Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update

Yves Vandenbrouck, Lydie Lane, Christine Carapito, Paula Duek, Karine Rondel, Christophe Bruley, Charlotte Macron, Anne Gonzalez de Peredo, Yohann Coute, Karima Chaoui, et al.

Looking for missing proteins in the proteome of human spermatozoa: an update

Yves Vandenbrouck[1,2,3,#,§], Lydie Lane[4,5,#], Christine Carapito[6], Paula Duek[5], Karine Rondel[7], Christophe Bruley[1,2,3], Charlotte Macron[6], Anne Gonzalez de Peredo[8], Yohann Couté[1,2,3], Karima Chaoui[8], Emmanuelle Com[7], Alain Gateau[5], Anne-Marie Hesse[1,2,3], Marlene Marcellin[8], Loren Méar[7], Emmanuelle Mouton-Barbosa[8], Thibault Robin[9], Odile Burlet-Schiltz[8], Sarah Cianferani[6], Myriam Ferro[1,2,3], Thomas Fréour[10,11], Cecilia Lindskog[12],Jérôme Garin[1,2,3], Charles Pineau[7,§].

1. *CEA, DRF, BIG, Laboratoire de Biologie à Grande Echelle, 17 rue des martyrs, Grenoble, F-38054, France*
2. *Inserm U1038, 17, rue des Martyrs, Grenoble F-38054, France*
3. *Université de Grenoble, Grenoble F-38054, France*
4. *Department of Human Protein Sciences, Faculty of medicine, University of Geneva, 1, rue Michel-Servet, 1211 Geneva 4, Switzerland*
5. *CALIPHO Group, SIB-Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, CH-1211 Geneva 4, Switzerland*
6. *Laboratoire de Spectrométrie de Masse BioOrganique (LSMBO), IPHC, Université de Strasbourg, CNRS UMR7178, 25 Rue Becquerel, 67087 Strasbourg, France*
7. *Protim, Inserm U1085, Irset, Campus de Beaulieu, Rennes, 35042, France*
8. *Institut de Pharmacologie et de Biologie Structurale, Université de Toulouse, CNRS, UPS, France*
9. *Proteome Informatics Group, Centre Universitaire d'Informatique, Route de Drize 7, 1227 Carouge, CH, Switzerland*
10. *Service de Médecine de la Reproduction, CHU de Nantes, 38 boulevard Jean Monnet, 44093 Nantes cedex, France*
11. *INSERM UMR1064, Nantes, France*
12. *Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden*

#: Contributed equally to this work

§: **Corresponding authors contact information:**
yves.vandenbrouck@cea.fr, tel: +33 (0)4 38 78 26 74, fax: (33) (0)4 38 78 50 32; Charles Pineau: charles.pineau@inserm.fr, tel: +33 (0)2 23 23 52 79

Email addresses:

YV: yves.vandenbrouck@cea.fr
LL: lydie.lane@sib.swiss
CC: ccarapito@unistra.fr
PD: paula.duek@sib.swiss
KR: karine.rondel@univ-rennes1.fr
CB: christophe.bruley@cea.fr
CM: c.macron@unistra.fr
AGP: anne.gonzalez-de-peredo@ipbs.fr
YC: yohann.coute@cea.fr
KC: karima.chaoui@ipbs.fr
EC: emmanuelle.com@univ-rennes1.fr
AG: alain.gateau@sib.swiss
AMH: anne-marie.hesse@cea.fr
MM: marlene.marcellin@ipbs.fr
LM: loren.mear@univ-rennes1.fr
EMB: emmanuelle.mouton@ipbs.fr
TB: thibault.robin@etu.unige.ch
SC: sarah.cianferani@unistra.fr
OBS: schiltz@ipbs.fr
MF: myriam.ferro@cea.fr
TF: thomas.freour@chu-nantes.fr
CL: cecilia.lindskog@igp.uu.se
JG: jerome.garin@cea.fr
CP: charles.pineau@inserm.fr

The mass spectrometry data used for proteomics have been deposited with the ProteomeXchange Consortium via the PRIDE partner repository under dataset identifier PXD003947

Reviewers can access our data using the following account details:
Project Name: Quest for missing proteins in the human spermatozoa: an update
Project accession: PXD003947
Project DOI: 10.6019/PXD003947
Reviewer account details:
Username: reviewer98361@ebi.ac.uk
Password: v8rRsNI3

Keywords: human proteome project, spermatozoon, missing proteins, mass spectrometry proteomics, immunohistochemistry, bioinformatics, data mining, cilia

Abstract

The Chromosome-Centric Human Proteome Project aims to identify proteins classed as « missing » in the neXtProt knowledgebase. In this article, we present an in-depth proteomics analysis of the human sperm proteome to identify testis-enriched missing proteins. Using a range of protein extraction procedures and LC-MS/MS analysis, we detected a total of 235 proteins (PE2-PE4) for which no previous evidence of protein expression was annotated. Through a combination of LC-MS/MS and LC-PRM analysis, data mining and immunohistochemistry, we were able to confirm the expression of 206 missing proteins (PE2-4) in line with current HPP guidelines (version 2.0). Parallel Reaction Monitoring (PRM) acquisition combined with synthetic heavy labeled peptides was used to target 36 « one-hit wonder » candidates selected on the basis of prior PSM assessment. Of this subset of candidates, 24 were validated with additional predicted and specifically targeted peptides. Evidence was found for a further 16 missing proteins using immunohistochemistry on human testis sections. The expression pattern for some of these proteins was specific to the testis, and they could potentially be valuable markers with applications in fertility assessment. Strong evidence was also found for the existence of 4 proteins labeled as “uncertain” (PE5); the status of these proteins should therefore be re-examined. Our results show how the use of a range of sample preparation techniques combined with MS-based analysis, expert knowledge and complementary antibody-based techniques can produce data of interest to the community. All MS/MS data are available via ProteomeXchange under identifier PXD003947. In addition to contributing to the Chromosome-Centric Human Proteome Project, we hope the

availability of these data will stimulate the continued exploration of the sperm proteome.

Introduction

The Chromosome-Centric Human Proteome Project (C-HPP) aims to catalogue the protein gene products encoded by the human genome, in a gene-centric manner [1]. As part of this project, neXProt [2] has been confirmed as the reference knowledgebase for human protein annotation [3]. Numerous initiatives were launched worldwide to search for so-called missing proteins - proteins predicted by genomic or transcriptomic analysis, but not yet validated experimentally by mass-spectrometry or antibody-based techniques. These proteins are annotated with a “Protein Existence” (PE) score of 2 when they are predicted by transcriptomics analysis, 3 when they are predicted by genomic analysis and have homologs in distant species, and 4 when they are only predicted by genomic analysis in human or other mammals. The most recent neXtProt release (2016-01-11) contains 2949 such missing proteins. It was suggested by Lane and collaborators [4] that proteins that have been systematically missed might be expressed only in a few organs or cell types. The very high number of testis-specific genes that have been described [5] supports the hypothesis that the testis is a promising organ in which to search for elements of the missing proteome [6] [7]. The testis’ main function is well known: to produce male gametes, known as spermatozoa (commonly called sperm). Human spermatozoa are produced at a rate of about 1,000 cells/sec [8] by a complex, intricate, tightly controlled and specialized process known as spermatogenesis [9] [10]. Spermiogenesis is the final stage of

spermatogenesis, which sees the maturation of spermatids into mature, motile spermatozoa. The fact that the numbers of couples consulting for difficulties related to conceiving has increased in recent years, and that sperm quality has been shown to be altered in one in seven men, for example with abnormal motility or morphology [11], makes further study of these cells even more topically relevant.

Large numbers of spermatozoa can be recovered in highly pure preparation through non-invasive procedures, making it possible to access the final proteome of the germ cell lineage, and providing access to a large number of germ cell-specific proteins. Thus, MS-based proteomics studies of spermatozoa have generated highly relevant data [12]. Knowledge of the mature sperm proteome will significantly contribute to sperm biology and help us to better understand fertility issues.

In a recent study [13], the Proteomics French Infrastructure (ProFI; www.profiproteomics.fr) described a step-by-step strategy combining bioinformatics and MS-based experiments to identify and validate missing proteins based on database search results from a compendium of MS/MS datasets. The datasets used were generated using 40 human cell line/tissue type/body fluid samples. In addition to the peptide- and protein-level false discovery rate (FDR), supplementary MS-based criteria were used for validation, such as peptide spectrum match (PSM) quality as assessed by an expert eye, spectral dot-product - calculated based on the fragment intensities of the native spectrum (endogenous peptide) and a reference spectrum (synthetic peptide) - and LC-SRM assays that were specifically developed to target proteotypic peptides.

Some of these criteria were also used in a concomitant study [14] involving trans-chromosome-based data analysis on a high quality mass spectrometry data set to catalogue missing proteins in total protein extracts from isolated human spermatozoa. This analysis validated 89 missing proteins based on version 1.0 of the HPP guidelines (http://www.thehpp.org/guidelines/). The distribution of two interesting candidates (C2orf57 and TEX37) was further studied by immunohistochemistry in the adult testis, and their expression was confirmed in postmeiotic germ cells. Finally, based on analyses of transcript abundance during human spermatogenesis, we concluded that it would be possible to characterize additional missing proteins in ejaculated spermatozoa.

The study presented in this paper originated with the Franco-Swiss contribution to the C-HPP initiative to map chromosomes 14 (France) and 2 (Switzerland) by identifying additional missing proteins. Here, we combine the search for proteins that are currently classed as “missing” with an extensive examination of the sperm proteome. A single pool of human spermatozoa was treated by a range of approaches, and the most recent version of the guidelines for the identification of missing proteins was followed (Deutsch et *al.*, submitted; http://www.thehpp.org/guidelines/). We thus performed an in-depth analysis of human sperm using different fractionation/separation protocols along with different protein extraction procedures. Through MS/MS analysis, 4727 distinct protein groups were identified that passed the 1% PSM-, peptide- and protein-level FDR thresholds. Mapping of unique peptides against the most recent neXtProt release (2016-01-11) revealed 235 proteins (201 PE2, 22 PE3, 12 PE4) that are still considered missing by the C-H , and nine proteins annotated with a PE5 (uncertain) status in neXtProt. 8

Additional MS-based strategies (spectral comparison and parallel reaction monitoring assays) were applied to validate some of these missing proteins. Data mining was also applied to determine which proteins would be selected for validation by immunohistochemistry on human testes sections.

Materials and methods

Ethics and donor consent

The study protocol "*Study of Normal and Pathological Human Spermatogenesis*" was approved by the local ethics committee. The protocol was then registered as No. PFS09-015 at the French Biomedicine Agency. Informed consent was obtained from donors where appropriate.

Sample collection and preparation

Human semen samples were collected from five healthy donors of unproven fertility at Nantes University Hospital (France). The donors gave informed consent for the use of their semen for research purposes, and samples were anonymized. Semen samples were all obtained on-site by masturbation following 2 to 7 days of sexual abstinence. After 30 min liquefaction at room temperature under gentle agitation, 1 ml of each sample was taken. Aliquots were pooled and a protease inhibitor mix (protease inhibitor cocktail tablets, complete mini EDTA-free, Roche, Meylan, France) was added according to the manufacturer's instructions. To separate sperm cells from seminal plasma and round cells, the pooled sperm sample was loaded onto 1 mL of a 50% suspension of silica particles (SupraSperm, Origio, Malov, Denmark) diluted in Sperm Washing medium (Origio, Malov, Denmark). The sample was centrifuged at 400 x g for 15 min at room temperature. The sperm pellet was then washed once by resuspension in 3 mL of Phosphate-buffered saline (PBS) and centrifuged again at 400 x g for 5 min at room temperature. The supernatant was removed and the cell pellet was flash frozen in liquid nitrogen.

Protein extraction, digestion and liquid chromatography-tandem mass spectrometry (LC-MS/MS) analyses

MS/MS analysis of pooled sperm was performed using four different protocols based on a range of protein extraction procedures: i.) total cell lysate followed by a 1D SDS-PAGE separation (23 gel slices); ii.) separation of Triton X-100 soluble and insoluble fractions followed by a 1D SDS-PAGE separation (20 gel slices per fraction); iii.) total cell lysate, in-gel digestion, peptides analyzed by nano-LC with long gradient runs; iv.) total cell lysate, in-gel digestion, peptides fractionated by high-pH reversed-phase (Hp-RP) chromatography. For all protocols, tryptic peptides were analyzed by high-resolution MS instruments (Q-Exactive). These experiments were performed by the three proteomics platforms making up ProFI (Grenoble, Strasbourg and Toulouse). A detailed description of the protein fractionation using Triton X-100, protein extraction and digestion, and liquid chromatography-tandem mass spectrometry (LC-MS/MS) analyses performed in this study can be found in Supplementary Material.

MS/MS data analysis

Peak lists were generated from the original LC-MS/MS raw data using the Mascot Distiller tool (version 2.5.1, Matrix Science). The Mascot search engine (version 2.5.1, Matrix Science) was used to search all MS/MS spectra against a database composed of *Homo sapiens* protein entries from UniProtKB/SwissProt (release 2015-10-30, 84362 protein coding genes sequences (canonical and isoforms)) and a list of contaminants frequently observed in proteomics analyses (the protein fasta file for these contaminants is available at ftp://ftp.thegpm.org/fasta/cRAP, it consists of 118

sequences). The following search parameters were applied: carbamidomethylation of cysteines was set as a fixed modification, and oxidation of methionines and protein N-terminal acetylation were set as variable modifications. Specificity of trypsin digestion was set for cleavage after K or R, and one missed trypsin cleavage site was allowed. The mass tolerances for protein identification on MS and MS/MS peaks were 5 ppm and 25 mmu, respectively. The FDR was calculated by performing the search in concatenated target and decoy databases in Mascot. Peptides identified were validated by applying the target-decoy approach, using Proline software (http://proline.profiproteomics.fr/), by adjusting the FDR to 1%, at PSM- and protein-level. At peptide level, only the PSM with the best Mascot score was retained for each peptide sequence. Spectra identifying peptides in both target and decoy database searches were first assembled to allow competition between target and decoy peptides for each MS/MS query. Finally, the total number of validated hits was computed as Ntarget+Ndecoy, the number of false positive hits was estimated as 2xNdecoy, and the FDR was then computed as 2xNdecoy/(Ntarget+Ndecoy). Proline software automatically determined a threshold Mascot e-value to filter peptides, and computed the FDR as described so as to automatically adjust it to 1%. At protein level, a composite score was computed for each protein group, based on the MudPIT scoring method implemented in Mascot: for each non-duplicate peptide identifying a protein group, the difference between its Mascot score and its homology threshold was computed and these “score offsets” were then summed before adding them to the average homology (or identity) thresholds for the peptide. Therefore, less significant peptide matches contributed less to the total protein score. Protein groups were filtered by applying a threshold to this MudPIT protein score to obtain a final

protein-level FDR of 1%. To optimize discrimination between true positive and true negative protein hits, the software applies a selection scheme approach, by adjusting the FDR separately for the subset of proteins identified by more than one validated peptide, and then for the single-peptide hits. In accordance with version 2.0.1 of the HPP data interpretation guidelines (Deutsch et *al.*, submitted; http://www.thehpp.org/guidelines/), individual result files from each of the five MS/MS datasets were combined, and a procedure to produce a protein-level FDR threshold of 1% was re-applied. This combination of result files created a single identification dataset from a set of identification results and was performed as follows: all PSM identified and validated at 1% were merged to create a unique combination of amino acid sequences and a list of PTMs located on that sequence which were aggregated in a single "representative" PSM. The newly created PSM were then grouped into proteins and protein families [41]. The resulting dataset therefore provides a non-redundant view of the identified proteins present in the original sample.

