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

Authors: Chocu, Sophie, Evrard, Bertrand, Lavigne, Régis, Rolland, Antoine D., Aubry, Florence, et al.

Source: Biology of Reproduction, 91(5)

Published By: Society for the Study of Reproduction

URL: https://doi.org/10.1095/biolreprod.114.122416

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

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

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

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

ABSTRACT

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

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

Received: 18 June 2014.
First decision: 8 July 2014.
Accepted: 11 August 2014.

eISSN: 1529-7268 http://www.biolreprod.org
ISSN: 0006-3363

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

INTRODUCTION

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

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

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

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

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

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

MATERIALS AND METHODS

Ethics Statement

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

Animals

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

Isolation of Testicular Cells

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

RNA-Sequencing Analysis

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

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

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

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

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

Creation of the Rat Nonredundant Reference Proteome Database

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

Mass Spectrometry Analysis

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

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

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

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

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

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

Refinement of Transfrag Selection

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

Statistical Filtration and Cluster Analysis

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

Multiple Alignments of Protein Sequences

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

Experimental Validation

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

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

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

In Situ Expression Analyses

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

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

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

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

Data Access

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

RESULTS

Experimental Design and PIT Workflow

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

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

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

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

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

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

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

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

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

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

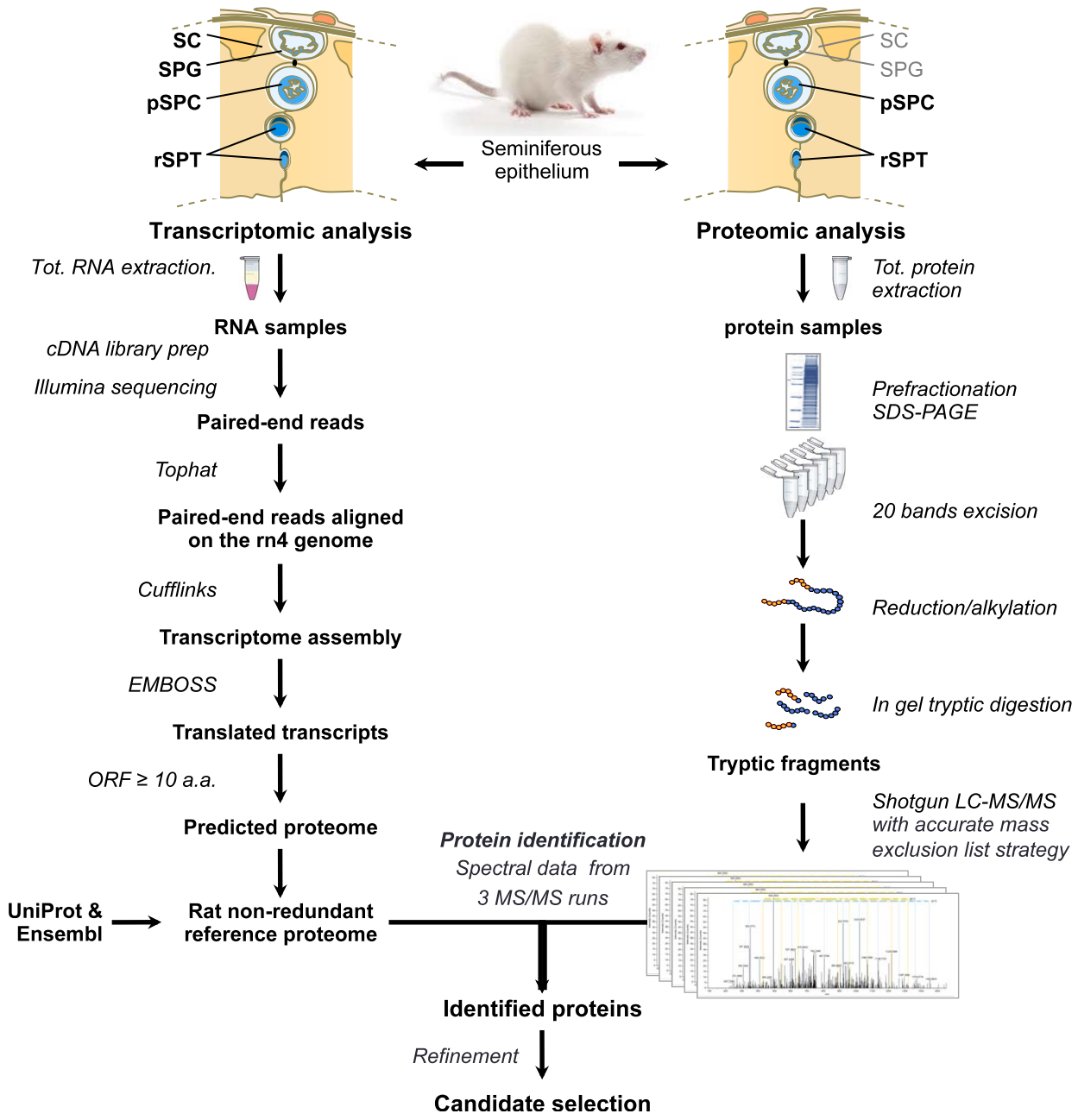

FIG. 1. Experimental design and PIT workflow. A schematic diagram of the strategy used to identify novel protein-coding loci by combining transcriptomic and proteomic data analysis.

