050_HUMAN	C1orf50	Uncharacterized protein C1orf50	Uncharacterized protein C1orf50	uPE1	DUF2452	Yes	No
A0A0G2JMZ2	A0A0G2JMZ2_HUMAN	-	Uncharacterized protein	Uncharacterized protein	N/A	Leucine-rich repeat-containing protein 37 family	-	-
Q86VG3	CK074_HUMAN	C11orf74	Uncharacterized protein C11orf74	Uncharacterized protein C11orf74 (Protein HEPIS)	uPE1	-	Yes	No
Q5JPI3-2	CC038_HUMAN	C3orf38	Isoform 2 of Uncharacterized protein	Uncharacterized protein C3orf38	PE1	NTF2-like_dom_sf	Yes	No
Q5TEA3	CT194_HUMAN	C20orf194	Uncharacterized protein C20orf194	Uncharacterized protein C20orf194	PE1	-	Yes	No
Q5U649	CL060_HUMAN	C12orf60	Uncharacterized protein C12orf60	Uncharacterized protein C12orf60	uPE1	DUF4533	Yes	No
Q8IXQ3	CI040_HUMAN	C9orf40	Uncharacterized protein	Uncharacterized protein C9orf40	uPE1	-	Yes	No
Q9H8K7	CJ088_HUMAN	C10orf88	Uncharacterized protein C10orf88	Uncharacterized protein C10orf88	uPE1	DUF4506; ABC_TRANSPORTER_1	Yes	No
Q8TB03	CX038_HUMAN	CXorf38	Uncharacterized protein CXorf38	Uncharacterized protein CXorf38	uPE1	DUF4559	No	

Table 1: List of the uncharacterized proteins identified in the human testis. Dark proteins annotated as uncharacterized in the UniProt release 2016-07 are provided together with the corresponding gene name. For each entry, the protein updated description in the UniProt release 2018-03, evidence level (neXtProt release 2018-01-17, N/A not applicable), domain/family (UniProt), and mRNA level expression in human testis (Expression Atlas) were provided. According to UniProt release 2016-07, 30 identified proteins were described as uncharacterized proteins. Two of these were chosen according to their expression pattern in reproductive organs: A0A0U1RQG5 and PNMA6E (A0A0J9YXQ4) which are testis-enriched proteins. A0A0U1RQG5 is testis-specific, whereas PNMA6E is also expressed at low level in the ovary and Fallopian tube.

further investigation: i) A0A0U1RQG5, for which the mRNA displayed a clear testis-specific expression profile and ii) A0A0J9XYQ4, for which the mRNA displayed an expression profile with a potential link to reproductive tissues. Indeed, the gene corresponding to A0A0J9XYQ4 was highly expressed in the testis, but also showed low expression in the female genital tract.

The status of the 30 uncharacterized proteins was updated in the last neXtProt release (2018-01-17). Eighteen were shown to exist at the protein level (PE1), of which 14 have an unknown function (uPE1) and are thus awaiting further characterization. Twenty-four are still annotated as "uncharacterized" (Table 1). A0A0U1RQG5 was not reviewed in neXtProt, whereas A0A0J9XYQ4 is evidenced at the protein level, known as PNMA6E, but with an unknown function. We further characterized these two dark proteins by designing immunogenic peptides for the production of specific polyclonal antibodies for immunohistochemistry studies of human testis sections.

A0A0U1RQG5, a novel cancer/testis antigen, is testis-specific and expressed in the germ cell lineage

A0A0U1RQG5 is a transcript of the AL772284.2 gene (Fig. 2A and B; Ensembl version 92) composed of three exons and located on chromosome Xq21.1 (position 119,073,226 to 119,076,373) on the forward strand. The transcript is 1,147 bps in length and encodes a protein of 324 aa (A0AOU1RQG5). No splice variant is referenced in the Ensembl database. The annotated transcript contains two "cancer/testis antigen 47" domains, suggesting that the A0A0U1RQG5 protein may be a cancer/testis antigen belonging to family 47. Furthermore, A0A0U1RQG5 shares 61.4% homology with CT47B and 60.3% with CT47A (data not show).

In our study, A0A0U1RQG5 was identified by two unique non-nested peptides with a length of 13 and 10 aa (Fig. 2C). According to the Expression Atlas data from several laboratories (Fig. 3A), the AL772284.2 gene transcript is only expressed in the adult testis. Data on the level of mRNA expression of the AL772284.2 gene was obtained from recently published RNA-Seq data on isolated human testicular cells, expressed as transcripts per million (TPM; Fig. 3B). AL772284.2 is expressed in spermatocytes and round spermatids. We produced a polyclonal antibody using two specific peptides for A0A0U1RQG5 (Fig. 2C). The peptides were blasted against human entries (UniProtKB Human database) to ensure that they did not share homology with other human proteins. The purified polyclonal antibody was used for immunohistochemistry studies of human testis sections (Fig. 3C). There was moderate immunoreactivity in the seminiferous epithelium at all stages of spermatogenesis, whereas the cytoplasmic lobes of elongated spermatids showed strong staining. In the interstitium, there was moderate staining of Leydig cells, a classical feature