Detection of missing proteins

The sequence of each peptide identified was searched in all the splicing isoform sequences present in neXtProt release 2016-01-11, using the pepx program developed in-house (https://github.com/calipho-sib/pepx). The method is based on a 6-mer amino acid index that is regenerated at each release; the 6 aa length was chosen as it significantly speeds up the mapping process. Leucine and isoleucine were considered equivalent. A peptide is considered to match an isoform sequence when all the 6-mers covering the peptide return the same sequence. Peptides were subsequently checked against the retrieved isoform sequence(s) to ensure an exact

string match. All matches to splicing isoforms derived from a single entry were considered relevant for the identification of the entry.

To further validate the identification of missing proteins, a second round of peptide-to-protein mapping was performed taking into account the 2.5 million variants described in neXtProt (SNPs and disease mutations). Currently, pepx only considers a single amino acid substitution or deletion in the 6-mer; substitutions and deletions more than 1 aa in length, as well as insertions, are not taken into account. Consequently, pepx returns a match if single amino acid variations in the isoform sequence are spaced at least 5 amino acids apart. Peptides matching more than one entry when variants were taken into account were excluded as they are potentially not proteotypic.

Data availability

All MS proteomics data, including reference files (readme, search database, .dat files) form a complete submission with the ProteomeXchange Consortium [15]. Data were submitted via the PRIDE partner repository under dataset identifiers PXD003947 and 10.6019/PXD003947.

Additional MS-based validation (MS/MS analysis of synthetic peptides, comparison of reference/endogenous fragmentation spectra and LC-PRM analysis).

Synthetic heavy labeled peptides were purchased (crude PEPotec™, Thermo Fisher Scientific) for 36 “one-hit wonder” candidates selected based on visual inspection of PSMs. The 36 peptides initially identified were synthesized along with two additional

predicted proteotypic peptides per protein, when possible. Thus, a total of 100 peptides were synthesized (Supplementary Table 4). The labeled peptides corresponding to the 36 peptides initially identified were mixed together and analyzed by LC-MS/MS (Q Exactive Plus, Thermo Fisher Scientific) to acquire HCD fragmentation spectra for comparison with the initial spectra, in the closest possible conditions. All MS/MS spectrum pairs are shown in Supplementary Figure 1. Following this step, targeted assays using a Parallel Reaction Monitoring (PRM) acquisition approach were developed on the same LC-MS/MS platform to target all 100 peptides, first in a total protein fraction prepared in stacking gel bands and subsequently in gel bands obtained from 1D SDS PAGE separation of the Triton X-100 insoluble proteins fraction. See Supplementary Material for details of MS experiments.

Data mining to select missing proteins for further characterization.

For each protein identified by MS, the tissue expression profile based on RNA sequencing analysis was retrieved from the Human Protein Atlas portal (version 14) (www.proteinatlas.org/). The evolutionary conservation profile was determined by a BLAST analysis using UniProtKB "Reference Proteomes" as target. In addition, homologs were systematically searched for in a number of ciliated organisms from distant groups including *Choanoflagellida* (*Salpingoeca, Monosiga*), Chlorophyta (*Micromonas, Volvox, Chlamydomonas*), Ciliophora (*Paramecium, Oxytricha, Stylonychia, Tetrahymena, Ichthyophthirius*), Trypanosomatidae (*Trypanosoma, Phytomonas, Leishmania, Angomonas, Leptomonas*), Cryptophyta (*Guillardia*), *Naegleria gruberi* and Flagellated protozoan (*Bodo saltans*). For each protein and all

its orthologs, all existing names, synonyms, and identifiers were collected from appropriate model organism databases. These names were used to query PubMed and Google. Proteins to be further validated by immunohistochemistry were selected based on a combination of criteria including antibody quality, available immunohistochemistry data in Protein Atlas (version 14), phenotype of mutant organisms, predicted or experimental biological function, tissue localization, interacting partners and phylogenetic profile. Uncharacterized proteins that are selectively expressed in testis or ciliated tissues and well conserved in ciliated organisms, interact with testis or cilia-related proteins, for which knockout model organisms show a reproduction phenotype, and for which high quality antibodies from the Human Protein Atlas were available were considered the best candidates for further validation.

Immunohistochemistry

To confirm the germline expression of proteins of interest, immunohistochemistry experiments were performed on human testes fixed in 4% paraformaldehyde and embedded in paraffin, as described [16]. Normal human testes were collected at autopsy at Rennes University Hospital from HIV-1-negative cadavers.

Paraffin-embedded tissues were cut into 4 µm-thick slices, mounted on slides and dried at 58 °C for 60 min. Immunohistochemical staining, using the Ventana DABMap and OMNIMap detection kit (Ventana Medical Systems, Tucson, USA), was performed on a Discovery Automated IHC stainer. Antigen retrieval was performed using proprietary Ventana Tris-based buffer solution, CC1, at 95 °C 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% H2O2, (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 missing proteins (Atlas Antibodies) diluted in TBS containing 0.2% Tween -20 (v/v) and 3% BSA (TBST--BSA). The antibody dilutions used are listed in Supplementary Table 6. Non-immune rabbit serum (1:1000) 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. Signal was enhanced using the Ventana DABMap Kit or Ventana OMNIMap kit. Sections were then counterstained for 16 min with hematoxylin (commercial solution, Microm), and for 4 min with bluing reagent (commercial solution, Microm) before rinsing with milliQ 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.

Results and discussion

Overall workflow

The overall workflow for the detection and validation of missing proteins is illustrated in Figure 1 and described in the Material and Methods section, with full details on sample preparation in Supplementary Materials. By applying this workflow, we produced a list of 235 "candidate missing protein" entries (PE2-4) and nine PE5 entries. This list was divided into two distinct subsets in line with version 2.0.1 of the HPP data interpretation guidelines (Deutsch et *al*., submitted;

http://www.thehpp.org/guidelines/): those validated by two or more distinct uniquely-mapping peptide sequences of length ≥ 9 amino acids, and those detected based on only one unique peptide of length ≥ 9 amino acids. Each PSM from the latter subset was then examined to seek additional MS-based evidence (PSM quality as assessed by an expert, comparison between endogenous and reference (synthetic peptide) fragmentation spectra and LC-PRM assays). In parallel, the full list of missing or uncertain protein entries (PE2-5) was mined by gathering additional information from public resources, bioinformatics analysis and the literature. This information was used to select a subset of high priority proteins for further immunohistochemistry analysis on human testes sections.

Analysis of the human sperm proteome

Because the workflow involved a range of enrichment strategies and separation protocols, including peptide pre-fractionation protocols based on high pH reverse phase (HpH-RP) chromatography that have been shown to be orthogonal to subsequent online reverse phase nano-LC separation of peptides [17], sensitivity was high and coverage extensive. This type of “cover all bases” approach has been shown to be particularly efficient for improving the detection of missing proteins [18]. Validation was subsequently performed for each results file (.dat) through the target-decoy approach [19], using the in-house developed Proline software (http://proline.profiproteomics.fr/), by adjusting the FDR to 1%, at PSM- and protein-level. In a second step, individual results files were combined for each dataset and a 1% protein-level FDR was applied to comply with the HPP data interpretation

guidelines, version 2.0.1 (http://www.thehpp.org/guidelines/, guideline #9) (see Materials and methods) . Supplementary Table 1 lists PSM-, peptide-, and protein-level FDR values along with the total number of true positives and false positives at each level for the five sperm sample preparation methods (individual results files for each fraction and after combination). After MS/MS data processing and filtering, a total of 4727 distinct protein groups passed the 1% PSM-, peptide- and protein-level criteria. Detailed information on the proteins identified from the five MS datasets are reported in Supplementary Table 2. Protein identification and their distribution across the five MS datasets were then compared to assess their contribution to total human sperm proteome data sets (Figure 1). The Venn diagram shows that 1526 proteins detected were present in all five datasets. In addition, each fractionation/separation method used in this study provided a significant added-value in terms of proteome coverage. Thus, Triton X-100 insoluble and soluble fractions, whole cell lysate analyzed by 1D SDS-PAGE, high-pH reverse phase peptide fractionation and long gradient runs allowed gains of 8.3, 6.7, 6.7, 3.8 and 0.8%, respectively. These results clearly emphasize the complementarity of the different enrichment techniques when seeking to obtain exhaustive (or as exhaustive as possible) proteome coverage.

Iin 2014, Amaral *et al.* [20] published a sperm proteome comprising 6198 proteins identified by combined MS-based analysis based on 30 LC-MS/MS proteomics studies. Crossing identifier lists between their data and the sperm proteome produced here revealed that our analysis yielded 1140 additional proteins. However, further investigation will be necessary to ensure a fair assessment as Amaral *et al.* applied different validation criteria to ours (*e.g.*, identification of at least two peptides with a protein-level FDR < 5%).

Focusing on missing proteins identified from the sperm proteome

Missing proteins were detected as described in Materials and Methods using the most recent neXtProt release (2016-01-11). The first step in this detection took all possible splice isoforms and I/L ambiguities into account, but no single amino-acid variants. This produced a list of 235 missing proteins (PE2-4) and nine uncertain proteins (PE5). Among the PE2-4 protein entries, 188 were identified and validated (<1% FDR) by at least two or more distinct uniquely-mapping peptide sequences of length ≥ 9 amino acids, while the remaining 47 were associated with only one unique peptide ≥ 9 amino acids (Table 1). We named this subset of missing proteins "one-hit wonders" and considered it separately for further MS-based analysis (see next section). In fact, 5 of these protein entries were identified with 2 (A2RUU4, Q5GH77, Q8NG35, Q8WTQ4) or 3 (Q6PI97) unique peptides, but since only one of these had a length equal to or greater than 9 amino-acids (Table 2), these entries were nevertheless considered to be "one-hit wonders". Among the full set of 235 proteins considered to be missing by neXtProt (PE2-4), 180 have no annotated function, while 56 are predicted to have at least one transmembrane helix (TMH). Full details (description, number of unique peptides, chromosome location, etc.) on the missing proteins are reported in Table 2 and Supplementary Table 3. The Venn diagram illustrated in Figure 3A shows how each fractionation protocol contributed to the identification of the whole set of missing proteins (PE5 included). Thirty-three proteins were detected in all five datasets, whereas 63 were specifically detected in a given dataset (30 in "insoluble fraction-1D gel", 9 in "soluble fraction-1D gel" gel, 17 in "whole lysate-1D gel", 6 in "peptide HpH-RP" and 1 in "whole lysate-long runs"**).**

We noticed that among the 30 proteins only detected in the insoluble fraction, around one-third (9 proteins) were annotated with at least one TMH (see Supplementary Table 3) illustrating the benefit of preliminary subcellular fractionation for the identification of hydrophobic proteins. Unsurprisingly, only a small number of proteins (8) with at least one TMH were detected in all five sperm datasets, and an even smaller number of them were specifically detected when protocols starting with the soluble fraction or a whole cell lysate were applied (5 in “soluble fraction-1D gel”, 3 in “whole lysate-1D gel”, 2 in “peptide HpH-RP” and none in “whole lysate-long runs”). The missing proteins identified in sperm were found to be distributed across all chromosomes, except chromosome Y and chromosome 21, with the highest number (21 proteins) coded by genes present on chromosome 1 (Figure 3B). In terms of coverage of missing proteins (PE5 included), around 80% were supported by two or more distinct uniquely-mapping peptide sequences of length ≥ 9 amino acids, with some proteins very well covered (up to 78 peptides). The other 20% of missing proteins (52 proteins) were identified by only one unique peptide sequence of length ≥ 9 amino acids (Figure 3C).

To comply with recent C-HPP guidelines (version 2.0.1; http://www.thehpp.org/guidelines/), we also considered alternative mappings of all peptides of length >= 9 amino acids mapping to PE2-5 proteins by taking the 2.5 million single amino acid variants available in neXtProt into account. This analysis indicated that 13 peptides mapping to 12 missing (PE2-4) proteins could correspond to an alternative peptide sequence. Peptide ‘TKMGLYYSYFK’ maps uniquely to DPY19L2P1 (Q6NXN4), but if reported SNPs are considered, it could also map to the PE1 proteins DPY19L2 (Q6NUT2) and DPY19L1 (Q2PZI1). Peptide

'TPPYQGDVPLGIR' maps uniquely to the PE2 protein SPAG11A (Q6PDA7), but if reported SNPs are considered, it could also map to the paralog SPAG11B (Q08648), which was also identified in this study with two other unique peptides. Likewise, the PE2 protein LRRC37A (A6NMS7) was identified by two peptides 'NAFEENDFMENTNMPEGTISENTNYNHPPEADSAGTAFNLGPTVK' and 'SKDLTHAISILESAK'. If reported SNPs are considered these peptides could also map to the PE2 protein LRRC37A2 (A6NM11) and the PE1 protein LRRC37A3 (O60309), respectively. Therefore, DPY19L2P1, SPAG11A and LRRC37A will need further investigation to validate their existence at protein level. Peptides 'QNVQQNEDASQYEESILTK' and 'QNVQQNEDATQYEESILTK' were both validated by PRM (see below) but only differ by one residue (S or T) and map to two close paralogs, RSPH10B2 (B2RC85) and RSPH10B (P0C881), respectively. An S71T variant has been reported, thus, the two paralogs cannot be distinguished based on these peptides. Since RSPH10B2 and RSPH10B only differ by three amino acids in total, it is very difficult to find other suitable unique peptides for validation. However, the identification of RSPH10B2 (B2RC85) was confirmed by PRM using the additional peptide 'EEEFNTWVNNTYVFFVNTLFHAYK'. The PE2 protein ZDHHC11 (Q9H8X9) was identified by 2 unique peptides, but one of them ('GVLQQGAGALGSSAQGVK') could also map to its paralog ZDHHC11B if SNP were considered. The six other peptides that lost their unicity when SNP were taken into account map to proteins for which there were more than three peptides of length $\geq$9 amino acids.

In the small group of nine proteins with a PE5 status (uncertain), three were identified and validated (<1% FDR) by at least two or more distinct uniquely-mapping peptide

sequences of length ≥ 9 amino acids. ATXN3L (Q9H3M9) was identified by six peptides - of which one would lose its unicity if SNPs were considered. This protein has now been characterized as a deubiquitinylase [22] [23] [24] [25], and its entry in UniProtKB/Swiss-Prot is currently under revision by the curators. HSP90AB4P (Q58FF6) and GK3P (Q14409), identified, respectively, by two and three peptides ≥ 9 amino acids in length are annotated as pseudogenes in most protein databases. Their status should be revised based on the results presented here.

The six others (PRSS41, LINC00521, FTH1P19, FAM205BP, SPATA31D3 and SPATA31D4) were detected with only one distinct uniquely-mapping peptide of length ≥ 9 amino acids. PRSS41 (Q7RTY9) is the ortholog of the recently-characterized testis-specific serine protease Prss41/Tessp-1 [21], and should no longer be considered as a pseudogene - the PE5 status of the entry is under revision by UniProtKB/Swiss-Prot curators. The putative uncharacterized protein encoded by LINC00521 (Q8NCU1) was detected by a 26-amino acid peptide for which no other match in the human proteome was found, even when possible variants were considered. The identification of the putative pseudogene FTH1P19 (P0C7X4) is also plausible since the peptide identified could only match the validated FTH1 protein (P02794) if both the rare D172N variant and an unknown S164A variant were considered. Nevertheless, according to the current HPP guidelines, the identifications of PRSS41, LINC00521 and FTH1P19 still need to be confirmed using other peptides.