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

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

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

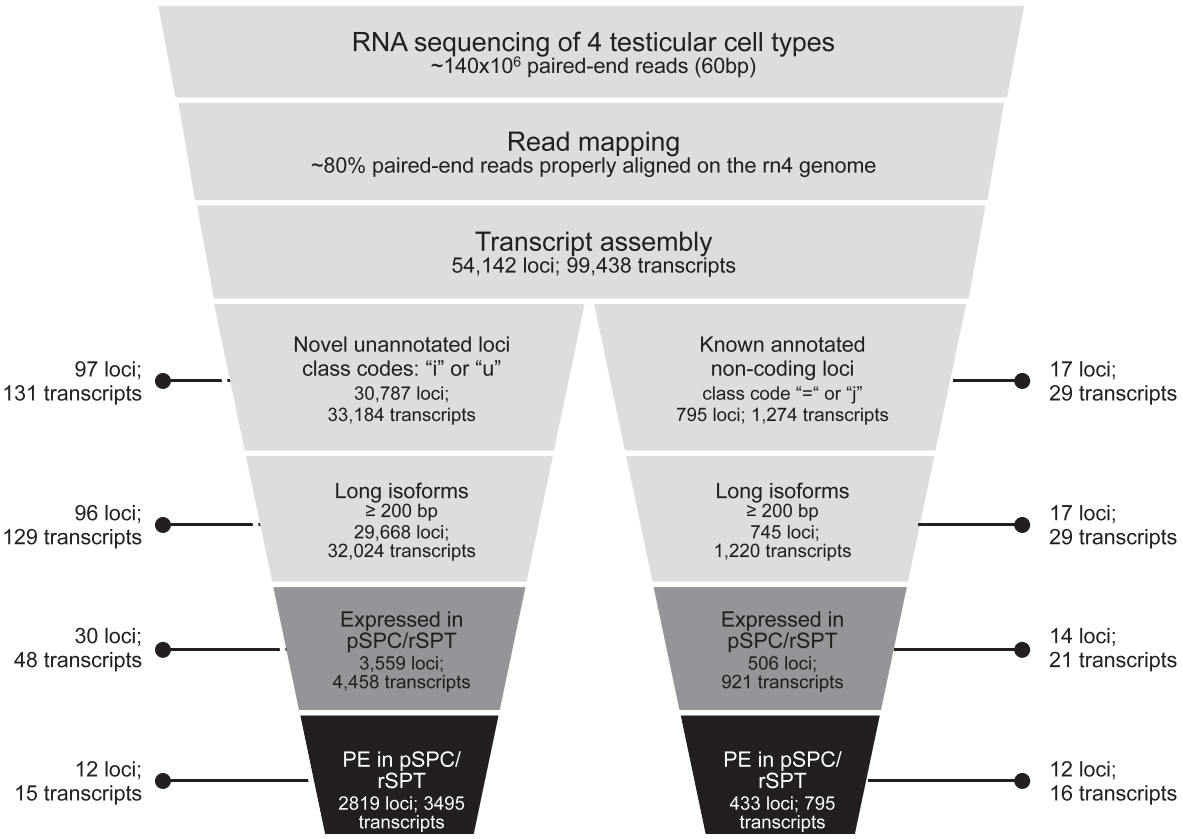

FIG. 2. PIT and refinement strategies used to select novel high-confidence protein-coding loci expressed in germ cells. From top to bottom, Illumina sequencing generated millions of 60-bp paired-end reads, which are aligned to the rat genome (release rn4) and subsequently assembled into cell-specific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