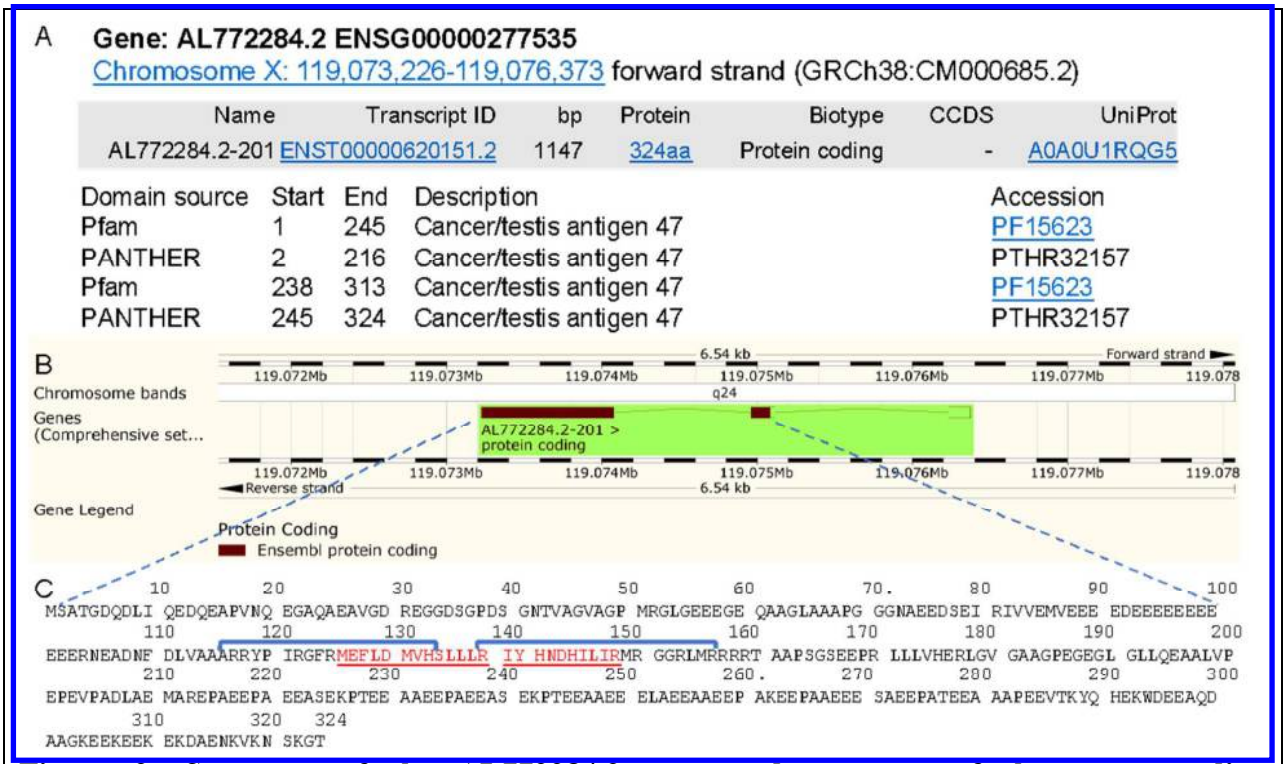


Figure 2: Structure of the AL772284.2 gene and sequence of the corresponding A0A0U1RQG5 protein. The AL772284.2 gene, identified in the human testis, is located on the forward strand on the X chromosome (A and B). The corresponding protein A0A0U1RQG5 contains CT47 domains (A). A0A0U1RQG5 was identified with 2 peptides of respectively 13 and 10 aa (C). The two immunogenic peptides used for antibody production were designed in regions that share 44% (peptide 1) and 63% (peptide 2) homology with CT47B and CT47A; 2 proteins which present the highest homology with A0A0U1RQG5 (C).

of the immunohistochemistry of testis sections, often corresponding to non-specific staining.

Our results show that expression of the AL772284.2 gene is testis-specific and the corresponding protein, A0A0U1RQG5, is germ cell-specific. A0A0U1RQG5 contains CT47 domains and is probably a novel cancer testis antigen (CTA). CTAs are overexpressed in tumors but restricted to immune privileged sites in normal tissues. CTA are classed into three groups: 1) specific to the

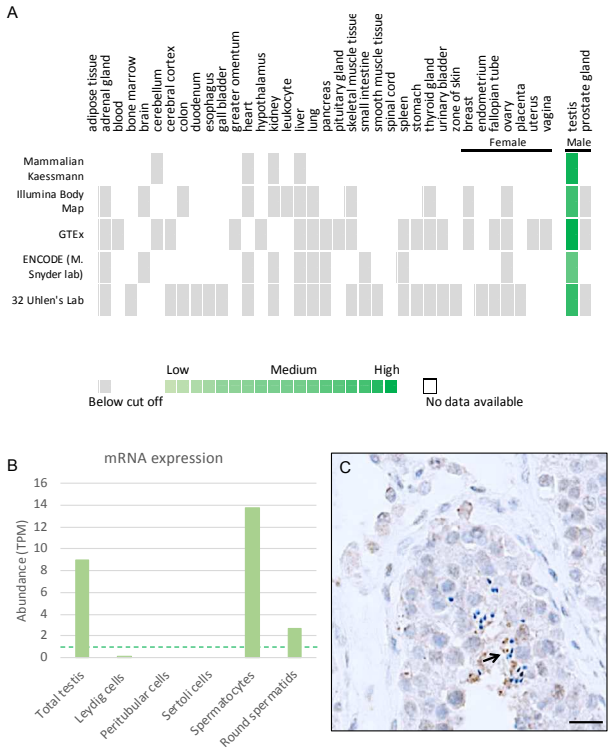


Figure 3: Expression of AL772284.2 gene in human tissues. (A) mRNA expression pattern in various human tissues. (B) mRNA level is expressed in TPM and the cut-off is set at 1 TPM (green dotted line). (C) Antibody staining validation in adult human testis. According to Expression Atlas, AL772284.2 mRNA expression was specifically found in the testis (A). In the testis, mRNA expression of the AL772284.2 gene was found in spermatocytes and round spermatids (B). The corresponding protein was detected by immunohistochemistry in transverse testis sections at stages IV to VI of the seminiferous epithelium[25] using a rabbit polyclonal antiserum designed against the 2 specific peptides (see Materials & Methods section and Fig. 2C). Non-immune serum was used as a negative control (data not shown). In the interstitium, a moderate staining was observed in cytoplasm of Leydig cells. Staining was clearly cytoplasmic and moderate in germ cells at all stages of their development. A very strong immunostaining was observed in the cytoplasm of late spermatids (C; arrow). Scale bars = 20 µm.

testis, 2) testis- and brain-specific, and 3) testis-specific and expressed at low levels in no more than two organs.[26] Our candidate, A0A0U1RQG5, is testis-specific, with a gene located on the X chromosome, similarly to half of all CTAs, such as MAGEs, leading to their being named CT-X genes. CT-Xs are more testis-restricted than other CTAs.[26] CTAs have been widely studied for their potential to target cancer cells in personalized immunotherapy (for review see[27, 28]). Recently, Babatunde and collaborators[29] reviewed the potential roles of CTAs in testis function. Most of the functions of CTAs functions remain unclear, but some appear to be involved in gametogenesis, sperm metabolism, or motility.[29]

PNMA6E is enriched in the human testis and is expressed during spermatogenesis

The second selected protein is paraneoplastic Ma antigen 6E (PNMA6E), which is testis-enriched (Table 1). The PNMA6E gene is located on chromosome Xq28 (position 153,395,640 to 153,401,420) on the reverse strand and has two splice variants: PNMA6E-201 and PNMA6E-202 (Fig. 4A and B). PNMA6E-201 was identified in our dataset by 11 non-nested peptides with a minimal length of nine aa. Nine are unique to the PNMA6E gene (two are shared with PNMA6F) and four are specific to the PNMA6E-201 isoform (Fig. 4C). The transcript is 2,192 bp in length and contains PNMA domains (Fig. 4A).