The three remaining identifications are more dubious. Indeed, the peptide identifying FAM205BP (Q63HN1) could also match FAM205A (Q6ZU69), a PE1 protein, if its

very common (30%) M499V variant form was considered. Likewise, as discussed in Jumeau et al. [14], the peptides mapping to SPATA31D4 (Q6ZUB0) and SPATA31D3 (P0C874) do not confidently identify one protein or the other. This issue cannot be resolved without access to data relative to the genomic sequences of the donors, as both variants SPATA31D3 R882G (dbSNP:rs815819) and SPATA31D4 G882R (dbSNP:rs138456481) may be present in the pooled sample studied.

Investigating one-hit wonder missing proteins using MS-based criteria

As one-hit wonder proteins could potentially correspond to incorrect PSM assignment or false positives that passed the 1% FDR threshold, the recent HPP guidelines recommend that additional MS-based analyses be performed to provide further proteomics evidence (Deutsch et *al.*, submitted; http://www.thehpp.org/guidelines/). We therefore investigated our subset of one-hit wonder missing or uncertain proteins using additional MS-based criteria, as described in [13]. The first criterion relied on a blinded inspection of each MS/MS spectrum for the 52 missing or uncertain proteins (47 PE2-PE4; 5 PE5), all of which were identified by a unique peptide $\geq$ 9 amino acids in length.

The quality control of the PSMs corresponding to these unique peptides was carried out by visual validation by at least two mass spectrometry experts from each of the three sites of the ProFI infrastructure (Grenoble, Strasbourg, Toulouse) and from the Protim core facility (Rennes). PSM quality was classed in two categories: low and high. A high classification was based on the following spectral features: (i) the

presence of y-ion and b-ion series; (ii) peak intensities; (iii) quality of the match between the experimental and theoretical spectra. A subset of 18 peptides were assigned a “high” quality tag by three out of the four sites and were therefore preferentially selected for further MS validation (data not shown). In addition to these 18 candidates, other one hit wonders were selected that were awarded a majority rather than a consensual “high” quality attribute. A supplementary filter was applied by retaining only peptides that could be synthesized (*i.e.,* peptides shorter than 25 amino acids). This filter was necessary as further validation steps required the availability of a synthetic peptide for each candidate to be validated. Thus, of the initial 52 “one-hit wonders” we ended up with a final list of 36 peptides mapping to 34 missing proteins (PE2-4) and 2 uncertain proteins (PE5) for further assessment (among the 18 that had a unanimous high-quality vote, 17 peptides were selected, see Supplementary Table 4). Synthetic labelled peptide versions of all 36 peptides were ordered to allow systematic comparison of identifications with a synthetic version of the peptide (*i.e.,* same charge, same instrument fragmentation conditions). The goal of this step was to increase confidence in the identification of each peptide. To allow readers to assess spectrum quality and peak intensity patterns between the endogenous peptide and its synthetic counterpart peptide for themselves, an example spectrum for synthetic peptides is shown alongside the naturally-derived peptides for protein entry Q9H693 in Figure 4A; likewise all other comparative spectra are presented in Supplementary Figure 1. To objectively assess peptide ‘VEAALPYWVPLSLRPR’ (protein entry Q9H693; shown in Figure 4A), we calculated the spectral dot-product score (SDPscore) [26] which corresponds to the spectral correlation score calculated for the intensities of all common singly-charged b- and y-

ions of the reference spectrum and the native spectrum, as in our previous studies [13].

A SDPscore of 0.954 was obtained for this peptide, indicating that its MS/MS fragmentation pattern is very similar to the pattern obtained for the reference synthetic peptide.

Additional experimental validation using targeted LC-PRM assays

In a final MS-based validation attempt, targeted MS assays were developed for the 36 candidate "one-hit wonder" proteins to try to re-detect their proteotypic peptide co-eluting with its synthetic labeled counterpart. Samples for these assays were prepared independently. In addition to the original unique peptide identified, two additional predicted proteotypic peptides were selected, when possible, and synthesized for all 36 proteins (Supplementary Table 4). A total of 100 labeled peptides were synthesized. These peptides were mixed with protein digests and the heavy and light forms were targeted for analysis using parallel reaction monitoring scanning on a high resolution Q-Orbitrap mass spectrometer. In a first attempt, all proteotypic peptides were targeted in an unfractionated total protein extract prepared in stacking gel bands. This allowed us to unambiguously detect perfectly coeluting specific light/heavy transition groups for 24 proteotypic peptides corresponding to 21 of the 36 proteins. Examples of light/heavy transition group coelution are presented for protein Q9H693 for its initial peptide (Figure 4B) and for one additional predicted peptide that was detected thanks to the increased sensitivity of targeted assays (Figure 4C). Subsequently, to attempt to validate more candidates, the remaining undetected peptides were targeted in the insoluble proteins fraction after preliminary

separation on 1D SDS-PAGE. Samples were prepared as previously described for the total proteome analysis (protocol ii). The insoluble fraction was chosen for these targeted assays as a majority of the one hit wonders (18 out of the 36, Supplementary Table 4) were identified in samples prepared by this protocol. Thus, our chances of validating peptides were greater with samples prepared by this protocol. These final MS experiments unambiguously validated 27 additional peptides belonging to 20 out of the 36 proteins, not all the same as the previously validated 21 proteins. Thus, in total, 24 out of the 36 proteins were validated with additional predicted and specifically targeted peptides. This successful validation with additional peptides confirms the utility of highly sensitive targeted assays compared to non-targeted data-dependent LC-MS/MS acquisitions, and shows that this approach is suitable for unambiguous validation of missing proteins. All the results obtained with targeted LC-PRM assays can be found in Supplementary Table 4 and Supplementary Figure 1.

Bioanalysis of the missing proteins

For all the missing proteins identified by MS-based analysis, the chromosomal location, PE status, predicted number of TMH, and functional annotation were retrieved from neXtProt (see Table 2 and Supplementary Table 3 for details). Using a previously described methodology [14], we also extensively mined publicly available transcriptome data, i.e., the "*Human testis gene expression program*" described by Chalmel and collaborators [5] to check whether the missing and uncertain proteins identified in the present study corresponded to genes carrying the testis signature,

related to the onset of human spermatogenesis, or whether they corresponded to genes expressed only at the very end of spermiogenesis.

Only 132 of the 235 missing proteins and three of the nine uncertain proteins identified in this study corresponded to genes referenced in the "*Human testis gene expression program*" [5], with an increasing expression in seminiferous tubules containing post-meiotic germ cells (Johnsen score >=7). Up to 76 (+ 2 uncertain) of these proteins corresponded to genes specifically expressed in the testis (SET), 31 (+ 1 uncertain) corresponded to genes preferentially expressed in the testis (PET), and 25 corresponded to genes with intermediate (IE) or ubiquitous (UE) expression in the testis (Supplemental Table 3). This information was not used to select candidates for validation, but it is important to help understand the results of the immunohistochemistry experiments. Of note is that 26 and 2 proteins corresponding respectively to SET and PET genes are present in the current list of 1057 testis-enriched proteins in Human Protein Atlas, an update of the initial list from from Djureinovic *et al.* [42]. That suggests the spermatozoon has great potential for the identification of additional missing proteins. It also shows that proteins that are not considered as highly enriched in the testis might concentrate in germ cells during late spermiogenesis and become accessible in the spermatozoa. However, it is also important to note that a significant subset of missing proteins identified in the present study corresponded to genes that are testis-specific but do not belong to the "*Human testis gene expression program*" as their expression is not enriched in seminiferous tubules containing post-meiotic germ cells (data not shown).

To date, up to 111 of the 244 PE2-5 proteins identified have been subjected to extensive data and literature mining (Supplementary Table 5). The information gleaned from this data mining was used to establish priorities for further antibody-based studies. The first selection criterion was based on transcriptomics analysis. RNA sequencing results from the Human Protein Atlas database indicated that 88 of these 111 proteins were either specifically expressed in testis, or were enriched in a small group of tissues in which testis has one of the two highest expression levels (column C). These proteins were assigned a very high score (dark green). The eight proteins which were enriched in a small group of tissues including testis, but for which expression levels in testis were not among the two highest ones were assigned with a lower score (light green). In contrast, a low (red) priority score was assigned to six proteins which were shown either to be specifically expressed in organs other than testis, or to be undetectable by RNA sequencing [5 43]. The remaining proteins are either ubiquitously expressed or expressed at low levels in testis and were assigned a neutral score (white).

The second selection criterion was based on phylogenetic profiling. We sought and report (column D) on the presence of homologs in *S. cerevisiae* and a number of ciliated organisms from distant groups. The 18 proteins for which homologs were present in at least three of the ciliated groups were assigned with a very high priority score (dark green) because we hypothesized that they could be involved in ciliogenesis. The 17 having homologs in one or two of the ciliated groups, but not in yeast were assigned a high priority score (green). In contrast, a low priority score (red) was assigned to the three proteins that were found to be conserved in yeast as they are not expected to play a specific role in ciliogenesis or human reproduction.

The third selection criterion was based on the phenotypes observed in knock-out mice (column E). The four proteins for which a deletion led to a reproduction phenotype, were assigned a very high priority score (dark green). The ten proteins whose deletion had no effect on fertility were assigned a low priority score (red).

The fourth selection criterion was based on information retrieved from literature mining (gene expression regulation, protein protein interactions, function). This information is summarized in column F. Proteins for which published information suggested a putative role in spermatogenesis, ciliogenesis, or cilia function (85 proteins) were assigned a very high (dark green, 28 proteins) or high (green, 56 proteins) priority score. In the course of literature mining, we noticed that up to 76 of these 111 proteins had already been detected by mass spectrometry in human testis [27] or sperm [28] [29] [30]. However, these identifications have not yet been curated by PeptideAtlas or neXtProt; thus, the existence of these proteins was not considered validated at the time of writing.

The fifth selection criterion was based on immunohistochemistry data retrieved from the Human Protein Atlas (column G). The nine proteins that were specifically observed in germ cells, or associated with ciliary or cytoskeletal structures were assigned a very high priority score (dark green). Twelve other proteins observed in testis or ciliated cells were assigned a high priority score (green). Conversely, four proteins which were not seen in testis but were clearly seen in other tissues were assigned a low priority score (red).

Based on the combination of all these criteria, each of the 111 PE2-5 proteins was assigned a score grading the potential relevance of the protein in sperm and testis biology as high/medium/low. Seven entries which were initially classed as of

high/medium interest were down-graded to low interest proteins because their localization in human sperm cells had already been published (column H). This left a total of 33 “high interest” proteins. Among them, we selected the 26 for which a Human Protein Atlas antibody was available which passed the Protein Arrays (PA) test with a single peak, corresponding to interaction only with its own antigen. We selected 12 additional proteins from among the 42 scored as “medium interest”, and one from among the 36 scored as “low interest” (columns I and J).

Orthogonal immunohistochemistry evidence

Immunohistochemical studies were undertaken to provide non-MS-based evidence for the expression of the 39 missing proteins selected based on this data mining process. Specific immunohistochemistry staining in human testes was obtained for 16 missing proteins using antibodies from the Human Protein Atlas without need for further technical improvement (see Supplemental Table 6). Results from these experiments show that all 16 selected proteins displayed immunoreactive signals at various intensities in germ cells at all stages of their development (Figures 5A, 5B). No staining above background levels was visible in interstitial cells or somatic cells in the seminiferous tubules for any of the antibodies (Figures 5A, 5B). The staining intensity for missing proteins increased significantly from pre-meiotic and meiotic germ cells onwards (for C19orf81, C20orf85, SSMEM1, CCT8L2, WDR88, SAMD15, WDR93 and NOXRED1) or post-meiotic germ cells onwards (for CXorf58, C17orf105, C10orf53, C10orf67, FAM187B, AXDND1, GOT1L1 and CFAP46). This profile was to be expected as all proteins, except C10orf67, C19orf81 and WDR93, corresponded to genes in the TGEP, and their expression was expected to gradually

increase in later stages of sexual maturation [5] (Supplemental Tables 3 and 6). The expression levels for genes coding for C10orf67, C19orf81 and WDR93 may be below the threshold required to be part of the TGEP, even though immunoreactivity was observed in the germ cell lineage.

All 16 missing proteins whose expression was demonstrated in situ deserve further study to determine their role in sperm biology. However, due to their very specific expression patterns, 5 of them call for an immediate focus. Indeed, based on our immunohistochemistry data, staining for CXorf58, C20orf85 and CFAP46 was concentrated in late spermatids at the level of the acrosome under formation, a sperm-specific organelle essential for fertilization, with CFAP46 and CXorf58 displaying spectacular annular staining. Expression of FAM187B and AXDND1 displayed a slightly different profile in the adult testis, with immunoreactivity concentrated in the cytoplasmic region of elongating and elongated spermatids undergoing intense remodeling. This staining profile has previously been shown to be associated with the expression of proteins playing a role in sperm maturation.

CXorf58 is an orphan protein with no curated functional comments in UniProtKB/Swiss-Prot; TargetP and MitoProt prediction programs predict a mitochondrial localization. Its expression in sperm cells was confirmed by PRM based on endogenous peptide 'SFFDEAPAFSGGR', detected only in the insoluble fraction, and the additional peptide 'DISAQIIQR'. The staining pattern observed in our immunohistochemistry experiment suggests that this protein is mainly located at the lower part of the head and midpiece of spermatids. This position is in favor of a link to mitochondria. Interestingly, mitochondria are grouped in the midpiece of mature

spermatozoa, and numerous studies support the proposal that these organelles are important for sperm function and fertilization (for a review, see [31]).

C20orf85 is also an orphan protein. It is expressed in the epithelium of the airways [32], with levels increasing sharply during mucociliary differentiation. Its expression clusters with that of genes involved in regulation of cytoskeletal organization and intracellular transport [33]. In the adult testis, C20orf85 immunoreactivity was very strongly concentrated in the acrosome as it formed in elongating spermatids. The protein might migrate further down to the midpiece or flagellum in mature sperm. In yeast two-hybrid experiments, the murine C20orf85 ortholog (1700021F07Rik) was shown to interact with CCNB1IP1, a putative ubiquitin E3 ligase that is essential for chiasmata formation, and hence fertility [34]. Together with these observations, the altered expression of C20orf85 in asthenozoospermic patients [20] suggests a possible role in sperm movement.

Finally, CFAP46 is the ortholog of Chlamydomonas FAP46, which is known to be part of the central apparatus of the cilium axoneme and to play a role in cilium movement [35]. In the adult testis, immunostaining for CFAP46 was annular in late spermatids, a pattern that is typical of migration that will continue further down the midpiece or flagellum. Interestingly, CFAP46 expression has been shown to be downregulated in patients with primary ciliary dyskinesia [36]. That, together with our observations, is in favor of a central role in sperm movement.

FAM187B is an orphan protein with no curated comments in UniProtKB/Swiss-Prot, except an indication that it is a transmembrane protein [30]. The very peculiar localization of the immunohistochemistry staining for this protein, in the cytoplasmic region of elongating spermatids, suggests that FAM187B may play a role in cytoplasm displacement and elimination that take place during spermiogenesis. Interestingly, FMA187B mRNA expression in sperm has been proposed as a valuable diagnostic indicator of sperm survival, fertility and capacity to promote early embryogenesis [37].

AXDND1 is an intracellular protein which is selectively expressed in the nasopharynx, bronchus, testis and fallopian tubes, according to HPA immunochemistry data. It is highly conserved in Vertebrates and in the Choanoflagellate *Salpingoeca*, and contains an axonemal dynein light chain domain (IPR019347). Interestingly, the outer arm dynein complex is the main propulsive force generator for ciliary/flagellar beating. The staining pattern for the protein, positive in the cytoplasmic region of elongating spermatids undergoing extensive remodelling, matches with a possible role for the protein in mobility of the sperm flagellum.