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

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

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

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

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

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

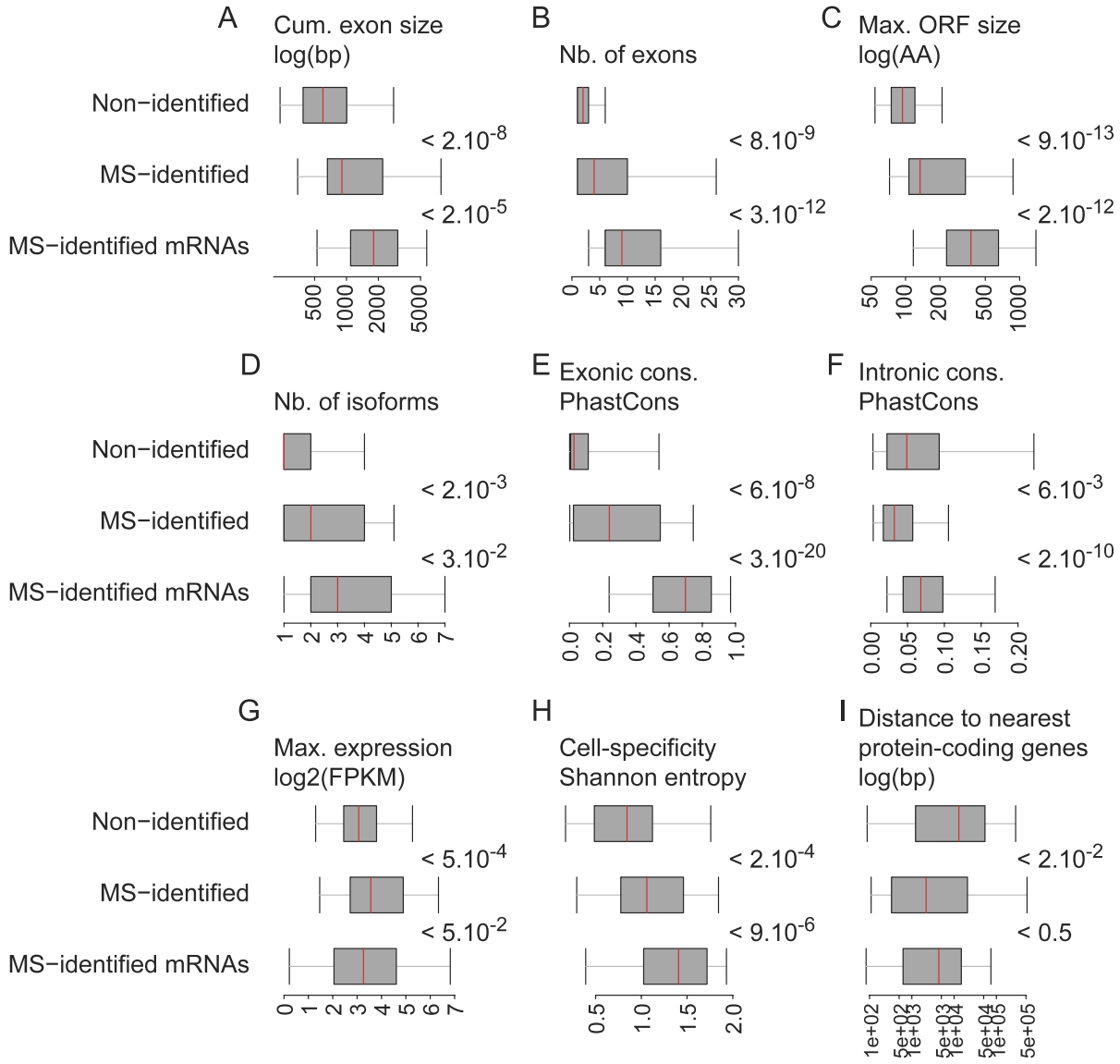

FIG. 3. Genomic and transcriptomic features of MS-identified transcripts. MS-identified TUTs, and known lncRNAs were compared to nonidentified TUTs and known lncRNAs and to MS-identified mRNAs. The box plots summarize the distributions of: cumulative exon length (**A**); number of exons (**B**); maximum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