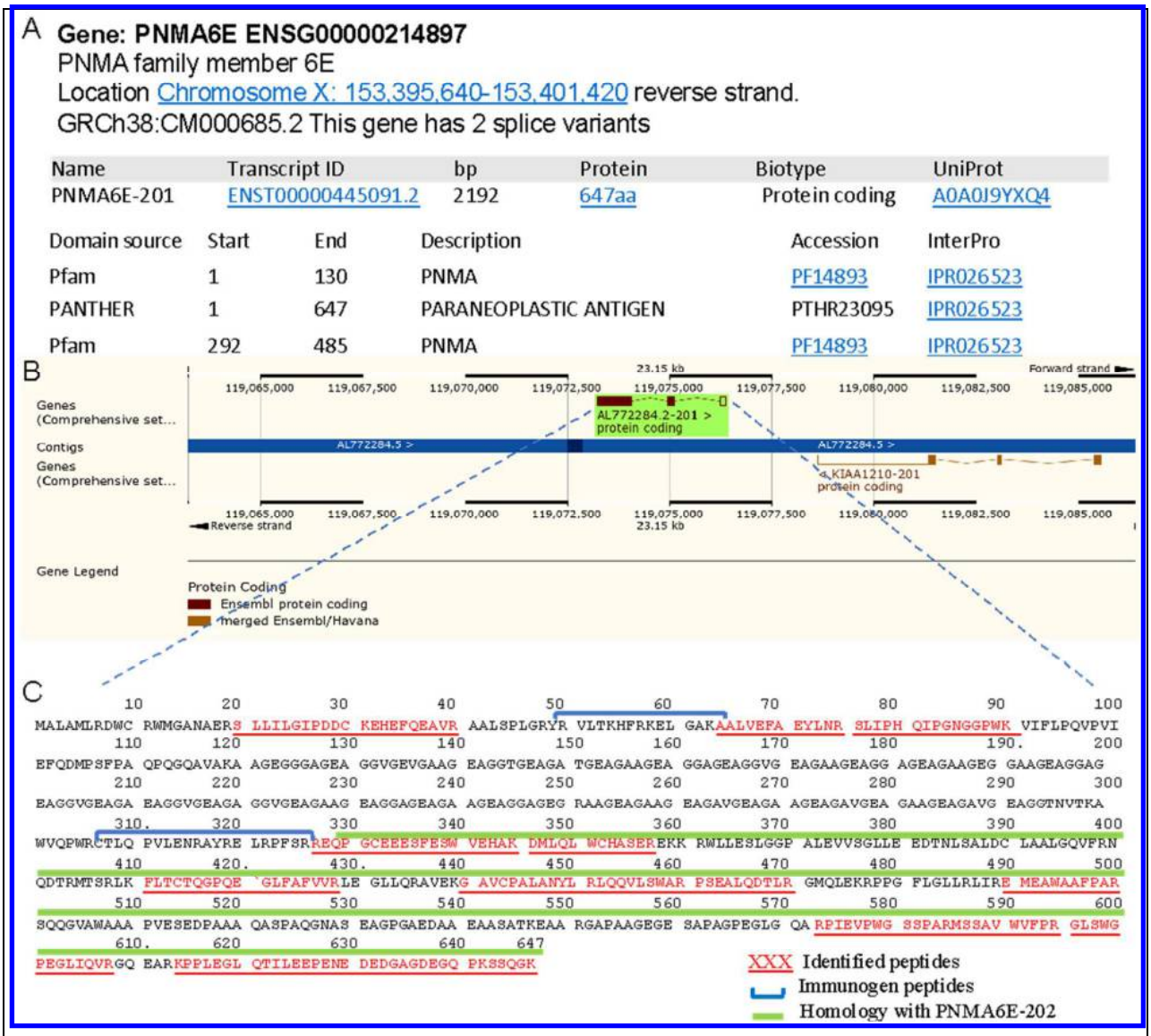


Figure 4: Structure of the PNMA6E gene and sequence of the corresponding protein. The gene PNMA6E identified in the human testis is located on the reverse strand of the X chromosome (genome reference GRCh38) (A). This gene has two splice variants, PNMA6E-202 and PNMA6E-201 (B). The second splice variant was identified in our dataset by11 peptides, of which four are specific (C). This protein contains a PNMA domain (A). The PNMA6E-201 and PNMA6E-202 variants share 317 amino acids (*i.e.,* 100% homology) (C). The two immunogenic peptides used for antibody production were designed within the specific region (1-329) of PNMA6E-201 (C).

The resulting protein, PNMA6E, has a length of 647 aa and a sequence of 322 aa in the C terminal part (325-647) shares 100% identity with PNMA6E-202 (Fig. 4C). Public expression data from the Expression Atlas shows that PNMA6E is highly expressed in the human testis (Fig. 5A) and present at low levels in the ovary and Fallopian tubes. The PNMA family is composed of 15 members, which are predominantly expressed in the brain (Fig. 6), whereas PNMA6E is only expressed in testis and, at a low level, in female organs (ovary and Fallopian tubes; Figs. 5, 6). We further investigated the expression of the PNMA6E gene using RNA-Seq data from isolated testicular cells. Its expression is restricted to germ cells (Fig. 5B). Furthermore, we performed immunohistochemistry experiments on human testis sections using a polyclonal antibody directed against two specific peptides of the PNMA6E protein (see Material and Methods section; Fig. 4C). PNMA6E showed nuclear staining (Fig. 5) in both cells of the interstitium and the seminiferous tubules. In the seminiferous epithelium, germ cells increasingly expressed this protein from premeiotic germ cells (*i.e.,* spermatogonia and primary spermatocytes) to pachytene spermatocytes and round spermatids. However, PNMA6E was no longer found in elongated spermatids. Leydig cells do not express the PNMA6E mRNA (Fig. 5B). Thus, the moderate immunostaining observed in Leydig cells probably corresponds to non-specific staining.

PNMA family members share several domains, with high sequence homology, and are quite conserved among species[30]. Particularly, PNMA6E shares 64% homology with its mouse orthologue. Although PNMA members share conserved domains, their expression patterns differ.[31] In contrast to other members of the PNMA family, PNMA6E expression is clearly restricted to the testis and, to a lower extent, to the female genital tract (Figs. 5A and 6).

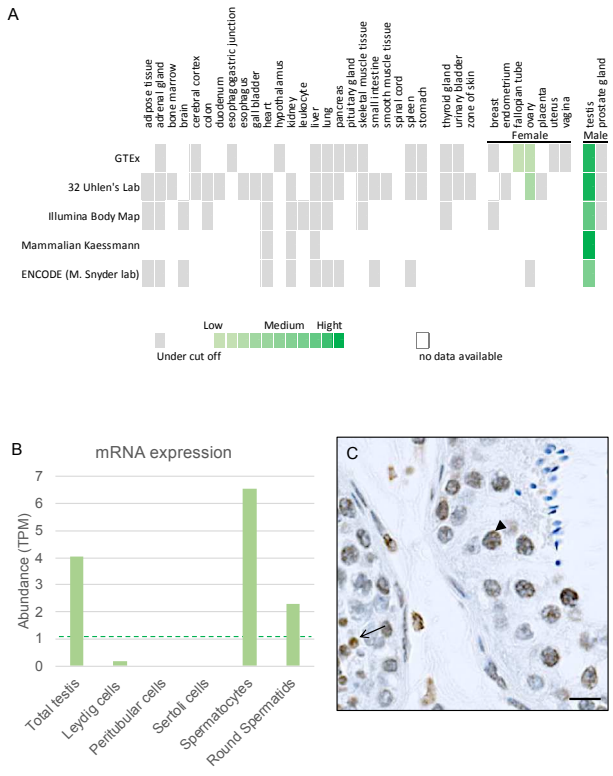


Figure 5: Expression of PNMA6E in human tissues. (**A) mRNA expression pattern in various human tissues.** According to Expression Atlas, PNMA6E mRNA expression is only found at a high level in the testis, whereas the ovaries and fallopian tubes also display positive, but low, expression. **(B) mRNA levels in different cell types of the testis.** The mRNA level is expressed in TPM and the cut-off is set at 1 TPM (green dotted line). In the testis, mRNA was detected only in male germ cells. **(C) Antibody staining validation of the PNMA6E protein in adult human testis**. The corresponding protein was detected in transverse testis sections at stages IV to VI of the seminiferous epithelium[25] using rabbit polyclonal antiserum designed against two specific peptides. Non-immune serum was used as a negative control (data not shown). In the interstitium, dense immunoreactivity was observed in the nucleus of Leydig cells. A significant number of peritubular cell nuclei also displayed a moderate signal. Immunoreactivity was more intense in germ cell nuclei, in which the signal increased from premeiotic germ cells to pachytene spermatocytes (arrowhead) and round spermatids (arrows), whereas there was no staining of elongated spermatids.

PNMAs also differ in their association with disease (for review see[31]): aberrant expression of PNMA1, 2, and 3 are correlated with paraneoplastic disorders or cancer[32], whereas MOAPS mediates apoptotic signaling.[31] Recently, mouse pnma5 was shown to be involved in meiotic progression in oocytes by controlling the phosphorylation of two important kinases, PDK1 and Gsk3beta.[33] The authors also demonstrated that pnma5 knockdown mice exhibit abnormal oocyte fertilization.