Conclusion

In this study, MS-based analysis of sperm samples detected 235 missing (PE2-4) and 9 uncertain (PE5) proteins. Among these, 206 missing and 4 uncertain proteins were validated with at least two or more distinct peptide sequences with ≥ 9 amino

acids that mapped only to a single protein entry, even when possible variants were considered. In line with version 2.0 of the HPP Data Interpretation Guidelines, these 210 proteins can therefore be considered as validated. Twenty-four of these proteins were confirmed by LC-PRM assays, and 16 by IHC on human testis sections. IHC studies not only allowed us to confirm the existence of the proteins in sperm, but also to hypothesize a biological role for some of them (*i.e.,* CXorf58, C20orf85, CFAP46, FAM187B and AXDND1). The combination of LC-PRM and IHC was clearly instrumental in validating two “one hit wonders”: CXorf58 and C19orf81.

The importance of considering possible variants was illustrated by the cases of eight proteins, including three PE5, identified with peptides that lost their proteotypicity when possible variants were considered. These eight proteins therefore cannot be considered validated with our data.

The remaining 26 proteins were detected with only one unique peptide ≥ 9 amino acids. Six of these peptides were confirmed by LC-PRM and for three others, manual inspection unanimously indicated high quality LC-MS spectra. Thus, these nine identifications are reported here with confidence. However, the current HPP guidelines require MS-based validation of additional peptides for these proteins, or antibody detection to definitively validate their existence. The 17 other peptides passed the FDR criterion, but visual examination of their spectra indicated insufficient quality to warrant further study. Hence, the identification of these 17 proteins can only be considered dubious.

The Swiss-French collaborative project investigating the human sperm proteome in the context of the C-HPP started three years ago. In our previous article [14], we

reported the detection of 94 PE2-5 proteins in sperm using an LTQ-Orbitrap XL mass spectrometer, with at least one peptide of 9 aa. This dataset was submitted to proteomeXchange, reanalysed by PeptideAtlas, combined with other datasets and used by neXtProt to validate protein existence, based on the stringent guidelines established in 2016. Finally, 54 of these 94 proteins, including three PE5, were validated and are now annotated PE1. It is remarkable that all 94 of these proteins were identified in the present study. Except for TMEM239 (Q8WW34), detected with a single 24-amino acid peptide, all proteins were detected with at least 2 peptides, and often many more (Supplementary Table 7). The coverage of each protein was considerably improved by the use of cutting edge instruments *(i.e.*, Q-Exactive; Thermo Scientific) and sample fractionation.

We are confident that the present data can be used to validate the existence of 210 missing or uncertain proteins and are looking forward to integration of these validations in neXtProt once they have been reanalyzed by PeptideAtlas and combined with data from the other C-HPP teams interested in the testis or sperm proteome. In the meantime, the investigation of the human sperm proteome continues in our laboratories together with extensive data mining on the remaining set of missing proteins presented in this study. The information gleaned will help to extend our knowledge on the potential roles of these proteins in sperm function and/or maturation. Indeed, some of the proteins identified here may present a high clinical potential, and could also benefit the Biology and Disease driven HPP (B/D-HPP) that aims to explore the impact of proteomic technologies applied to a focused area of life science and health [38].

Notes:

The authors declare no competing financial interest.

Tables

Table 1: Description of missing proteins (PE2-PE4) detected in this study.

Total number of missing proteins (PE2-PE4)	Missing proteins with at least two unique, non nested peptides >=9AA	Missing proteins with only one unique peptide >=9AA	Missing proteins with no annotated function in Uniprot	Missing proteins with at least one transmembrane domain
235	188	47	180	56

Table 2: List of missing proteins (PE2-4) identified in the five MS/MS datasets. Accession numbers and number of transmembrane helices (No. TMH) were retrieved from UniprotKB; Gene names and chromosome location are as referenced in neXtProt.

Entry	Gene names	no. of unique peptide	chromosome	neXtProt PE level	no. of TMH
A0AVI2	FER1L5	1	2q11.2	Evidence at transcript level	1
A1A4V9	CCDC189 C16orf93	10	16p11.2	Evidence at transcript level	0
A1L453	PRSS38 MPN2	2	1q42.13	Evidence at transcript level	0
A2RUU4	CLPSL1 C6orf127	2	6p21.31	Evidence at transcript level	0
A4D1F6	LRRD1	13	7q21.2	Evidence at transcript level	0
A4D256	CDC14C CDC14B2	1	7p12.3	Evidence at transcript level	1
A4QMS7	C5orf49	5	5p15.31	Evidence at transcript level	0
A5D8W1	CFAP69 C7orf63	25	7q21.13	Evidence at transcript level	0
A5PLK6	RGSL1 RGSL RGSL2	5	1q25.3	Evidence at transcript level	1
A6H8Z2	FAM221B C9orf128	7	9p13.3	Evidence at transcript level	0
A6NCJ1	C19orf71	10	19p13.3	Predicted	0
A6NCL2	LRCOL1	2	12q24.33	Evidence at transcript level	0
A6NCM1	IQCA1L IQCA1P1	2	7q36.1	Inferred from homology	0
A6NCN8		3	12p13.33	Predicted	0
A6NE01	FAM186A	6	12q13.12	Evidence at transcript level	0
A6NE52	WDR97 KIAA1875	11	8q24.3	Evidence at transcript level	0
A6NEN9	CXorf65	4	Xq13.1	Predicted	0
A6NF34	ANTXRL	8	10q11.22	Inferred from homology	1
A6NFU0	FAM187A	4	17q21.31	Inferred from homology	1
A6NFZ4	FAM24A	4	10q26.13	Inferred from homology	0
A6NGB0	TMEM191C	1	22q11.21	Inferred from homology	1
A6NGY3	C5orf52	2	5q33.3	Evidence at transcript level	0

A6NI87	CBY3	3	5q35.3	Inferred from homology	0
A6NIV6	LRRIQ4 LRRC64	14	3q26.2	Inferred from homology	0
A6NJI9	LRRC72	5	7p21.1	Evidence at transcript level	0
A6NLX4	TMEM210	1	9q34.3	Inferred from homology	1
A6NM11	LRRC37A2	3	17q21.31	Evidence at transcript level	1
A6NMS7	LRRC37A LRRC37A1	2	17q21.31	Evidence at transcript level	1
A6NN90	C2orf81	9	2p13.1	Inferred from homology	0
A6NNE9	MARCH11	1	5p15.1	Evidence at transcript level	2
A6NNW6	ENO4 C10orf134	8	10q25.3	Evidence at transcript level	0
A6NNX1	RIIAD1 C1orf230	3	1q21.3	Predicted	0
A7E2U8	C4orf47 Chr4_1746	12	4q35.1	Evidence at transcript level	0
A8MTL0	IQCF5	3	3p21.2	Evidence at transcript level	0
A8MTZ7	C12orf71	2	12p11.23	Predicted	0
A8MV24	C17orf98	5	17q12	Predicted	0
A8MYZ5	IQCF6	5	3p21.2	Inferred from homology	0
A8MZ26	EFCAB9	5	5q35.1	Inferred from homology	0
B1AJZ9	FHAD1 KIAA1937	2	1p36.21	Evidence at transcript level	0
B1ANS9	WDR64	35	1q43	Evidence at transcript level	0
B2RC85	RSPH10B2	1	7p22.1	Evidence at transcript level	0
B2RV13	C17orf105	4	17q21.31	Evidence at transcript level	0
B4DYI2	SPATA31C2 FAM75C2	7	9q22.1	Evidence at transcript level	1
C9J6K1	C19orf81	1	19q13.33	Predicted	0
D6REC4	CFAP99	1	4p16.3	Inferred from homology	0
H3BNL8	C6orf229	5	6p22.3	Predicted	0
H3BTG2-2	C1orf234	4	1p36.12	Evidence at transcript level	0
O95214	LEPROTL1 My047 UNQ577/PRO1139	1	8p12	Evidence at transcript level	4
O95473	SYNGR4	9	19q13.33	Evidence at transcript level	4

P0C221	CCDC175 C14orf38	13	14q23.1	Predicted	0
P0C5Z0	H2AFB2; H2AFB3 H2ABBD H2AFB	1	Xq28	Evidence at transcript level	0
P0C7A2	FAM153B	1	5q35.2	Evidence at transcript level	0
P0C7I6	CCDC159	9	19p13.2	Evidence at transcript level	0
P0C7M6	IQCF3	5	3p21.2	Evidence at transcript level	0
P0C7X4	FTH1P19 FTHL19	1	Xp21.1	Uncertain	0
P0C874	SPATA31D3 FAM75D3	2	9q21.32	Uncertain	1
P0C875	FAM228B	6	2p23.3	Evidence at transcript level	0
P0C881	RSPH10B	1	7p22.1	Evidence at transcript level	0
P0C8F1	PATE4	8	11q24.2	Evidence at transcript level	0
P0CW27	CCDC166	6	8q24.3	Predicted	0
P0DJG4	THEGL	10	4q12	Evidence at transcript level	0
P0DKV0	SPATA31C1 FAM75C1	7	9q22.1	Evidence at transcript level	1
P49223	SPINT3	3	20q13.12	Inferred from homology	0
Q08648-4		3	8p23.1	Evidence at transcript level	
Q0P670	C17orf74	13	17p13.1	Evidence at transcript level	1
Q0VAA2	LRRC74A C14orf166B LRRC74	18	14q24.3	Evidence at transcript level	0
Q14409	GK3P GKP3 GKTB	4	4q32.3	Uncertain	0
Q14507	EDDM3A FAM12A HE3A	3	14q11.2	Evidence at transcript level	0
Q17R55	FAM187B TMEM162	6	19q13.12	Evidence at transcript level	1
Q2TAA8	TSNAXIP1 TXI1	20	16q22.1	Evidence at transcript level	0
Q2WGJ8	TMEM249 C8orfK29	3	8q24.3	Evidence at transcript level	2
Q32M84	BTBD16 C10orf87	20	10q26.13	Evidence at transcript level	0
Q3KNT9	TMEM95 UNQ9390/PRO34281	1	17p13.1	Evidence at transcript level	1
Q3SY17	SLC25A52 MCART2	1	18q12.1	Evidence at transcript level	6
Q3ZCV2	LEXM LEM C1orf177	9	1p32.3	Evidence at transcript level	0
Q494V2	CCDC37	7	3q21.3	Evidence at transcript level	0

Q495T6	MMEL1 MELL1 MMEL2 NEP2	24	1p36.32	Evidence at transcript level	1
Q499Z3	SLFNL1	18	1p34.2	Evidence at transcript level	0
Q4G0N8	SLC9C1 SLC9A10	4	3q13.2	Evidence at transcript level	16
Q4G1C9	GLIPR1L2	10	12q21.2	Evidence at transcript level	1
Q4ZJI4	SLC9B1 NHEDC1	3	4q24	Evidence at transcript level	1
Q502W6	VWA3B	7	2q11.2	Evidence at transcript level	0
Q502W7	CCDC38	15	12q23.1	Evidence at transcript level	0
Q537H7	SPATA45 C1orf227 HSD-44 HSD44	3	1q32.3	Inferred from homology	0
Q53FE4	C4orf17	2	4q23	Evidence at transcript level	0
Q53SZ7	PRR30 C2orf53	12	2p23.3	Evidence at transcript level	0
Q58FF6	HSP90AB4P	2	15q21.3	Uncertain	0
Q5BJE1	CCDC178 C18orf34	3	18q12.1	Evidence at transcript level	0
Q5GAN3	RNASE13	6	14q11.2	Evidence at transcript level	0
Q5GH77	XKR3 XRG3	2	22q11.1	Evidence at transcript level	10
Q5H913	ARL13A	3	Xq22.1	Evidence at transcript level	0
Q5H9T9	FSCB C14orf155	8	14q21.2	Evidence at transcript level	0
Q5I0G3	MDH1B	14	2q33.3	Evidence at transcript level	0
Q5JRC9	FAM47A	3	Xp21.1	Evidence at transcript level	0
Q5JU00	TCTE1	10	6p21.1	Evidence at transcript level	0
Q5JU67	C9orf117	17	9q34.11	Evidence at transcript level	0
Q5JWF8	ACTL10 C20orf134	7	20q11.22	Evidence at transcript level	0
Q5SQS8	C10orf120	7	10q26.13	Evidence at transcript level	0
Q5SY80	C1orf101	11	1q44	Evidence at transcript level	1
Q5T0J7	Tex35 C1orf49	6	1q25.2	Evidence at transcript level	0
Q5T1A1	DCST2	2	1q21.3	Evidence at transcript level	6
Q5T1B0	AXDND1 C1orf125	19	1q25.2	Evidence at transcript level	0
Q5T7R7	C1orf185	4	1p32.3	Evidence at transcript level	1

Q5TBE3	C9orf153	2	9q21.33	Predicted	0
Q5TEZ5	C6orf163	22	6q15	Predicted	0
Q5TFG8	ZC2HC1B C6orf94 FAM164B	2	6q24.2	Evidence at transcript level	0
Q5TGP6-2	MROH9 C1orf129	3	1q24.3	Evidence at transcript level	0
Q5VTH9	WDR78	21	1p31.3	Evidence at transcript level	0
Q5VZ72	IZUMO3 C9orf134	8	9p21.3	Inferred from homology	1
Q5VZQ5	TEX36 C10orf122	6	10q26.13	Evidence at transcript level	0
Q5XX13	FBXW10	2	17p11.2	Evidence at transcript level	0
Q63HN1	FAM205BP C9orf144 C9orf144A FAM205B	1	9p13.3	Uncertain	0
Q68DN1	C2orf16	62	2p23.3	Evidence at transcript level	0
Q68G75	LEMD1	2	1q32.1	Evidence at transcript level	1
Q6ICG8	WBP2NL PAWP	11	22q13.2	Evidence at transcript level	0
Q6IPT2-2	FAM71E1	2	19q13.33	Evidence at transcript level	0
Q6NXN4	DPY19L2P1	1	7p14.2	Evidence at transcript level	3
Q6NXP6	NOXRED1 C14orf148	2	14q24.3	Evidence at transcript level	0
Q6P2C0	WDR93	6	15q26.1	Evidence at transcript level	0
Q6P2D8	XRRA1	6	11q13.4	Evidence at transcript level	0
Q6PDA7-2	SPAG11A EP2 HE2	1	8p23.1	Evidence at transcript level	0
Q6PI97	C11orf88	3	11q23.1	Evidence at transcript level	0
Q6PIY5	C1orf228 NCRNA00082	14	1p34.1	Evidence at transcript level	0
Q6UW60	PCSK4 PC4 UNQ2757/PRO6496	1	19p13.3	Evidence at transcript level	1
Q6UWQ5	LYZL1 LYC2 UNQ648/PRO1278	2	10p11.23	Evidence at transcript level	0
Q6UXN7	TOMM20L UNQ9438/PRO34772	1	14q23.1	Evidence at transcript level	1
Q6V702	C4orf22	10	4q21.21	Evidence at transcript level	0
Q6ZMY6	WDR88 PQWD	4	19q13.11	Evidence at transcript level	0
Q6ZNQ3	LRRC69	1	8q21.3	Evidence at transcript level	0
Q6ZRH7	CATSPERG C19orf15	15	19q13.2	Evidence at transcript level	1