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

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

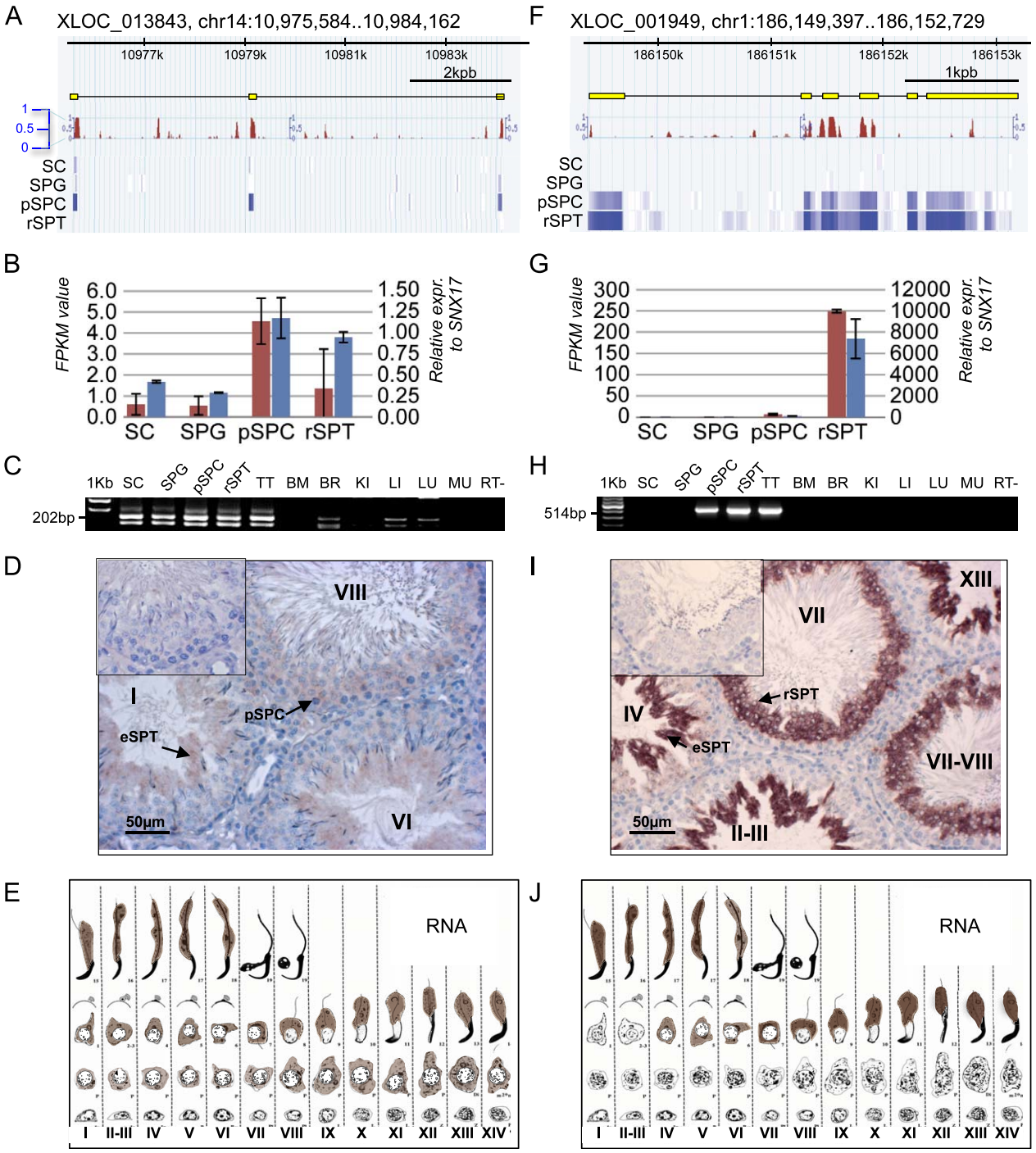

FIG. 4. Cell-specific expression patterns of XLOC_013843 (VAMP9) and XLOC_001949 (T-ENOL) transcripts. The expression patterns of XLOC_013843 (**A**–**E**) and XLOC_001949 (**F**–**J**) were investigated. **A** and **F**) The gene structure for both (yellow boxes correspond to exons), the sequence conservation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