In conclusion, PNMA6E expression is restricted to the genital tract. The protein is highly enriched in the testis, particularly in germ cells up to meiosis. The expression of PNMA6E in only the human testis, and to a lower extent in the ovary and Fallopian tubes, favors a role in gametogenesis, similar to PNMA5, which is considered to be a female fertility factor[33].However, this awaits further investigation.

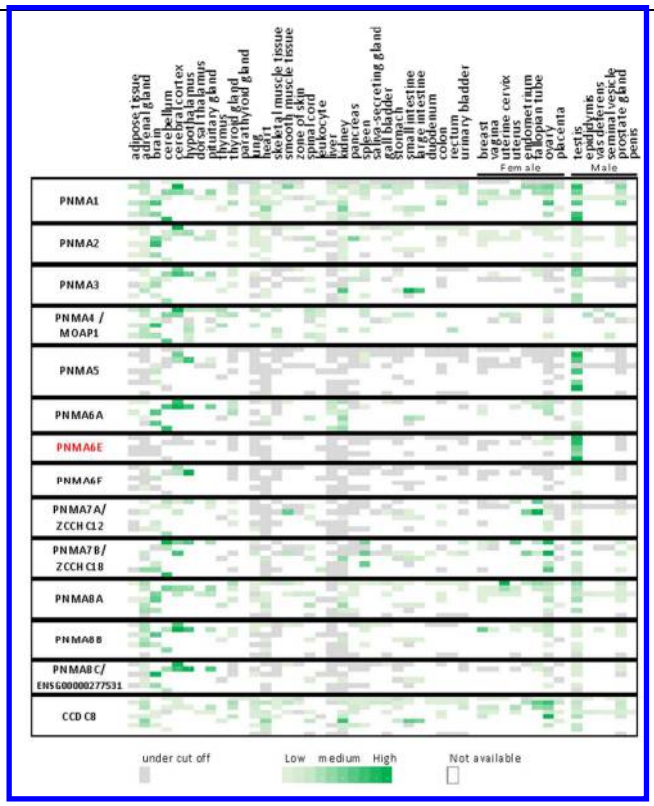


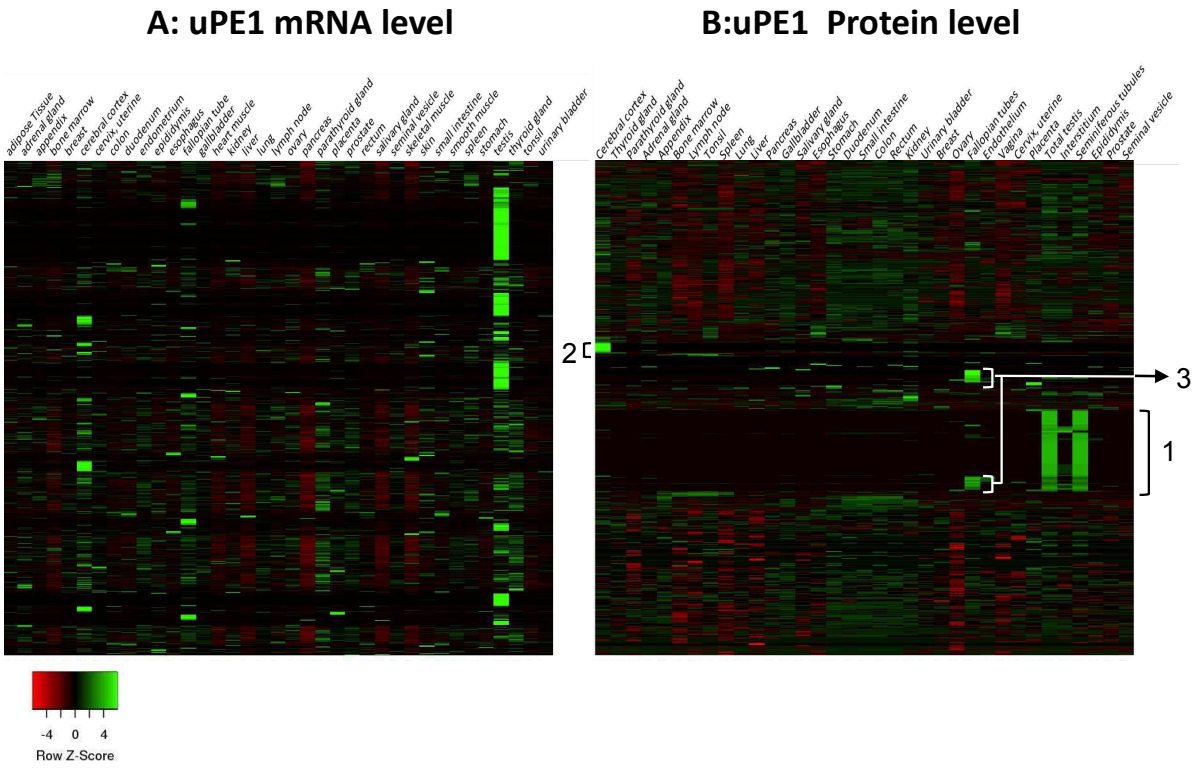
Figure 6: Human tissue expression of the PNMA family gene members. mRNA expression data were obtained from the Expression Atlas. Expression of 14 of the known members (PNMA6B is a pseudogene not shown here) of the PNMA family is shown in various human tissues, including that from the female and male reproductive tracts. All PNMA genes, except PNMA6E, are expressed in brain tissues. PNMA1, PNMA5, and PNMA6E display a testis-enriched pattern, whereas PNMA6E is exclusively expressed in the testis and, to a lower extent, in the ovary and Fallopian tube.

uPE1 and MPs are highly expressed in the testis

In the last part of our study, we investigated the expression pattern of other dark proteins: uPE1 and missing proteins (PE2-4) (Figs. 7 and 8). We examined the mRNA expression of the 19,613 genes and the level of the 13,205 proteins from the HPA database. The status of the evidence for the existence of the protein was assigned to each entry (neXtProt release 2018-01-17). Among the 19,613 genes, 1,229 correspond to uPE1 and 1,822 correspond to MPs (PE2-4), others are PE1, PE5 or unreferenced in neXtProt. Among the 13,205 proteins found in HPA, 754 are considered to be uPE1 whereas 424 are MPs. We found that uPE1 genes were highly expressed in the testis using a clustering method (Fig. 7A). At the protein level, many uPE1 also showed high expression levels in the testis (Fig. 7B). Unsurprisingly, most were found at high levels in seminiferous tubules, suggesting that uPE1 are mainly expressed in the germ lineage.

Two other significant enriched expression clusters could be observed in the cerebral cortex and Fallopian tubes. Finally, we sought uPE1 identified in our testis dataset (Fig. 7C). Among the 1,260 uPE1 referenced in the latest neXtProt release, 238 were identified in the present study, corresponding to approximately 19% of the current uPE1 in neXtProt.

We used a clustering method to also compare the 1,822 MP genes found in the HPA database (Figs. 8A, B and C). Most genes corresponding to MPs were testis-enriched, similar to uPE1 genes. At the protein level, the 424 MPs found in the HPA database also displayed a testis-enriched pattern and were, to a large extent, highly expressed in seminiferous tubules, similar to uPE1, and thus most likely expressed in germ cells. Some MPs were also enriched in the cerebral cortex, placenta, and epididymis.