Q6ZUB0	SPATA31D4 FAM75D4	1	9q21.32	Uncertain	1
Q6ZUB1	SPATA31E1 C9orf79 FAM75E1	44	9q22.1	Evidence at transcript level	1
Q6ZUG5		10	3q21.3	Evidence at transcript level	0
Q6ZVS7	FAM183B	4	7p14.1	Evidence at transcript level	0
Q7RTY9	PRSS41 TESSP1	1	16p13.3	Uncertain	0
Q7Z2V1	C16orf82	2	16p12.1	Evidence at transcript level	0
Q7Z4T8	GALNTL5 GALNT15	1	7q36.1	Evidence at transcript level	1
Q7Z4W2	LYZL2	2	10p11.23	Evidence at transcript level	0
Q7Z5J8	ANKAR	15	2q32.2	Evidence at transcript level	1
Q7Z7B7	DEFB132 DEFB32 UNQ827/PRO1754	1	20p13	Inferred from homology	0
Q86TZ1-2	TTC6	3	14q21.1	Evidence at transcript level	0
Q86UG4	SLCO6A1 OATP6A1 SLC21A19	15	5q21.1	Evidence at transcript level	12
Q86VE3	SATL1	2	Xq21.1	Evidence at transcript level	0
Q86VS3	IQCH	9	15q23	Evidence at transcript level	0
Q86WZ0	HEATR4	3	14q24.3	Evidence at transcript level	0
Q86X67	NUDT13	1	10q22.2	Evidence at transcript level	0
Q8IUB5	WFDC13 C20orf138 WAP13	2	20q13.12	Inferred from homology	0
Q8IVL8	CPO	3	2q33.3	Evidence at transcript level	0
Q8IVU9	C10orf107	7	10q21.2	Evidence at transcript level	0
Q8IWF9	CCDC83 HSD9	2	11q14.1	Evidence at transcript level	0
Q8IXM7	ODF3L1	12	15q24.2	Evidence at transcript level	0
Q8IXW0	LMNTD2 C11orf35	10	11p15.5	Evidence at transcript level	0
Q8IYJ2	C10orf67	4	10p12.2	Evidence at transcript level	0
Q8IYM0	FAM186B C12orf25	15	12q13.12	Evidence at transcript level	0
Q8IYU4	UBQLNL	1	11p15.4	Evidence at transcript level	0
Q8IYW2	CFAP46 C10orf123 C10orf124 C10orf92 C10orf93 TTC40	34	10q26.3	Evidence at transcript level	0
Q8N0W5	IQCK	9	16p12.3	Evidence at transcript level	0

Q8N309	LRRC43	10	12q24.31	Evidence at transcript level	0
Q8N456	LRRC18 UNQ9338/PRO34010	7	10q11.23	Evidence at transcript level	0
Q8N4B4	FBXO39 FBX39	1	17p13.1	Evidence at transcript level	0
Q8N4L4	SPEM1 C17orf83	10	17p13.1	Evidence at transcript level	1
Q8N4P6	LRRC71 C1orf92	15	1q23.1	Evidence at transcript level	0
Q8N5S1	SLC25A41	6	19p13.3	Evidence at transcript level	6
Q8N5S3	C2orf73	4	2p16.2	Evidence at transcript level	0
Q8N5U0	C11orf42	3	11p15.4	Evidence at transcript level	0
Q8N5W8	FAM24B	1	10q26.13	Inferred from homology	0
Q8N688	DEFB123 DEFB23 UNQ1963/PRO4485	1	20q11.21	Evidence at transcript level	0
Q8N6G2	TEX26 C13orf26	9	13q12.3	Evidence at transcript level	0
Q8N6K0	TEX29 C13orf16	3	13q34	Evidence at transcript level	1
Q8N6M8	IQCF1	6	3p21.2	Evidence at transcript level	0
Q8N6V4	C10orf53	1	10q11.23	Inferred from homology	0
Q8N7B9	EFCAB3	20	17q23.2	Evidence at transcript level	0
Q8N7C7	RNF148	1	7q31.32	Evidence at transcript level	1
Q8N7X0	ADGB C6orf103 CAPN7L	32	6q24.3	Evidence at transcript level	0
Q8N7X2-4	C9orf173	9	9q34.3	Evidence at transcript level	0
Q8N801	C2orf61	8	2p21	Evidence at transcript level	0
Q8N9W8	FAM71D	1	14q23.3	Evidence at transcript level	0
Q8N9Z9	LMNTD1 IFLTD1	1	12p12.1	Evidence at transcript level	0
Q8NA56	TTC29	13	4q31.22	Evidence at transcript level	0
Q8NA66	CNBD1	1	8q21.3	Evidence at transcript level	0
Q8NA69	C19orf45	10	19p13.2	Evidence at transcript level	0
Q8NCQ7	PROCA1	5	17q11.2	Evidence at transcript level	0
Q8NCU1	LINC00521	1	14q32.12	Uncertain	0
Q8ND07	BBOF1 C14orf45 CCDC176	7	14q24.3	Evidence at transcript level	0

Q8ND61	C3orf20	8	3p25.1	Evidence at transcript level	1
Q8NDH2	CCDC168 C13orf40	9	13q33.1	Evidence at transcript level	0
Q8NE28	STKLD1 C9orf96 SGK071	14	9q34.2	Evidence at transcript level	0
Q8NEA5	C19orf18	4	19q13.43	Evidence at transcript level	1
Q8NEE8	TTC16	7	9q34.11	Evidence at transcript level	0
Q8NEX6	WFDC11 WAP11	1	20q13.12	Inferred from homology	0
Q8NG35	DEFB105A BD5 DEFB105 DEFB5; DEFB105B	2	8p23.1	Evidence at transcript level	0
Q8NHS2	GOT1L1	6	8p11.23	Evidence at transcript level	0
Q8NHU2	CFAP61 C20orf26	33	20p11.23	Evidence at transcript level	0
Q8NHX4	SPATA3 TSARG1	5	2q37.1	Evidence at transcript level	0
Q8TBY8	PMFBP1	78	16q22.2	Evidence at transcript level	0
Q8TBZ9	C7orf62	12	7q21.13	Evidence at transcript level	0
Q8TD35	LKAAEAR1 C20orf201	2	20q13.33	Evidence at transcript level	0
Q8WTQ4	C16orf78	2	16q12.1	Evidence at transcript level	0
Q8WVZ1	ZDHHC19	2	3q29	Evidence at transcript level	4
Q8WVZ7	RNF133	3	7q31.32	Evidence at transcript level	1
Q8WW18	C17orf50	1	17q12	Evidence at transcript level	0
Q8WWF3	SSMEM1 C7orf45	6	7q32.2	Evidence at transcript level	1
Q8WXQ8	CPA5	10	7q32.2	Evidence at transcript level	0
Q96E66	LRTOMT LRRC51	12	11q13.4	Evidence at transcript level	0
Q96KW9	SPACA7 C13orf28	3	13q34	Evidence at transcript level	0
Q96L03	SPATA17	8	1q41	Evidence at transcript level	0
Q96L15	ART5 UNQ575/PRO1137	4	11p15.4	Evidence at transcript level	0
Q96LI9	CXorf58	1	Xp22.11	Evidence at transcript level	0
Q96LM5	C4orf45	5	4q32.1	Evidence at transcript level	0
Q96LU5	IMMP1L	3	11p13	Evidence at transcript level	0
Q96M20	CNBD2 C20orf152	6	20q11.23	Evidence at transcript level	0

Q96M60	FAM227B C15orf33	1	15q21.2	Evidence at transcript level	0
Q96M69	LRGUK	11	7q33	Evidence at transcript level	0
Q96M83	CCDC7 BIOT2	7	10p11.22	Evidence at transcript level	0
Q96M86	DNHD1 C11orf47 CCDC35 DHCD1 DNHD1L UNQ5781/PRO12970	13	11p15.4	Evidence at transcript level	0
Q96N23	CFAP54 C12orf55 C12orf63	38	12q23.1	Evidence at transcript level	0
Q96PP4	TSGA13	4	7q32.2	Evidence at transcript level	0
Q96SF2	CCT8L2 CESK1	5	22q11.1	Evidence at transcript level	0
Q9BYW3	DEFB126 C20orf8 DEFB26	5	20p13	Evidence at transcript level	0
Q9BZ19	ANKRD60 C20orf86	1	20q13.32	Inferred from homology	0
Q9BZJ4	SLC25A39 CGI-69 PRO2163	3	17q21.31	Evidence at transcript level	6
Q9GZN6	SLC6A16 NTT5	6	19q13.33	Evidence at transcript level	12
Q9H1M3	DEFB129 C20orf87 DEFB29 UNQ5794/PRO19599	3	20p13	Evidence at transcript level	0
Q9H1P6	C20orf85	5	20q13.32	Evidence at transcript level	0
Q9H1U9	SLC25A51 MCART1	1	9p13.1	Evidence at transcript level	6
Q9H3M9	ATXN3L ATX3L MJDL	7	Xp22.2	Uncertain	0
Q9H3V2	MS4A5 CD20L2 TETM4	1	11q12.2	Evidence at transcript level	4
Q9H3Z7	ABHD16B C20orf135	2	20q13.33	Inferred from homology	0
Q9H579-2	MROH8 C20orf131 C20orf132	18	20q11.23	Evidence at transcript level	0
Q9H5F2	C11orf1	5	11q23.1	Evidence at transcript level	0
Q9H693	C16orf95	1	16q24.2	Evidence at transcript level	0
Q9H7T0	CATSPERB C14orf161	22	14q32.12	Evidence at transcript level	4
Q9H8X9	ZDHHC11 ZNF399	2	5p15.33	Evidence at transcript level	4
Q9H943	C10orf68	1	10p11.22	Evidence at transcript level	0
Q9NUD7	C20orf96	1	20p13	Evidence at transcript level	0
Q9NZM6	PKD2L2	2	5q31.2	Evidence at transcript level	6
Q9P1V8	SAMD15 C14orf174 FAM15A	19	14q24.3	Evidence at transcript level	0
Q9P1Z9-2	CCDC180 C9orf174 KIAA1529	22	9q22.33	Evidence at transcript level	1

Q9P2S6	ANKMY1 TSAL1 ZMYND13	19	2q37.3	Evidence at transcript level	0
Q9UKJ8	ADAM21	3	14q24.2	Evidence at transcript level	1
Q9ULG3	KIAA1257	1	3q21.3	Evidence at transcript level	0
Q9Y238	DLEC1 DLC1	22	3p22.2	Evidence at transcript level	0
Q9Y581	INSL6 RIF1	2	9p24.1	Evidence at transcript level	0
W5XKT8	SPACA6 SPACA6P UNQ2487/PRO5774	2	19q13.41	Evidence at transcript level	1

Figure legends

Figure 1: Flowchart illustrating the strategies used to identify and validate missing proteins detected in the human sperm proteome.

Figure 2: Contribution of the different fractionation protocols to identification of spermatozoa proteins. Upper part: Venn diagram created with the jvenn web application [39] illustrating overlap between the five fractionation protocols. WL 1D gel: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), WL LR: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs, HpH-RP: total cell lysate, in-gel digestion of proteins, and peptide fractionation by high-pH reversed-phase (HpH-RP) chromatography, Soluble and Insoluble: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by a 1D SDS-PAGE separation of proteins (20 gel slices per fraction). Lower part: bar chart representing the total number of proteins identified in each MS/MS dataset.

Figure 3: Missing proteins detected in the sperm proteome. A. Venn diagram illustrating the overlap between the five different fractionation protocols and the contribution of each fraction to the detection of missing proteins. Venn diagram was created with the jvenn web application [39]. WL-1D-gel: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), WL-LGR: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs, WL-HP-RP: total cell lysate, in-gel digestion of proteins, and peptide fractionation by

high-pH reversed-phase (HpH-RP) chromatography, Soluble and Insoluble: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by a 1D SDS-PAGE separation of proteins (20 gel slices per fraction). Lower part: bar charts representing the number of missing proteins identified in each MS/MS dataset. Under the bar chart: information related to the number of missing proteins identified specifically in each (1) or shared by 2, 3, 4 or all 5 datasets. B. Distribution of missing proteins according to the chromosomal location of their genes (retrieved from neXtProt). C. Distribution of missing proteins according to the number of proteotypic peptides: numbers beside each portion of the pie indicate the number of unique peptides that mapped onto missing proteins (max. number of unique peptides: 56; see Table 2 and Supplementary Table 3 for details).

Figure 4: MS validation of one-hit wonder missing protein, example of protein Q9H693 (C16orf95). **A.** Missing protein sequence with the unique peptide identified highlighted in blue; MS/MS spectrum of the endogenous peptide correlated with the MS/MS spectrum of its labelled synthetic counterpart and spectral correlation score calculated (SDPscore). **B.** MS2 traces extracted from LC-PRM data from an unfractionated total protein extract with dotp calculated for the endogenous and labelled peptides and rdotp calculated for light/heavy correlation. **C.** MS2 traces extracted from LC-PRM data from an unfractionated total protein extract. Analysis was targeted to detect an additional predicted peptide for protein Q9H693. The dotp for the endogenous and labelled peptide and the rdotp for light/heavy correlation were calculated.

Figure 5A: Antibody staining for orphan proteins CXorf58 (Q96LI9), C19orf81 (C9J6K1), C17orf105 (B2RV13), C20orf85 (Q9H1P6), C10orf53 (Q8N6V4), C10orf67 (Q8IYJ2), FAM187B (Q17R55) and SSMEM1 (Q8WWF3) in adult human testis. Proteins were detected in transverse testis sections at stages IV to VI of the seminiferous epithelium [40] using polyclonal antibodies from the Human Protein Atlas specific for CXorf58 (HPA031543) (A), C19orf81 (HPA060238) (B), C17orf105 (HPA053028) (C), C20orf85 (HPA058271) (D), C10orf53 (HPA037951) (E), C10orf67 (HPA038131) (F), FAM187B (HPA014687) (G), SSMEM1 (HPA026877) (H). Non-immune serum was used as a negative control (data not shown). In all testis sections, a more or less intense antibody staining signal was visible in germ cells at all stages and for all proteins (A to H). CXorf58 immunoreactivity was very strong in the headpiece of late spermatids (A; arrow). C19orf81 presented strong staining in the cytoplasm of spermatogonia (arrowhead) and of late spermatids (B; arrow). Intense C17orf105 staining was visible in the cytoplasm of late spermatids (C; arrow). C20orf85 immunoreactivity displayed as an intense granular staining in pachytene spermatocytes (arrowhead) and concentrated with a very strong signal in the acrosome of elongating spermatids (arrows) (D). C10orf53 immunoreactivity appeared concentrated in the cytoplasm of late spermatids (E; arrows). A punctiform signal was visible for C10orf67 in the cytoplasm of elongating spermatids (F; arrows). FAM187B immunoreactivity concentrated in the cytoplasm of elongating spermatids (G; arrows). SSMEM1 displayed intense staining in pachytene spermatocytes (arrowhead) and in elongating spermatids (arrows) (H). Scale bars = 20 μm.