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

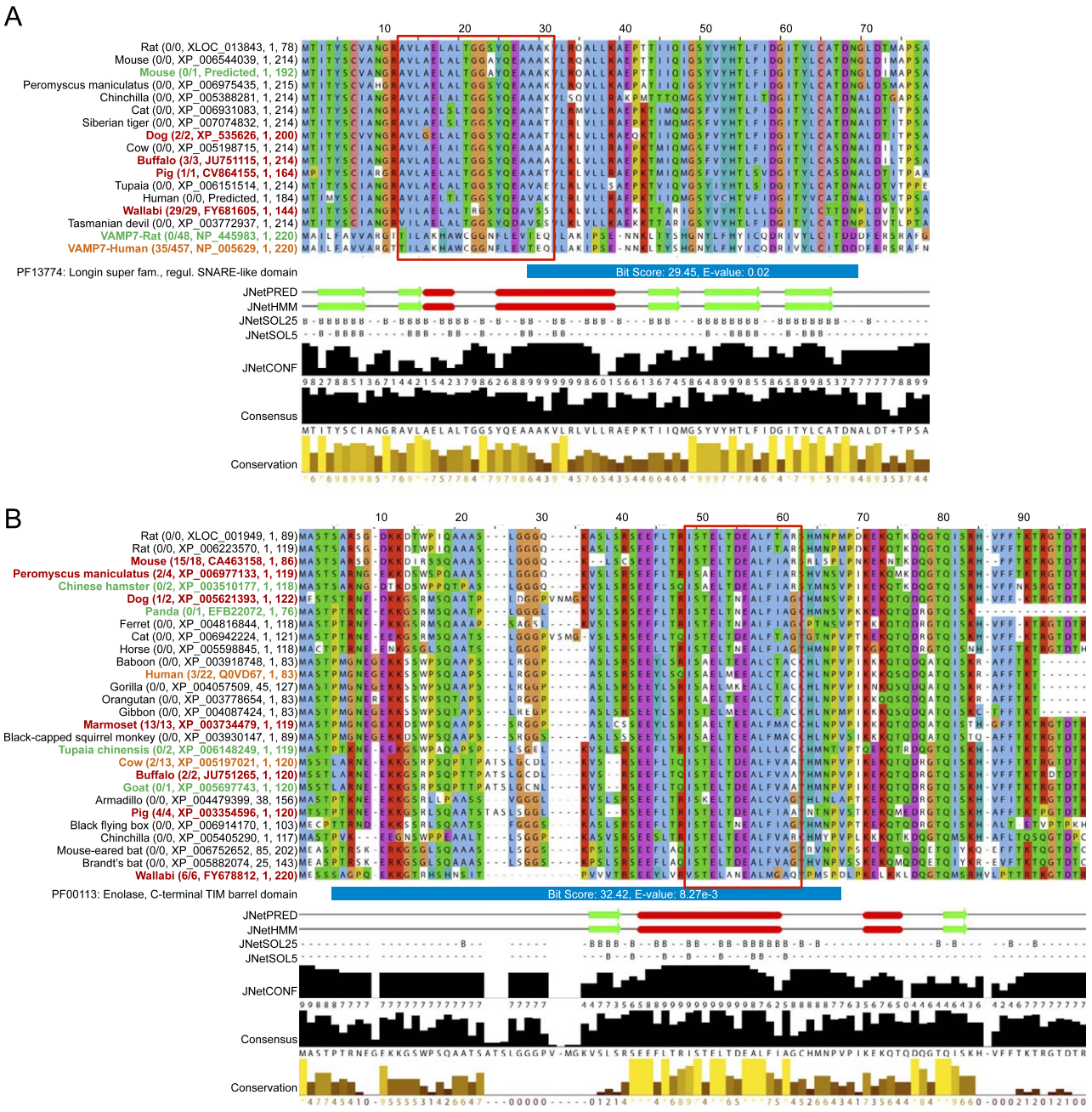

FIG. 5. Sequence conservation of the two proteins selected for detailed analysis, XLOC_013843 (VAMP9), and XLOC_001949 (T-ENOL). Predicted orthologous proteins of XLOC_013843 (**A**) and XLOC_001949 (**B**) in several mammalian species were retrieved using NCBI Blast programs with RefSeq, UniProt, and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

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

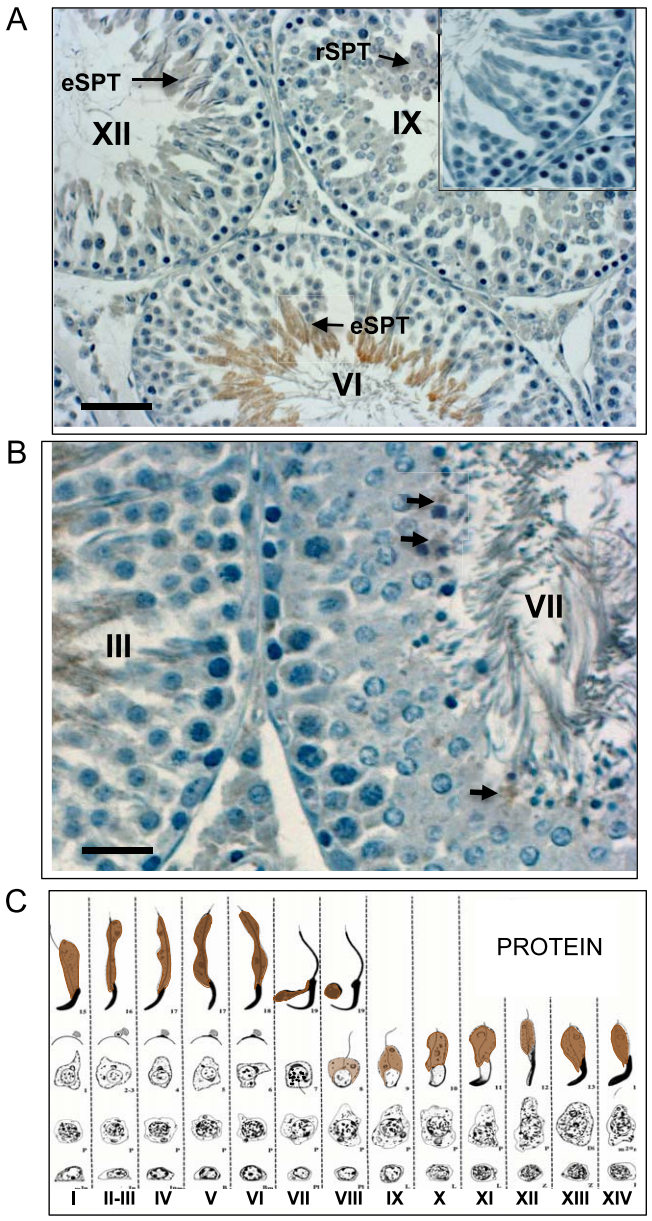

FIG. 6. Validation of the expression of XLOC_001949 (T-ENOL) by detection of a protein. **A** and **B**) Immunolocalization of the XLOC_001949 protein in adult rat testis sections as revealed with the anti $_{rec}$XLOC_001949 rabbit polyclonal antiserum. Serial rat testis sections were similarly probed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

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

DISCUSSION

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ACKNOWLEDGMENT

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

REFERENCES

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

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

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