A: uPE1 mRNA level
B:uPE1 Protein level
adipose Tissue
adrenal gland
appendix
bone marrow
breast
cerebral cortex
cervix, uterine
colon
duodenum
endometrium
epididymis
esophagus
fallopian tube
gallbladder
heart muscle
kidney
liver
lung
lymph node
ovary
pancreas
parathyroid gland
placenta
prostate
rectum
salivary gland
seminal vesicle
skeletal muscle
skin
small intestine
smooth muscle
spleen
stomach
testis
thyroid gland
tonsil
urinary bladder
Cerebral cortex
Thyroid gland
Parathyroid gland
Adrenal gland
Appendix
Bone marrow
Lymph node
Tonsil
Spleen
Lung
Liver
Pancreas
Gallbladder
Salivary gland
Esophagus
Stomach
Duodenum
Small intestine
Colon
Rectum
Kidney
Urinary bladder
Breast
Ovary
Fallopian tubes
Endothelium
Vagina
Cervix, uterine
Placenta
Total testis
Interstitium
Seminiferous tubules
Epididymis
Prostate
Seminal vesicle
2
3
1
-4
0
4
Row Z-Score

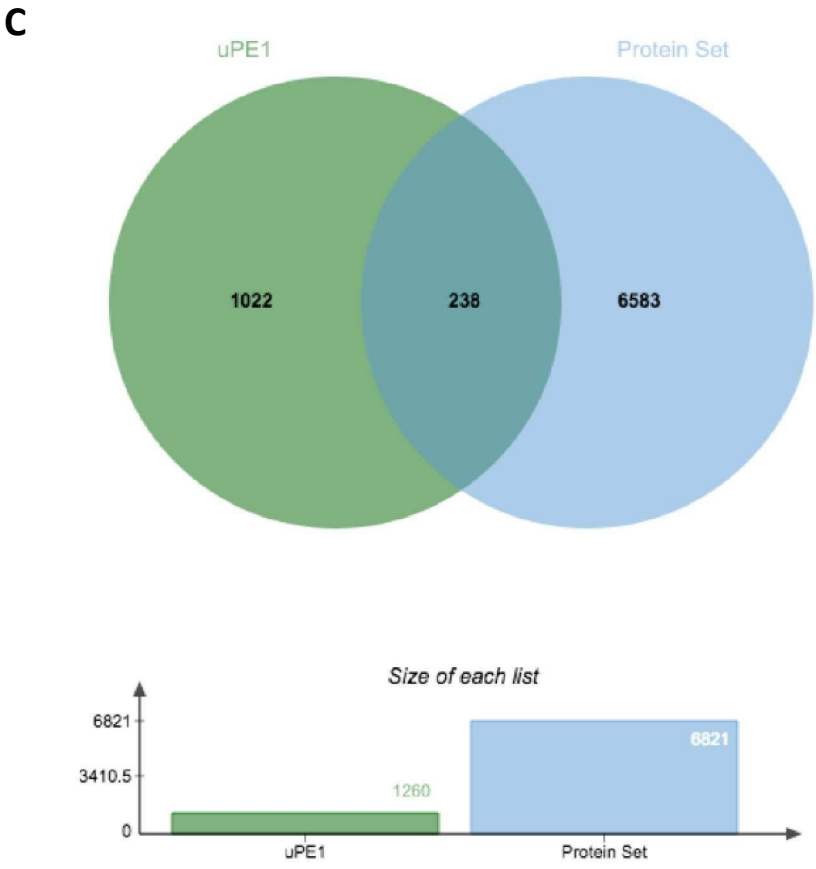

C
uPE1
Protein Set
1022
238
6583
Size of each list
6821
3410.5
0
1260
6821
uPE1
Protein Set

Figure 7: Testicular enrichment of uPE1 proteins in humans. Data on the (A) mRNA expression of 19,613 genes and (B) the staining pattern of 13,205 proteins were obtained from the HPA database. (C) Our data set (Prot set) was compared to the uPE1 list from neXtProt (release 2018-01-17) using a Venn diagram. The protein evidence status was assigned to each entry (neXtProt release 2018-01-17). (A) Heat map representing the clustering of gene expression of corresponding uPE1 proteins in 37 human tissues. Many of the uPE1 genes are testis enriched. (B) Among the 13,205 proteins found in the HPA database, 754 are uPE1 proteins. Clustering showed a testis enriched cluster (1). These uPE1 were enriched in the testis and most were also found at high levels in the seminiferous tubules. Other significantly enriched clusters were observed in the cerebral cortex (2) and Fallopian tubes (3). (C) 238 proteins identified in the human testis are uPE1, corresponding to approximately 19% of the 1,260 uPE1 referenced in neXtProt.

Our results clearly show the great potential of the human testis, not only to search for additional missing proteins, but also to characterize the function of uPE1 in a biological context. The dark proteins studied here are, to a large extent, testis-enriched and their expression could be traced to the seminiferous tubules. This confirms previous studies showing that most testis-enriched genes are expressed by germ cells.[12] Although seminiferous tubules account for 90% of the total testis mass, approximately two-thirds of the cells in these tubules are haploid spermatids at various stages of development. This suggests that testis-enriched uPE1 and MPs are primarily expressed in germ cells and thus play a role in spermatogenesis. Other interesting clusters for characterization of dark proteins are indeed the cerebral cortex, the Fallopian tubes, the placenta,

and the epididymis (Figs. 7 and 8), suggesting their involvement in biological mechanisms specific to these organs. In