Figure 5B: Antibody staining for CCT8L2 (Q96SF2), AXDND1 (Q5T1B0), WDR88 (Q6ZMY6), GOT1L1 (Q8NHS2), CFAP46 (Q8IYW2), SAMD15 (Q9P1V8), WDR93 (Q6P2C0) and NOXRED1 (Q6NXP6) in adult human testis. Proteins were detected in transverse testis sections at stages IV to VI of the seminiferous epithelium [40] using polyclonal antibodies from the Human Protein Atlas specific for CCT8L2 (HPA039268) (A), AXDND1 (HPA071114) (B), WDR88 (HPA041916) (C), GOT1L1 (HPA028778) (D), CFAP46 (HPA038034) (E), SAMD15 (HPA030673) (F), WDR93 (HPA048112) (G) and NOXRED1 (HPA055658) (H). Nonimmune serum was used as a negative control (data not shown). In all testis sections, a more or less intense signal for all proteins was visible in germ cells at all stages (A to H). CCT8L2 staining was clearly cytoplasmic in pachytene spermatocytes (arrowhead) and late spermatids (arrow) (A). AXDND1 immunoreactivity was intense in the cytoplasm of late spermatids (arrow; B). WDR88 immunoreactivity concentrated with a very strong signal in the cytoplasm of elongating spermatids (arrow; C). A very strong GOT1L1 immunostaining was observed in the cytoplasm of late spermatids (D; arrow). CFAP46 immunoreactivity was very strong and ring-shaped in the headpiece of late spermatids (arrow; E). SAMD15 immunoreactivity was intense in pachytene spermatocytes (arrow; F). A strong immunoreactivity was observed for WDR93 in the cytoplasm of pachytene spermatocytes (arrowhead) and in the cytoplasm of elongating spermatids (arrow) (G). NOXRED1 immunoreactivity was exclusively cytoplasmic and the signal increased from premeiotic germ cells to pachytene spermatocytes (arrowhead) and round spermatids (arrows) (H). Scale bars = 20 μm.

Supporting Information:

This material is available free of charge via http://pubs.acs.org/.

Supplementary Material: 1. Shotgun LC-MS/MS analyses. 2. LC-MS/MS analysis of labeled synthetic peptides and comparison of fragmentation spectra. 3. How targeted LC-PRM assays were developed (details).

Supplementary Figure 1: MS/MS spectra for the 36 endogenous peptides and their synthetic reference counterparts combined with LC-PRM results for the 36 peptides, additional predicted proteotypic peptides and their labeled synthetic counterparts. MS validation results are presented for all 36 selected one-hit wonder proteins in the same way as data for protein Q9H693 are presented in Figure 4 of the manuscript. A. Missing protein sequence with the unique identified peptide highlighted in blue; MS/MS spectrum of the endogenous peptide correlated with the MS/MS spectrum of its labeled synthetic counterpart. B. MS2 traces extracted from LC-PRM data with dotp calculated for the endogenous and labeled peptides and rdotp calculated for light/heavy correlation. C. MS2 traces extracted from LC-PRM data targeting additional predicted peptides. The dotp for the endogenous and labeled peptide(s) and the rdotp for light/heavy correlation were calculated.

Supplementary Table 1: PSM-, peptide-, and protein-level FDR values along with the total number of expected true- and false-positives at each level for each sperm proteome dataset and the combined dataset (tab 1: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), tab 2: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs,

tab 3: total cell lysate, in-gel digestion of proteins, and peptide fractionation by high-pH reversed-phase (HpH-RP) chromatography; tabs 4 and 5: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by 1D SDS-PAGE separation of proteins (20 gel slices per fraction); tab 6: combined dataset corresponding to the combination of results for the five proteome datasets).

Supplementary Table 2: List of proteins identified and validated with a protein-level 1% FDR for each fraction (detailed information): tab 1: total cell lysate followed by a 1D SDS-PAGE separation of proteins (23 gel slices), tab 2: total cell lysate, in-gel digestion of proteins, and total peptide analysis by nanoLC with long gradient runs, tab 3: total cell lysate, in-gel digestion of proteins, and peptide fractionation by high-pH reversed-phase (HpH-RP) chromatography; tabs 4 and 5: fractionation of proteins into Triton X-100-soluble and -insoluble fractions, followed by 1D SDS-PAGE separation of proteins (20 gel slices per fraction).

Supplementary Table 3: Missing (PE2-4) and uncertain (PE5) proteins detected in the sperm proteome: detailed information. Accession numbers, entry description, molecular weight (MW), protein length (length) number of transmembrane domains (No. TMH), subcellular location and function (CC field) were retrieved from UniprotKB; Gene names and chromosome location are as referenced in neXtProt. Coverage (protein coverage in %) and number of unique peptides mapping to missing proteins, proteins seen (yes)/not seen (not) in each of the five MS/MS datasets acquired in this study are reported. The expression annotations of the Human testis gene expression program (TGEP; [5]) were also reported when available

(SET: proteins produced by genes specifically expressed in the testis; PET: proteins produced by genes preferentially expressed in the testis; IE: proteins produced by genes with intermediate expression in the testis; UE: proteins produced by genes with ubiquitous expression in the testis). The « testis-enriched gene » status in the Human Protein Atlas version 15 is also provided.

Supplementary Table 4: Tab1: List of one-hit wonder missing proteins identified and selected for further spectral comparison (MS/MS) and PRM validation. Tab2: Extended list of 100 peptides selected for validation of the 36 one hit wonders and synthesized as crude labeled peptides.

Supplementary Table 5: List of 111 PE2-5 proteins for which complete data mining was performed, showing the rationale for their prioritization for subsequent antibody-based studies. Entry accession numbers (column A) and gene names (column B) were retrieved from UniProtKB, transcript abundance was retrieved from HPA (column C), phylogenetic profiles in ciliated organisms were determined by Blast analysis on UniProtKB “Reference proteomes” (column D), knock-out mice phenotypes were retrieved from MGI (column E), associated publications that were not annotated in neXtProt were searched in PubMed (column F), and immunohistochemistry data was retrieved from HPA (column G). For all these criteria, a four-color grading system was adopted. Dark green and green cells were retained as positive criteria, red cells as negative ones. Based on these criteria, a score of relevance (high/medium/low) for the implication of the proteins in spermatogenesis has been assigned (column H). The existence of a suitable

antibody in HPA is reported column I, and the list of proteins that were finally selected for IHC is provided in column J.

Supplementary Table 6: List of missing proteins for which IHC was successful, HPA antibody names and dilutions used.

Supplementary Table 7: List of the 94 missing proteins detected in Jumeau *et al.* (2015) [14] and confirmed in the present study. The number of peptides identified in each study is reported in columns C and E. The protein existence (PE) status of the entries in the 2015 and 2016 neXtProt reference releases is reported in columns B and F.

Authors' contributions: YV, LL and CP co-coordinated the study. YV, LL, CP, AGP, CC, CB, SC, OBS, MF and JG conceived and designed the experiments and analyses. TF and KR performed spermatozoa preparation. CC, CM, AGP, YC, MM and KC performed the sample preparation and MS/MS analysis. YV, CB, AMH, EMB, LL, TR and AG processed and analyzed MS/MS datasets. CC and CM performed MS/MS spectra comparison and PRM assays. PD, YV, CP, AG, LM, EC and LL performed bioinformatics analysis and data/literature mining on identified proteins and selected candidates for IHC studies. Immunohistochemical studies were done by KR, CP and CL. YV, LL, PD, CC, CM, CP prepared the figures, tables and supporting information. YV, LL, CC, CP drafted the manuscript. All the authors approved the final version of the manuscript.

Acknowledgements:

This work was partially funded through the French National Agency for Research (ANR) (grant ANR-10-INBS-08; ProFI project, “Infrastructures Nationales en Biologie et Santé”; “Investissements d’Avenir” call). This work was also supported by grants from Biogenouest and *Conseil Régional de Bretagne* awarded to CP. We are grateful to Monique Zahn and Maighread Gallagher-Gambarelli for suggestions on language usage. We thank Marine Seffals for technical support with immunohistochemistry experiments on the H2P2 core facility (Université de Rennes 1, US18, UMS3480 Biosit, Biogenouest, Rennes, France). We particularly thank Blandine Guével, Véronique Dupierris, Mélanie Lagarrigue, Jean-Philippe Menetrey, Mathieu Schaeffer, Régis Lavigne and Christine Kervarrec for their technical assistance.

References

1. 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.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S., The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature biotechnology* **2012,** *30* (3), 221-3.

2. Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Teixeira, D.; Zhang, Y.; Lane, L.; Bairoch, A., The neXtProt knowledgebase on human proteins: current status. *Nucleic acids research* **2015,** *43* (Database issue), D764-70.

3. Omenn, G. S.; Lane, L.; Lundberg, E. K.; Beavis, R. C.; Nesvizhskii, A. I.; Deutsch, E. W., Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *Journal of proteome research* **2015,** *14* (9), 3452-60.

4. 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. *Journal of proteome research* **2014,** *13* (1), 15-20.

5. 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. *Human reproduction* **2012,** *27* (11), 3233-48.

6. Uhlen, M.; Fagerberg, L.; Hallstrom, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, A.; Kampf, C.; Sjostedt, 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.; Ponten, F., Proteomics. Tissue-based map of the human proteome. *Science* **2015,** *347* (6220), 1260419.

7. Uhlen, M.; Hallstrom, B. M.; Lindskog, C.; Mardinoglu, A.; Ponten, F.; Nielsen, J., Transcriptomics resources of human tissues and organs. *Molecular systems biology* **2016,** *12* (4), 862.

8. 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), 211-7.

9. Eddy, E. M., Male germ cell gene expression. *Recent progress in hormone research* **2002,** *57*, 103-28.

10. Jégou, B.; Pineau, C.; Dupaix, A., Paracrine control of testis function. In *Male Reproductive Function*, Wang, C., Ed. Kluwer Academic: Berlin, 1999; pp 41-64.

11. Krausz, C., Male infertility: pathogenesis and clinical diagnosis. *Best practice & research. Clinical endocrinology & metabolism* **2011,** *25* (2), 271-85.

12. Rolland, A. D.; Jégou, B.; Pineau, C., Testicular development and spermatogenesis: harvesting the postgenomics bounty. *Advances in experimental medicine and biology* **2008,** *636*, 16-41.

13. Carapito, C.; Lane, L.; Benama, M.; Opsomer, A.; Mouton-Barbosa, E.; Garrigues, L.; Gonzalez de Peredo, A.; Burel, A.; Bruley, C.; Gateau, A.; Bouyssie, D.; Jaquinod, M.; Cianferani, S.; Burlet-Schiltz, O.; Van Dorsselaer, A.; Garin, J.; Vandenbrouck, Y., Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *Journal of proteome research* **2015,** *14* (9), 3621-34.

14. 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), 3606-20.

15. Vizcaino, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Rios, 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-Bartolome, 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), 223-6.

16. Com, E.; Rolland, A. D.; Guerrois, M.; Aubry, F.; Jégou, B.; Vallet-Erdtmann, V.; Pineau, C., Identification, molecular cloning, and cellular distribution of the rat homolog of minichromosome maintenance protein 7 (MCM7) in the rat testis. *Molecular reproduction and development* **2006,** *73* (7), 866-77.

17. Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C., Orthogonality of separation in two-dimensional liquid chromatography. *Analytical chemistry* **2005,** *77* (19), 6426-34.

18. Kitata, R. B.; Dimayacyac-Esleta, B. R.; Choong, W. K.; Tsai, C. F.; Lin, T. D.; Tsou, C. C.; Weng, S. H.; Chen, Y. J.; Yang, P. C.; Arco, S. D.; Nesvizhskii, A. I.; Sung, T. Y.; Chen, Y. J., Mining Missing Membrane Proteins by High-pH Reverse-Phase StageTip Fractionation and Multiple Reaction Monitoring Mass Spectrometry. *Journal of proteome research* **2015,** *14* (9), 3658-69.

19. Elias, J. E.; Gygi, S. P., Target-decoy search strategy for mass spectrometry-based proteomics. *Methods in molecular biology* **2010,** *604*, 55-71.

20. Amaral, A.; Castillo, J.; Ramalho-Santos, J.; Oliva, R., The combined human sperm proteome: cellular pathways and implications for basic and clinical science. *Human reproduction update* **2014,** *20* (1), 40-62.

21. Yoneda, R.; Kimura, A. P., A testis-specific serine protease, Prss41/Tessp-1, is necessary for the progression of meiosis during murine in vitro spermatogenesis. *Biochemical and biophysical research communications* **2013,** *441* (1), 120-5.

22. Weeks, S. D.; Grasty, K. C.; Hernandez-Cuebas, L.; Loll, P. J., Crystal structure of a Josephin-ubiquitin complex: evolutionary restraints on ataxin-3 deubiquitinating activity. *The Journal of biological chemistry* **2011,** *286* (6), 4555-65.

23. Buus, R.; Faronato, M.; Hammond, D. E.; Urbe, S.; Clague, M. J., Deubiquitinase activities required for hepatocyte growth factor-induced scattering of epithelial cells. *Current biology : CB* **2009,** *19* (17), 1463-6.

24. Sacco, J. J.; Yau, T. Y.; Darling, S.; Patel, V.; Liu, H.; Urbe, S.; Clague, M. J.; Coulson, J. M., The deubiquitylase Ataxin-3 restricts PTEN transcription in lung cancer cells. *Oncogene* **2014,** *33* (33), 4265-72.

25. Ge, F.; Chen, W.; Qin, J.; Zhou, Z.; Liu, R.; Liu, L.; Tan, J.; Zou, T.; Li, H.; Ren, G.; Chen, C., Ataxin-3 like (ATXN3L), a member of the Josephin family of deubiquitinating enzymes, promotes breast cancer proliferation by deubiquitinating Kruppel-like factor 5 (KLF5). *Oncotarget* **2015,** *6* (25), 21369-78.

26. Ye, D.; Fu, Y.; Sun, R. X.; Wang, H. P.; Yuan, Z. F.; Chi, H.; He, S. M., Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate. *Bioinformatics* **2010,** *26* (12), i399-406.

27. Liu, M.; Hu, Z.; Qi, L.; Wang, J.; Zhou, T.; Guo, Y.; Zeng, Y.; Zheng, B.; Wu, Y.; Zhang, P.; Chen, X.; Tu, W.; Zhang, T.; Zhou, Q.; Jiang, M.; Guo, X.; Zhou, Z.; Sha, J., Scanning of novel cancer/testis proteins by human testis proteomic analysis. *Proteomics* **2013,** *13* (7), 1200-10.

28. Wang, G.; Guo, Y.; Zhou, T.; Shi, X.; Yu, J.; Yang, Y.; Wu, Y.; Wang, J.; Liu, M.; Chen, X.; Tu, W.; Zeng, Y.; Jiang, M.; Li, S.; Zhang, P.; Zhou, Q.; Zheng, B.; Yu, C.; Zhou, Z.; Guo, X.; Sha, J., In-depth proteomic analysis of the human sperm reveals complex protein compositions. *Journal of proteomics* **2013,** *79*, 114-22.

29. Pilarski, L. M.; Gillitzer, R.; Zola, H.; Shortman, K.; Scollay, R., Definition of the thymic generative lineage by selective expression of high molecular weight isoforms of CD45 (T200). *European journal of immunology* **1989,** *19* (4), 589-97.

30. Wang, G.; Wu, Y.; Zhou, T.; Guo, Y.; Zheng, B.; Wang, J.; Bi, Y.; Liu, F.; Zhou, Z.; Guo, X.; Sha, J., Mapping of the N-linked glycoproteome of human spermatozoa. *Journal of proteome research* **2013,** *12* (12), 5750-9.

31. Rajender, S.; Rahul, P.; Mahdi, A. A., Mitochondria, spermatogenesis and male infertility. *Mitochondrion* **2010,** *10* (5), 419-28.

32. Hackett, N. R.; Butler, M. W.; Shaykhiev, R.; Salit, J.; Omberg, L.; Rodriguez-Flores, J. L.; Mezey, J. G.; Strulovici-Barel, Y.; Wang, G.; Didon, L.; Crystal, R. G., RNA-Seq quantification of the human small airway epithelium transcriptome. *BMC genomics* **2012,** *13*, 82.

33. Ross, A. J.; Dailey, L. A.; Brighton, L. E.; Devlin, R. B., Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *American journal of respiratory cell and molecular biology* **2007,** *37* (2), 169-85.

34. Strong, E. R.; Schimenti, J. C., Evidence Implicating CCNB1IP1, a RING Domain-Containing Protein Required for Meiotic Crossing Over in Mice, as an E3 SUMO Ligase. *Genes* **2010,** *1* (3), 440-51.

35. Brown, J. M.; Dipetrillo, C. G.; Smith, E. F.; Witman, G. B., A FAP46 mutant provides new insights into the function and assembly of the C1d complex of the ciliary central apparatus. *Journal of cell science* **2012,** *125* (Pt 16), 3904-13.

36. Geremek, M.; Zietkiewicz, E.; Bruinenberg, M.; Franke, L.; Pogorzelski, A.; Wijmenga, C.; Witt, M., Ciliary genes are down-regulated in bronchial tissue of primary ciliary dyskinesia patients. *PloS one* **2014,** *9* (2), e88216.

37. Georgiadis, A. P.; Kishore, A.; Zorrilla, M.; Jaffe, T. M.; Sanfilippo, J. S.; Volk, E.; Rajkovic, A.; Yatsenko, A. N., High quality RNA in semen and sperm: isolation, analysis and potential application in clinical testing. *The Journal of urology* **2015,** *193* (1), 352-9.

38. Aebersold, R.; Bader, G. D.; Edwards, A. M.; van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S., The biology/disease-driven human proteome project (B/D-HPP):

enabling protein research for the life sciences community. *Journal of proteome research* **2013,** *12* (1), 23-7.

39. Bardou, P.; Mariette, J.; Escudie, F.; Djemiel, C.; Klopp, C., jvenn: an interactive Venn diagram viewer. *BMC bioinformatics* **2014,** *15*, 293.

40. Clermont, Y., The cycle of the seminiferous epithelium in man. *The American journal of anatomy* **1963,** *112*, 35-51.

41. Nesvizhskii, A. I.; Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Molecular and Cellular Proteomics* **2005**, 4(10), 1419-1440.

42. Djureinovic, D.; Fagerberg, L.; Hallstrom, B.; Danielsson, A.; Lindskog, C.; Uhlen, M.; Ponten, F. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Molecular Human Reproduction* **2014**, 20(6), 476-488.

43. Petit, F. G.; Kervarrec, C.; Jamin, S. P.; Smagulova, F.; Hao, C.; Becker, E.; Jegou, B.; Chalmel, F.; Primig, M. Combining RNA and protein profiling data with network interactions identifies genes associated with spermatogenesis in mouse and

Figure 1. Flowchart illustrating the strategies used to identify and validate missing proteins detected in the human sperm proteome.

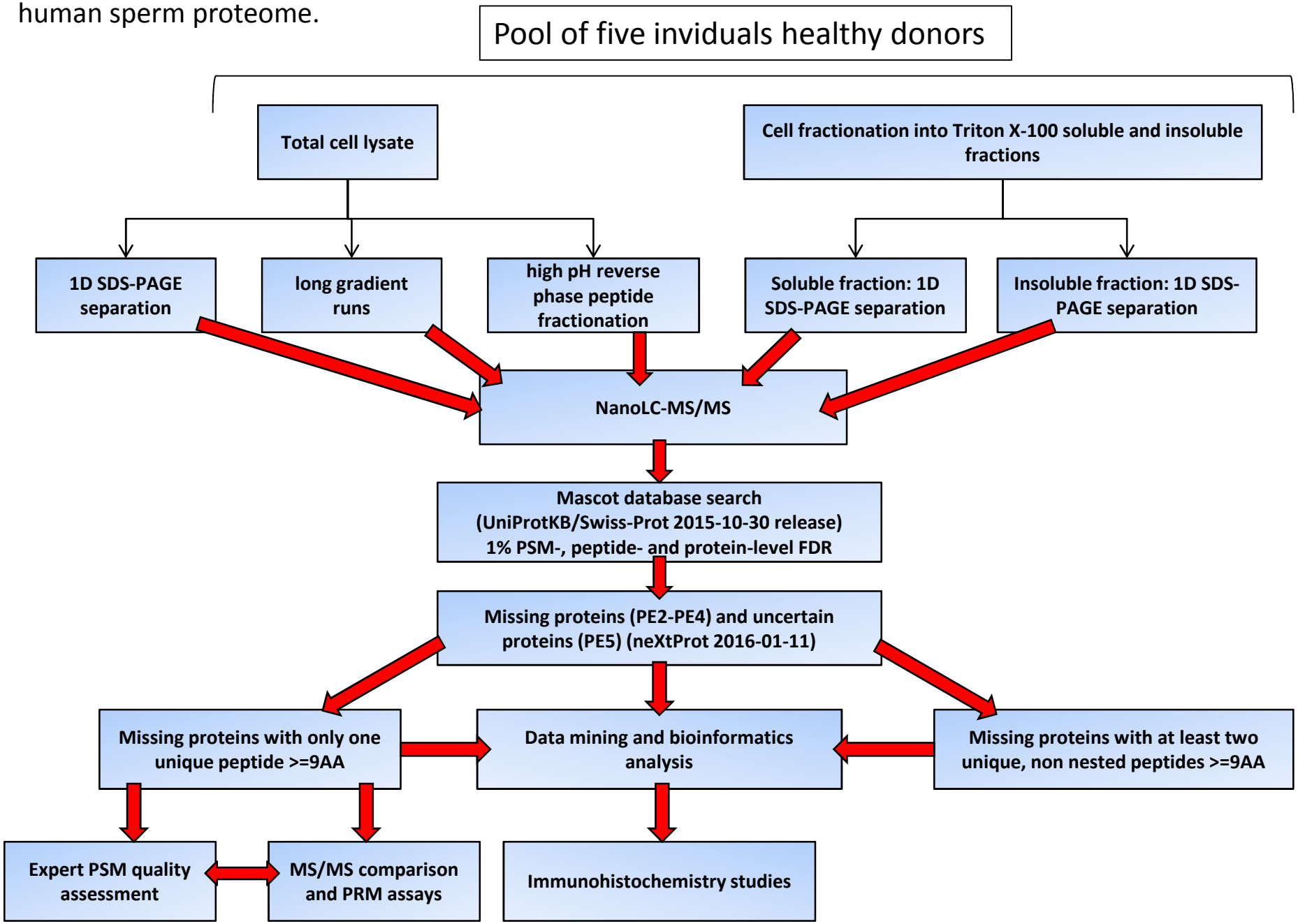


Figure 2. Contribution of the different fractionation protocols to identification of spermatozoa proteins.

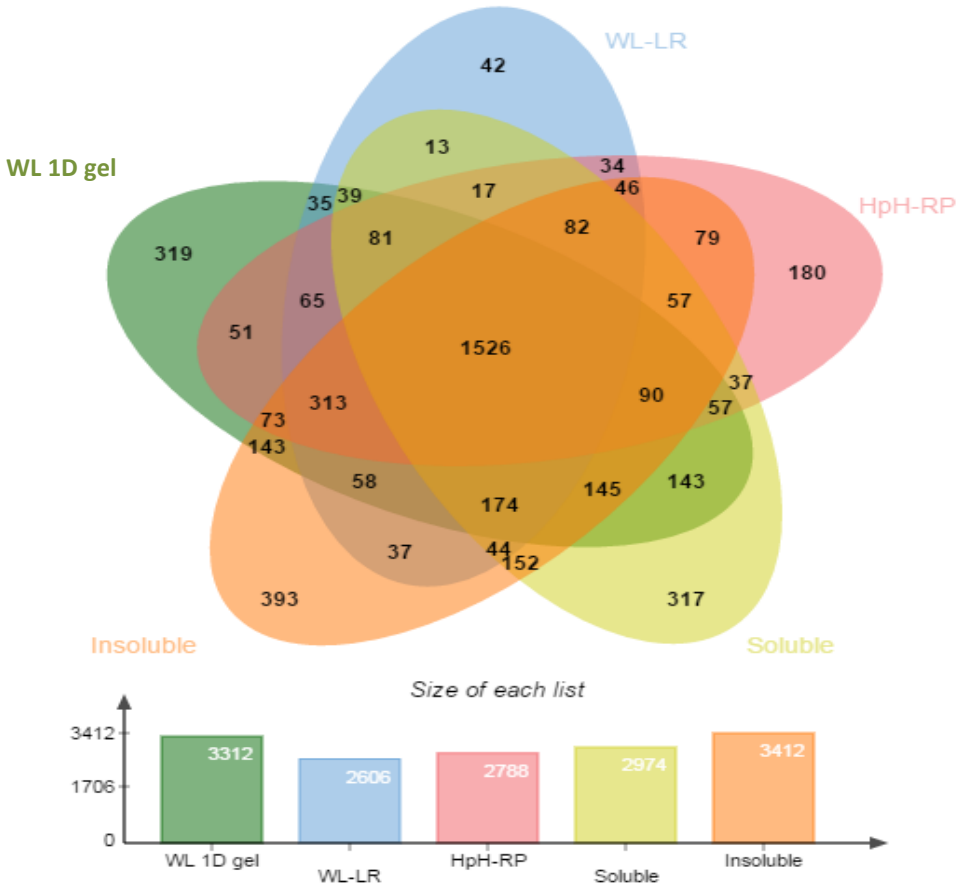


Figure 3. Missing proteins detected in the sperm proteome.

(A)

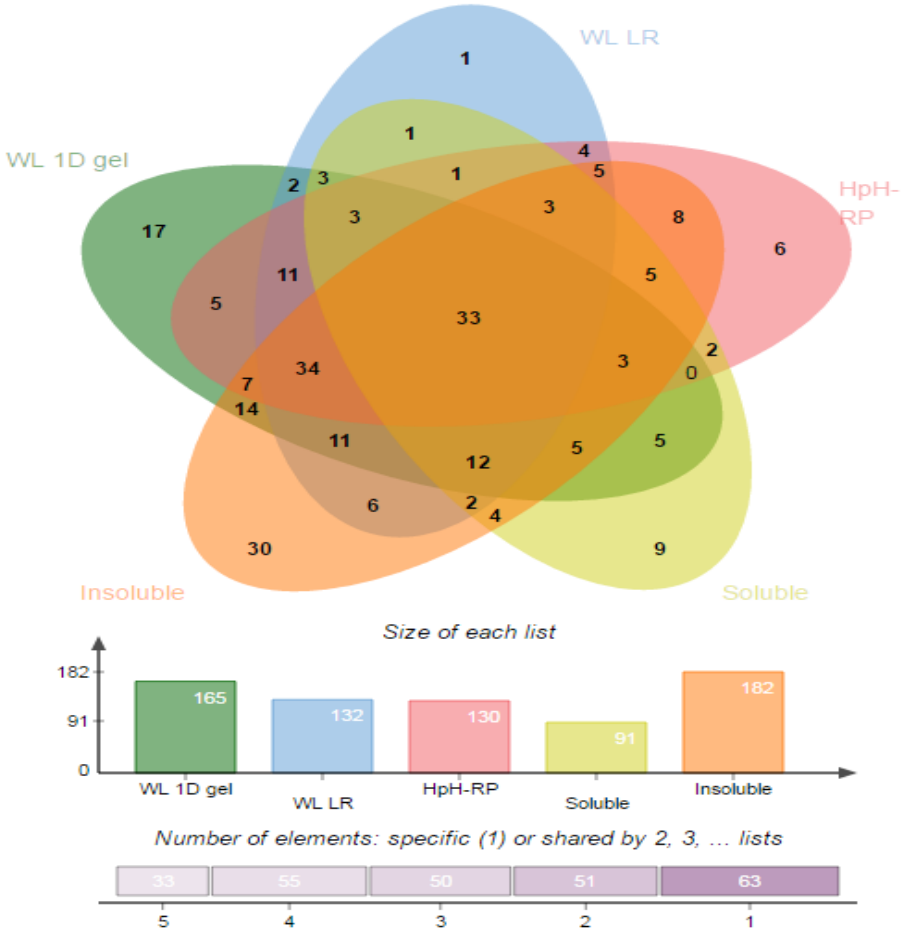
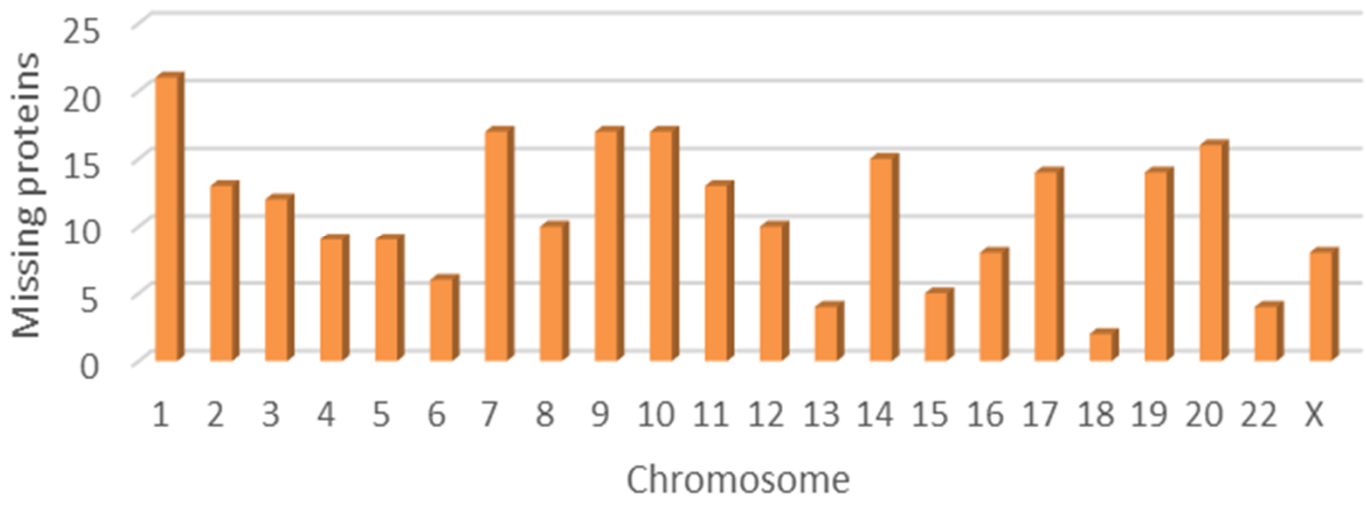


Figure 3. Missing proteins detected in the sperm proteome.