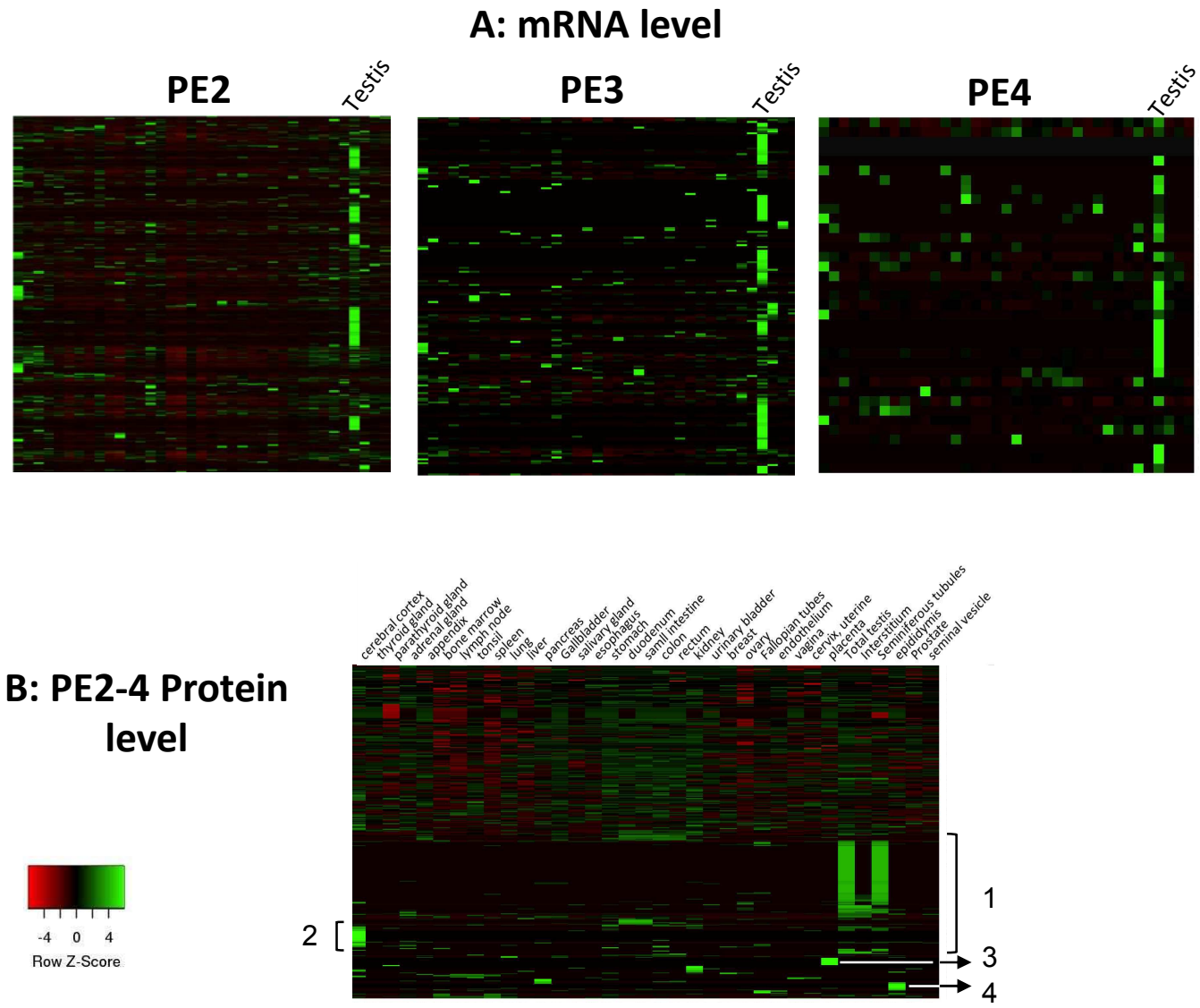


Figure 8: Updated status of missing proteins in the human testis. Data on the (A) mRNA expression levels of 19,613 genes and (B) staining pattern of 13,205 proteins were obtained from the HPA database. The protein evidence status was assigned to each entry (neXtProt release 2018-01-17). (A) Heat maps representing the clustering of gene expression corresponding to missing proteins (PE2-4) in 37 human tissues. Many of the MP genes are testis-enriched. (B) Among the 13,205 proteins found in the HPA database, 424 are MPs (PE2-4). Clustering showed

a testis-enriched cluster (1). These MPs were enriched in the testis and most were also found at high levels in the seminiferous tubules. Other significantly enriched clusters were observed in the cerebral cortex (2), placenta (3), and epididymis (4).

such a context, future studies in our lab will involve collecting information about the proteins and genes from the clusters highlighted in this study to associate them with organ-related functions.

CONCLUSION

Here, we investigated the global expression of dark proteins in the human testis, focusing on uncharacterized proteins, as well as uPE1 and MPs. We were unable to find evidence of new missing proteins for validation in our dataset according to the latest C-HPP Guidelines. Only nine missing proteins were potentially identified by single peptide that pass through the automatic validation, among which 5 MPs were evidenced with one unique peptide (supplementary Table 2), suggesting that the dynamic range in a total testis extract is still too large to access low-copy number proteins by conventional pre-fractionation and LC-MS/MS analysis. Although, we have clearly reached the limit of detection of shotgun mass spectrometry, our results confirm that the testis is a promising organ to search for additional MPs, as previously suggested[8, 10, 11]. Monitoring the expression of genes corresponding to missing proteins in the whole testis is indeed a relevant prerequisite to search for MPs. However, such an approach could be misleading, because even though MP genes may be highly expressed in the testis, expression of the corresponding proteins could vary or they could be expressed in particular subcellular locations that render them difficult to access and identify. In such a context, the use of a targeted-MS approach was shown to be efficient in the search for selected testis-enriched MPs, as carried out by Carapito and collaborators on ejaculated spermatozoa[17]. Equalization of

protein abundance using the Proteominer approach may also be relevant for the identification of low expressed MPs[34], particularly for the identification of membrane or nuclear MPs. The use of RNA-Seq information on MP gene expression will be an important asset to design proteotypic peptides for selected protein candidates prior to targeted-MS analyses. The search for MPs in isolated testicular cell extracts and/or subcellular compartments has indeed been envisioned but is still technically challenging due to limited routine access to human testes.

The C-HPP consortium also focuses on the characterization of uPE1 and, more widely, dark proteins. According to the Protein Structure group, the dark proteome consists of proteins for which the experimental structure has never been determined and which are inaccessible to homology modeling. More than half of dark proteins are not intrinsic disorder proteins, have no transmembrane domain, and have a low compositional bias, as shown by Perdigão *et al.*[35, 36]. These authors have recently created the Dark Proteome Database DPD (http://darkproteome.ws) to better understand and characterize the Dark Proteome. The database presently provides a general map for the dark proteome but needs to be further implemented with additional annotation and prediction sources. Undoubtedly, such a database will be important for aiding the characterization of dark proteins, regardless of the definition used.

Determining the biological functions of dark proteins in the human testis is a very long-term project. Although genome editing may open new perspectives in the coming years[37], the CRISPR/Cas9 system is far from being routine and allowing large-scale functional studies of spermatogenesis.[38] A more realistic approach will be to preferentially select candidate proteins based on their expression pattern. Thus, priority should be given to dark proteins specifically expressed during meiosis and spermiogenesis, two unique processes of the testis. As an example,

the PNMA6E expression pattern observed here suggests that the protein may play an important role during meiosis up to early spermiogenesis. Late spermatid- and spermatozoa-specific proteins may also play a role in sperm biology (*e.g.*, mobility, interaction with the oocyte). In both cases, gene invalidation studies could be considered. The selection criteria will be the existence of an orthologous gene in the mouse, with no paralogs, and coordination with the International Mouse Phenotyping Consortium (http://www.mousephenotype.org) to accelerate the production of knockout mice on the orthologous gene of interest. However, demonstration of a link between a defect in the expression of a specific protein and/or a defect with its post-translational modification and human infertility will take several years. Coordination with clinicians of Assisted Reproductive Technologies will also be necessary to determine the most relevant sperm pathologies, based on observed phenotypes, and then the collection of the cohort of samples. Coutton *et al.*[39] succeeded in using *Trypanosoma brucei* gene invalidation to rapidly study the function of proteins suspected to be important for sperm flagella. The authors demonstrated that the use of such an original model, which shares an extremely conserved flagellar structure with mammals, for gene invalidation could induce severe flagella defects. Indeed, a comparable approach could be envisioned for a selected set of testicular dark proteins.

Here, we demonstrated that uPE1, MPs, and other uncharacterized proteins are, to a large extent, highly expressed in the testis and, more precisely, within the seminiferous tubules for uPE1 and MPs, even if we could only identify 30 uncharacterized proteins. Nevertheless, we identified 238 uPE1, corresponding to 19% of the uPE1 referenced in neXtProt. Our results clearly show that the testis is an organ of choice, not only for the identification of MPs, but also for the study and characterization of uPE1 and other uncharacterized proteins. Interestingly, in addition to somatic protein isoforms and due to the presence of the germ cell lineage, the testis is

also suspected to express an array of germinal isoforms, a significant number of which being produced through alternative splicing. It might be wise to also explore that domain. Only a few laboratories worldwide work on testicular function and this is probably why many dark proteins, expressed in the testis, are still uncharacterized or missing. In addition, the testis is a complex organ with an array of very specific functions and unique phenomena (*e.g.,* chromatin remodeling, repackaging, and transcriptional reprogramming in haploid germ cells) which involve specific genes and proteins[9, 10]. We hope that the presented data will significantly aid the further characterization of dark proteins. Finally, thorough characterization of dark proteins in the testis will contribute to a better understanding of the molecular mechanisms underlying spermatogenesis and should yield striking discoveries in the context of male infertility.

Supporting Information

Table S1: List of the uncharacterized proteins identified in the human testis (xlsx).