(B)



(C)

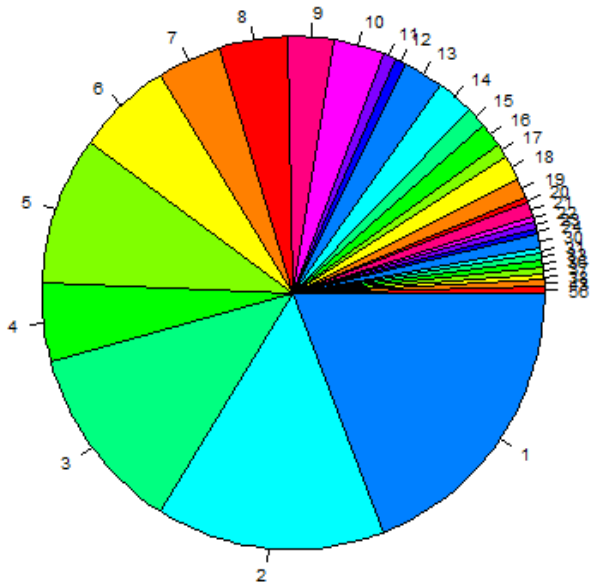


Figure 4. MS validation of one-hit wonder missing protein, example of protein Q9H693 (C16orf95)

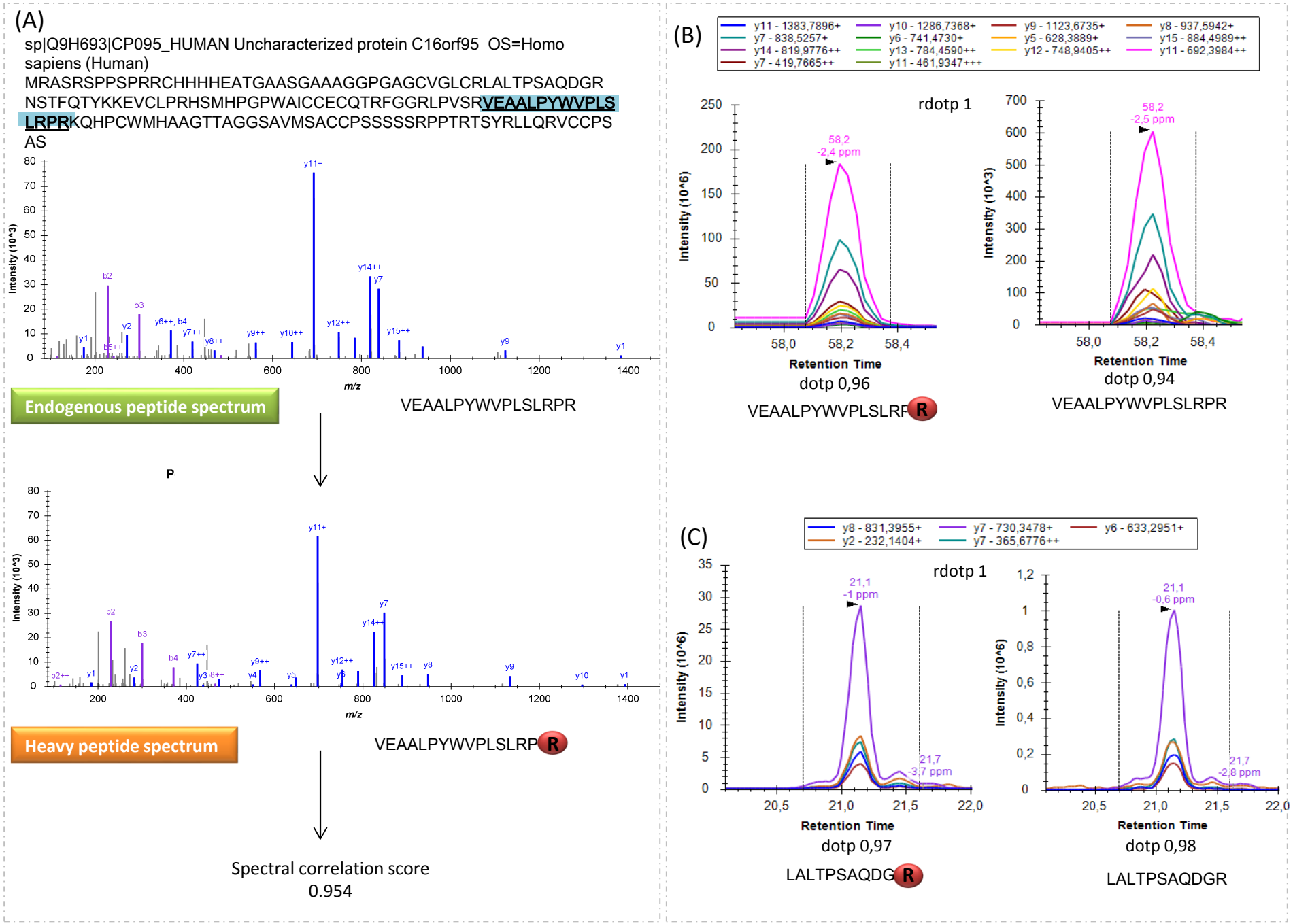


Figure 5A:

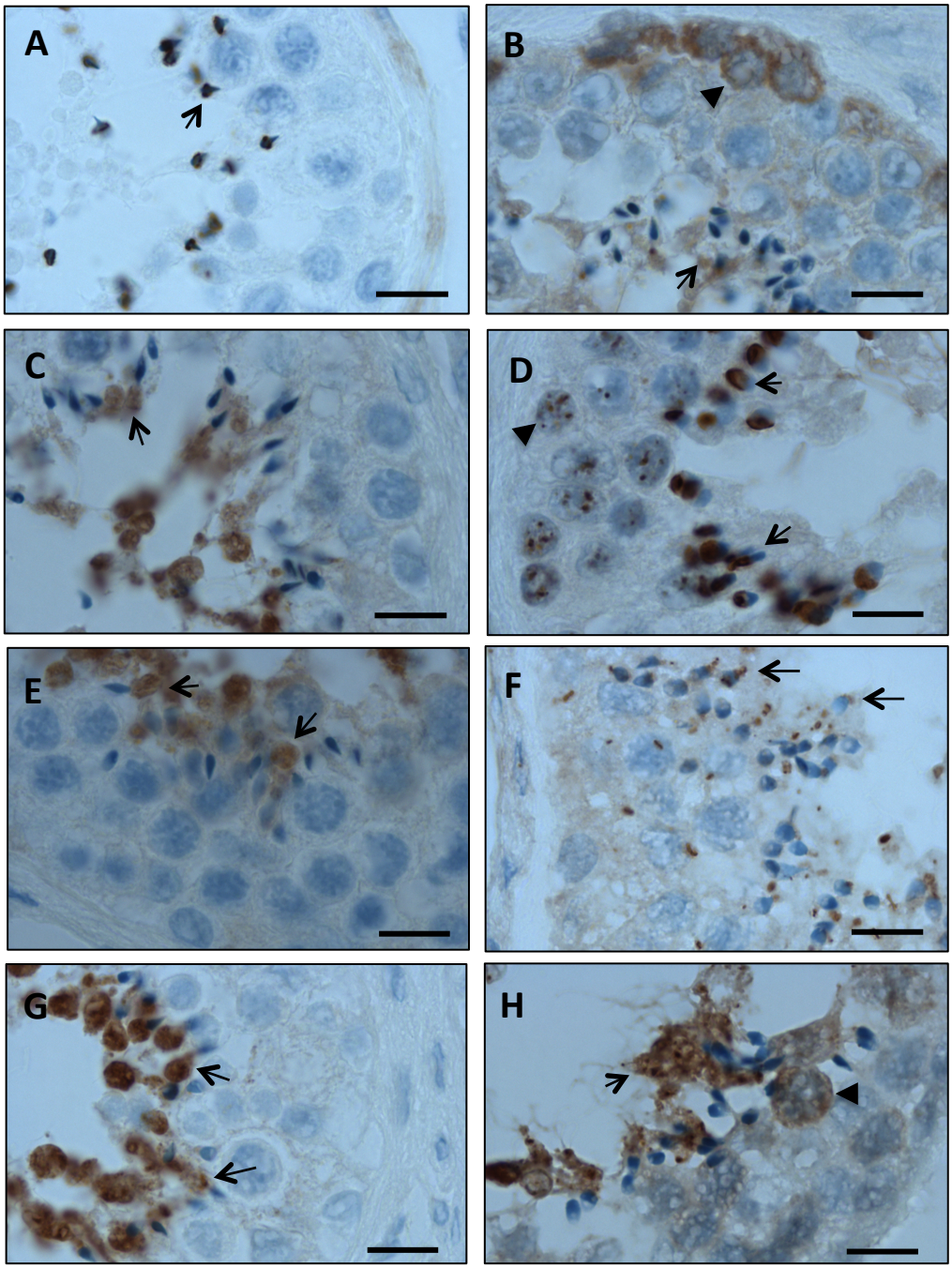
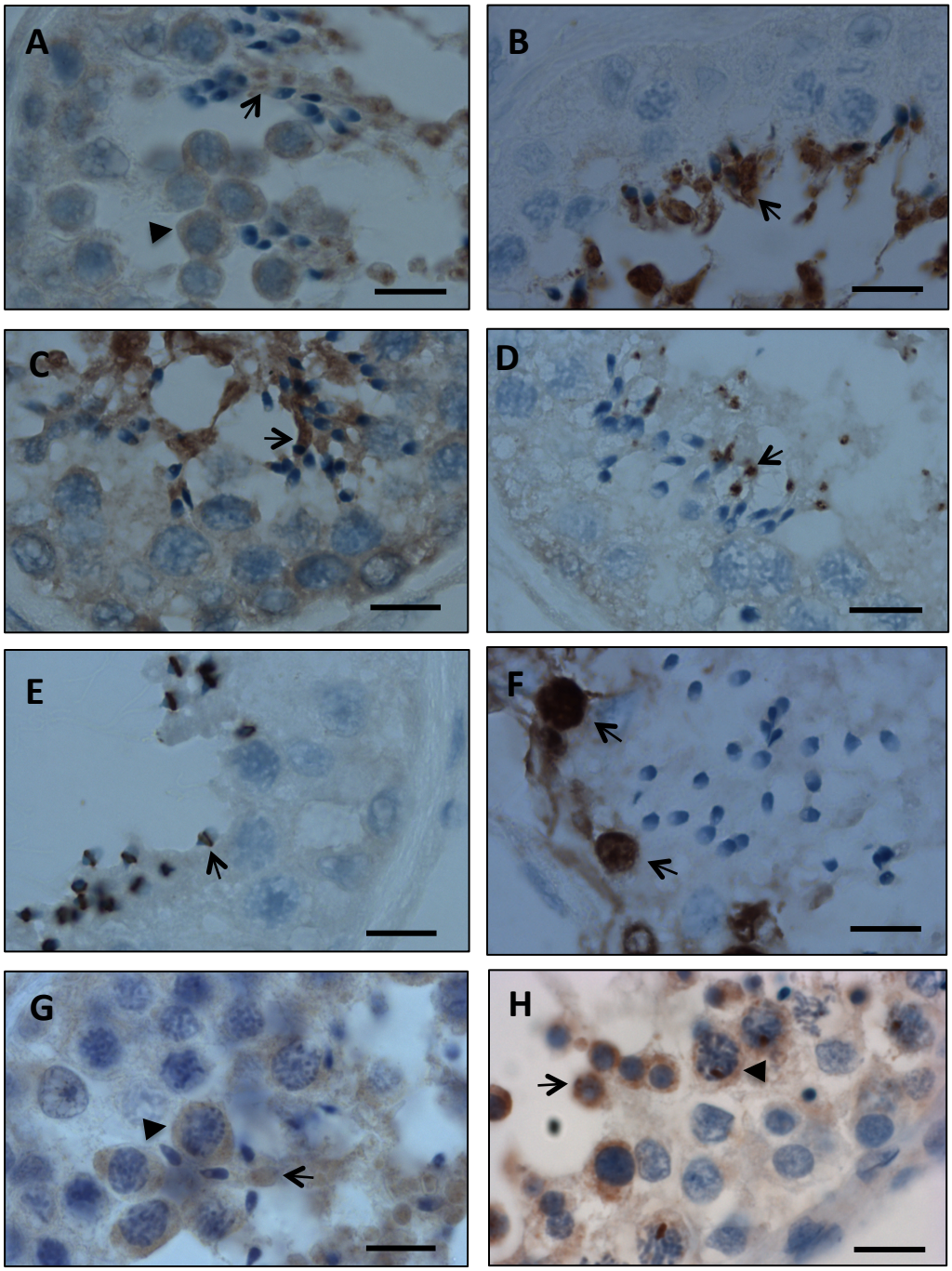
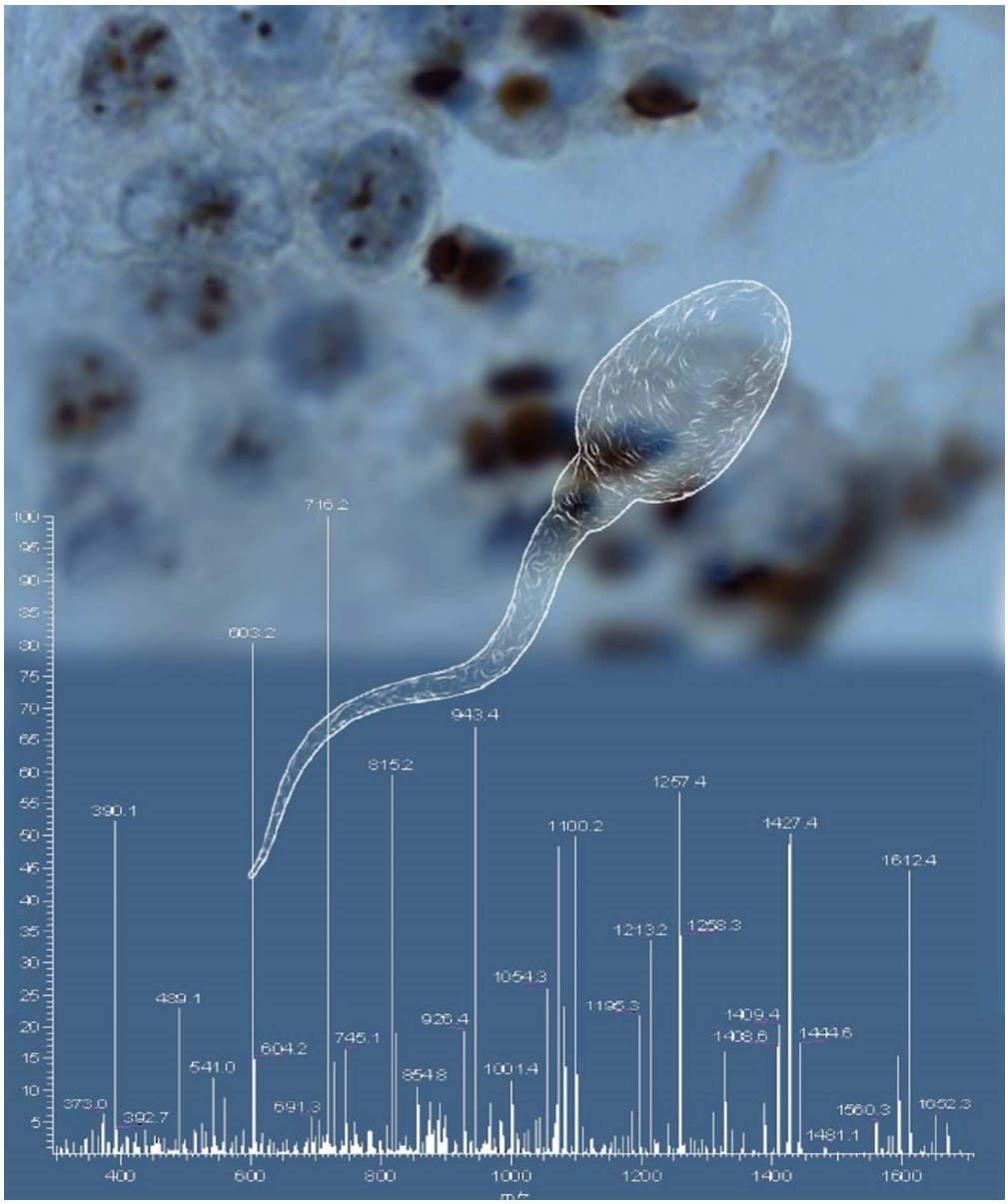


Figure 5B:





For TOC Only
For TOC Only
67x80mm (300 x 300 DPI)