Table S2: Information on the 9 Missing Protein identified with single peptide (xlsx).

Table S3: Values for FDR calculation (xlsx)

Mass Spectrometry proteomics data: PXD009598 (ProteomeXchange Consortium).

AUTHOR INFORMATION

Corresponding Author

*Charles Pineau: **charles.pineau@inserm.fr**, Tel: +33 (0)2 2323 5279

ORCID number: 0000-0002-7461-5433

Author Contributions

CP, EC, and NMel co-coordinated the study. JVC provided the testis sample. NMel, EC, and CP conceived and designed the experiments and analyses. JVC, ML, RL, and BG performed the sample preparation. NM and JD performed the sample fractionation and MS/MS analysis. ML, EC, and NMel processed and analyzed the MS/MS data sets. NMel, EC, and LG performed the bioinformatics analysis and data/literature mining of the identified proteins and selected candidates. NM designed the immunogen peptides and immunohistochemical studies. The IHC experiments were performed by PB. NMel, EC and CP prepared the figures, tables, and Supporting Information. NMel and CP drafted the manuscript.

Notes

The authors declare no competing financial interest.

Funding Sources

This work was partially funded through the Fondation pour la Recherche Médicale (FRM Grant DBS20140930778). This work was also supported by grants from Biogenouest and the Conseil Régional de Bretagne, awarded to CP.

ACKNOWLEDGMENTS

We are grateful to Gauthier Husson (Université de Strasbourg, CNRS, IPHC UMR 7178, LSMBO, Strasbourg, France) for recalibration of the raw MS/MS data. We also thank the GenOuest bioinformatics Core Facility (UMR6074 IRISA CNRS/INRIA/Université de Rennes1) for access to the Galaxy web platform at galaxy.genouest.org used for the RNA-Seq data analysis.

REFERENCES

(1) Weinbauer, G. F.; Gromoll, J.; Simoni, M.; Nieschlag, E. Physiology of Testicular Function. *Andrology* : male reproductive health and dysfunction. E. Nieschlag, H.M. Behre (Eds.) **1997**, 25-57

(2) Eddy, E. M. Male Germ Cell Gene Expression. *Recent Progress in Hormone Research* **2002**, *57* (1), 103–128.

(3) Jégou, B.; Pineau, C.; Dupaix, A. Paracrine Control of Testis Function. *Male Reproductive Function Mineralogical Society Series* **1999**, 41–64.

(4) Baker, M. A.; Nixon, B.; Naumovski, N.; Aitken, R. J. Proteomic insights into the maturation and capacitation of mammalian spermatozoa. *Systems Biology in Reproductive Medicine* **2012**, 58 (4), 211–217.

(5) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C.; Corthals, G. L.; Costello, C. E.; Deutsch, E.W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G.H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlén, M.; Wu, C.H.; Yamamoto, T.; Paik, Y.K.; Omenn, G.S. The human proteome project: Current state and future direction. *Molecular and Cellular Proteomics* **2011**, 10 (7), M111.009993.

(6) Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Research* 2017, *45* (D1), D177-D182.

(7) Na, K.; Shin, H.; Cho, J.-Y.; Jung, S. H.; Lim, J.; Lim, J.-S.; Kim, E. A.; Kim, H. S.; Kang, A. R.; Kim, J. H.; Shin, J.M.; Jeong, S.K.; Kim, C.Y.; Park, J.Y.; Chung, H.M.; Omenn, G.S.; Hancock, W.S.; Paik, Y.K. Systematic Proteogenomic Approach To Exploring a Novel Function for NHERF1 in Human Reproductive Disorder: Lessons for Exploring Missing Proteins. **2017**, 16 (12), 4455–4467.

(8) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins. **2014**, 13 (1), 15–20.

(9) Son, C. G.; Bilke, S.; Davis, S.; Greer, B.T.; Wei, J.S.; Whiteford, C.C.; Chen, Q.R.; Cenacchi, N.; Khan, J. Database of mRNA gene expression profiles of multiple human organs. *Genome Research* **2005**, 15 (3), 443–450.

(10) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjöstedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C.A.; Odeberg, J.; Djureinovic, D.; Takanen, J.O.; Hober, S.; Alm, T.; Edqvist, P.H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J.M.; Hamsten, M.; von Feilitzen, K.; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; von Heijne, G.; Nielsen, J.; Pontén, F. Tissue-based map of the human proteome. *Science* **2015**, 347 (6220), 1260419–1260427.

(11) Uhlén, M.; Hallström, B. M.; Lindskog, C.; Mardinoglu, A.; Pontén, F.; Nielsen, J. Transcriptomics resources of human tissues and organs. *Molecular Systems Biology* 2016, *12* (4), 862–873.

(12) 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. *Molecular Human Reproduction* **2014**, 20 (6), 476–488

(13) Wei, W.; Luo, W.; Wu, F.; Peng, X.; Zhang, Y.; Zhang, M.; Zhao, Y.; Su, N.; Qi, Y.; Chen, L.; Zhang, Y.; Wen, B.; He, F.; Xu, P. Deep Coverage Proteomics Identifies More Low-Abundance Missing Proteins in Human Testis Tissue with Q-Exactive HF Mass Spectrometer. **2016**, 15 (11), 3988–3997.

(14) Wang, Y.; Chen, Y.; Zhang, Y.; Wei, W.; Li, Y.; Zhang, T.; He, F.; Gao, Y.; Xu, P. Multi-Protease Strategy Identifies Three PE2 Missing Proteins in Human Testis Tissue. *Journal of Proteome Research* 2017, 16 (12), 4352–4363

(15) Jumeau, F.; Com, E.; Lane, L.; Duek, P.; Lagarrigue, M.; Lavigne, R.; Guillot, L.; Rondel, K.; Gateau, A.; Melaine, N.; Guével, B.; Sergeant, N.; Mitchell, V.; Pineau, C. Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *Journal of Proteome Research* **2015**, 14 (9), 3606–3620.

(16) Vandenbrouck, Y.; Lane, L.; Carapito, C.; Duek, P.; Rondel, K.; Bruley, C.; Macron, C.; Peredo, A. G. D.; Couté, Y.; Chaoui, K.; Com, E.; Gateau, A.; Hesse A.M.; Marcellin, M.; Méar, L.; Mouton-Barbosa, E.; Robin, T.; Burlet-Schiltz, O.; Cianferani, S.; Ferro, M.; Fréour, T.; Lindskog, C.; Garin, J.; Pineau, C. Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. *Journal of Proteome Research* **2016**, 15 (11), 3998–4019

(17) Carapito, C.; Duek, P.; Macron, C.; Seffals, M.; Rondel, K.; Delalande, F.; Lindskog, C.; Fréour, T.; Vandenbrouck, Y.; Lane, L.; Pineau; C. Validating Missing Proteins in Human

Sperm Cells by Targeted Mass-Spectrometry- and Antibody-based Methods. *Journal of Proteome Research* **2017**, *16* (12), 4340–4351.

(18) Schaeffer, M.; Gateau, A.; Teixeira, D.; Michel, P-A.; Zahn-Zabal, M.; Lane, L. The neXtProt peptide uniqueness checker: a tool for the proteomics community. *Bioinformatics* **2017**, 33(21):3471-3472.

(19) Jégou, B.; Sankararaman, S.; Rolland, A.D.; Reich, D.; Chalmel, F. Meiotic Genes Are Enriched in Regions of Reduced Archaic Ancestry. *Molecular Biology and Evolution* **2017**, 34 (8), 1974–1980.

(20) Leinonen, R.; Sugawara, H.; Shumway, M. The Sequence Read Archive. *Nucleic Acids Research* **2010**, 39 (Database issue), D19-D21.

(21) Bray, N. L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **2016**, 34 (5), 525–527.

(22) Bardou, P.; Mariette, J.; Escudié, F.; Djemiel, C.; Klopp, C. jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* **2014**, 15 (1), 293-100.

(23) Babicki, S.; Arndt, D.; Marcu, A.; Liang, Y.; Grant, J. R.; Maciejewski, A.; Wishart, D. S. Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Research* **2016**, 44 (W1), W147-53.

(24) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P.A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R.J.; Kraus, H.J.; Albar, J.P,.; Martinez-Bartolomé, S.; Apweiler, R., Omenn, G.S.; Martens, L.; Jones, A.R.; Hermjakob, H. ProteomeXchange

provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* **2014**, 32 (3), 223–226.

(25) Clermont, Y. The cycle of the seminiferous epithelium in man. *American Journal of Anatomy* **1963**, 112 (1), 35–51.

(26) Hofmann, O.; Caballero, O. L.; Stevenson, B. J.; Chen, Y.-T.; Cohen, T.; Chua, R.; Maher, C. A.; Panji, S.; Schaefer, U.; Kruger, A.; Lehvaslaiho, M.; Carninci, P.; Hayashizaki, Y.; Jongeneel, C.V.; Simpson, A.J.; Old, L.J.; Hide, W. Genome-wide analysis of cancer/testis gene expression. *Proceedings of the National Academy of Sciences* **2008**, 105 (51), 20422–20427.

(27) Whitehurst, A. W. Cause and Consequence of Cancer/Testis Antigen Activation in Cancer. *Annual Review of Pharmacology and Toxicology* **2014**, 54 (1), 251–272.

(28) Salmaninejad, A.; Zamani, M. R.; Pourvahedi, M.; Golchehre, Z.; Hosseini Bereshneh, A.; Rezaei, N. Cancer/Testis Antigens: Expression, Regulation, Tumor Invasion, and Use in Immunotherapy of Cancers. *Immunological Investigations* **2016**, *45* (7), 619–640.

(29) Babatunde, K.A.; Najafi, A.; Salehipour, P.; Modarressi, M.H.; Mobasheri, M.B. Cancer/Testis genes in relation to sperm biology and function. *Iranian Journal of Basic Medical Sciences* **2017,** (9), 967-974.

(30) Aken, B.L.; Achuthan, P.; Akanni, W.; Amode, M.R.; Bernsdorff, F.; Bhai, J.; Billis, K.; Carvalho-Silva, D.; Cummins, C.; Clapham, P.; Gil, L.; Girón, C.G.; Gordon, L.; Hourlier, T.; Hunt, S.E.; Janacek, S.H.; Juettemann, T.; Keenan, S.; Laird, M.R.; Lavidas, I.; Maurel, T.; McLaren, W.; Moore, B.; Murphy, D.N.; Nag, R.; Newman, V.; Nuhn, M.; Ong, C.K.; Parker,

A.;, Patricio, M.; Riat, H.S.; Sheppard, D.; Sparrow, H.; Taylor, K.; Thormann, A.; Vullo, A.; Walts, B.; Wilder, S.P.; Zadissa, A.; Kostadima, M.; Martin, F.J.; Muffato, M.; Perry, E.; Ruffier, M.; Staines, D.M.; Trevanion, S.J.; Cunningham, F.; Yates, A.; Zerbino, D.R.; Flicek, P. Ensembl 2017. *Nucleic Acids Research* **2016**, 45 (D1), D635-D642.

(31) Pang, S. W.; Lahiri, C.; Poh, C. L.; Tan, K. O. PNMA family: Protein interaction network and cell signalling pathways implicated in cancer and apoptosis. *Cellular Signalling* **2018**, *45*, 54–62.

(32) Sahashi, K; Sakai, K.; Mano, K.; Hirose, G. Anti-Ma2 antibody related paraneoplastic limbic/brain stem encephalitis associated with breast cancer expressing Ma1, Ma2, and Ma3 mRNAs. *Journal of Neurology, Neurosurgery & Psychiatry* **2003**, *74* (9), 1332–1335.

(33) Zhang, X.-L.; Liu, P.; Yang, Z.-X.; Zhao, J.-J.; Gao, L.-L.; Yuan, B.; Shi, L.-Y.; Zhou, C.-X.; Qiao, H.-F.; Liu, Y.-H.; Ying, X.-Y.; Zhang, J.-Q.; Ling, X.-F.; Zhang, D. Pnma5 is essential to the progression of meiosis in mouse oocytes through a chain of phosphorylation. *Oncotarget* **2017**, *8* (57), 96809-96825.

(34) Li, S.; He, Y.; Lin, Z.; Xu, S.; Zhou, R.; Liang, F.; Wang, J.; Yang, H.; Liu.; Ren, Y. Digging More Missing Proteins Using an Enrichment Approach with ProteoMiner. *Journal of Proteome Research* **2017**, 16 (12):4330-4339.

(35) Perdigão, N.; Rosa, A. C.; O'Donoghue, S. I. The Dark Proteome Database. *BioData Mining* **2017**, *10* (1).

(36) Perdigão, N.; Heinrich, J.; Stolte, C.; Sabir, K. S.; Buckley, M. J.; Tabor, B.; Signal, B.; Gloss, B. S.; Hammang, C. J.; Rost, B.; Schafferhans, A.; O'Donoghue, S. Unexpected

features of the dark proteome. *Proceedings of the National Academy of Sciences* **2015**, *112* (52), 15898–15903.

(37) Vassena, R.; Heindryckx, B.; Peco, R.; Pennings, G.; Raya, A.; Sermon, K. ; Veiga, A. Genome engineering through CRISPR/Cas9 technology in the human germline and pluripotent stem cells. *Human Reproduction Update* **2016**, 22 (4), 411–419.

(38) Kherraf, Z.E.; Conne, B.; Amiri-Yekta, A.; Kent, M.C.; Coutton, C.; Escoffier, J.; Nef, S.; Arnoult, C.; Ray, P.F. Creation of knock out and knock in mice by CRISPR/Cas9 to validate candidate genes for human male infertility, interest, difficulties and feasibility. *Molecular and Cellular Endocrinology* **2018**, 468, 70-80

(39) Coutton, C.; Vargas, A.S.; Amiri-Yekta, A.; Kherraf, Z.E.; Ben Mustapha, S.F.; Le Tanno, P.; Wambergue-Legrand, C.; Karaouzène, T.; Martinez, G.; Crouzy, S.; Daneshipour, A.; Hosseini, S.H.; Mitchell, V.; Halouani, L.; Marrakchi, O.; Makni, M.; Latrous, H.; Kharouf, M.; Deleuze, J.F.; Boland, A.; Hennebicq, S.; Satre, V.; Jouk, P.S.; Thierry-Mieg, N.; Conne, B.; Dacheux, D.; Landrein, N.; Schmitt, A.; Stouvenel, L.; Lorès, P.; El Khouri, E.; Bottari, S.P.; Fauré, J.; Wolf, J.P.; Pernet-Gallay, K.; Escoffier, J.; Gourabi, H.; Robinson, D.R.; Nef, S.; Dulioust, E.; Zouari, R.; Bonhivers, M.; Touré, A.; Arnoult, C.; Ray, P.F. Mutations in CFAP43 and CFAP44 cause male infertility and flagellum defects in Trypanosoma and human. *Nature Communications* **2018**, 9 (1), 686.

Table of content (TOC) / Abstract Graphic

For TOC Only