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

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

Sophie Chocu, Bertrand Evrard, Régis Lavigne, Antoine D. Rolland, Florence Aubry, Bernard Jégou, Frédéric Chalmel, and Charles Pineau

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 some of these 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 validating testicular expression studies. 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. 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.

INTRODUCTION

Spermatogenesis is a specialized and dynamic process facilitating the transmission of genetic inheritance. 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. Consistent with the complexity of the process, the testis is one of the most complex organs in the body.

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. 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. 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. 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. 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).

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 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, the spectra 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 tests 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 testis and epididymis samples and for testicular cell isolation, in situ hybridization, and immunochemical experiments.” Animals were purchased from Elevage Janvier.

**Isolation of Testicular Cells**

Pachyteste spermatocytes (gSPC) 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 previously described methods [40, 41]. For RNA extraction, protein extraction, Western blot experiments, and isofracton 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] mm4 [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 (mm4, 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 (AJIS) for each TopHat run. The junction outputs from individual TopHat runs were pooled and added to the AJIS to allow TopHat to use junction information from all samples. TopHat was run again for each sample using the resulting AJIS. The output of this second run comprised the final alignment.

**Ab initio transcriptome assembly.** Individual sample alignments for each testis and 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.

**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 Transseq 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 MgCl₂, pH 7.4). A cocktail of protease inhibitors with 1 mM EDTA, 0.5 mM dithiothreitol (DTT), 1 mM 4-(2-aminoethyl) benzene sulfonyl fluoride hydrochloride, 10 mM trans-epoxysuccinyl-L-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 x g for 4°C for 10 min to remove cellular debris. The supernatants were then centrifuged at 105,000 x g at 4°C for 1 h. The supernatants containing the soluble proteins were kept on ice, and the pellets were recovered in 100 mM NaCl, 0.6 U/mL nuclease, and 2% (v/v) Nonidet P-40. These suspensions were again centrifuged at 105,000 x 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 prefraccionation and digestion.** Aliquots of 100 μg of total proteins from spermatocytes and spermatids were denatured at 70°C for 10 min in a LDES (NaPage 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 = 60000 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 data-dependent elution profile) and a charge state of 2 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 the 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 by ProteinProspector (University of California) and/or the Thermo RunView software, 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 HighFidelity DNA Polymerase (Qiagen) for cloning, in a Peltier thermocycler (Labgene). PCR products were then resolved on 1% 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.

Experimental Validation

RT-PCR and real-time quantitative RT-PCR. Complementary DNA was obtained from all 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 qPCR 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 HighFidelity DNA Polymerase (Qiagen) for cloning, in a Peltier thermocycler (Labgene). PCR products were then resolved on 1% 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 BamHI-HF in the appropriate CutSmart Buffer (New England Biolabs) to obtain the sequences of interest as BamHI 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-thiogalactopyranoside to induce recombinant protein production. The bacterial cells were harvested and lysed, and the soluble XLOC_001949 6 His-tagged recombinant protein (XLOC_001949) present in the supernatant was purified by affinity on Ni-NTA agarose slurry (Qiagen), using 250 mM imidazole elution buffer according to the manufacturer’s protocol. The purity of 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 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 solution.
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/Microtek) 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 Qiagucc Quick Gel Extraction Kit (Qiagen), inserted into the pCRII-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 BamHI linearized vector) and antisense (Sp6 RNA polymerase on XhoI 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 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 dataset identifier PXD000872. Sequences of the identified peptides were mapped on the rat genome using BLAT [62] and subsequently deposited in the Ensembl (32,971 known and 44,993 predicted protein 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 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 4,999 nonredundant proteins in pSPC (16,611 peptides; 4,056 proteins) and/or rSPT (11,755 peptides; 3,061 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 IncRNAs (506 loci) (Fig. 2, right). Mass spectrometry identified translation products for 69 of these transcripts (44 loci), including 48 TUTs (30 loci) and 21 IncRNAs (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 IncRNAs (12 loci) that were preferentially expressed in pSPC and/or rSPT (patterns P4–P6) (Fig. 2).

The Features of MS-Identified Transcripts Diverge Significantly from Those Typical of Known IncRNAs
TUTs share many genomic characteristics with known IncRNAs 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 IncRNAs 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...
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; \(P < 2.10^{−12}\)) (Fig. 3A). The number of exons for MS-identified (\(med = 4\) exons) was significantly greater than for non-MS-identified TUTs and lncRNAs (\(med = 2\)) 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 IncRNAs (\(q1 = 1.0\), \(med = 2.0\), \(q3 = 4.0\) ) 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].
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, \( P = 6.10^{-2} \) and 0.0 for intron, \( P = 6.10^{-2} \)) 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 3.3; \( P < 5.10^{-2} \)) 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 \( Q = 1.1 \)) significantly lower than that for non-MS-identified transcripts (0.8; \( P < 2.10^{-2} \)) 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 = 12946, q3 = 53684) were about six times farther from any protein-coding genes than were MS-identified TUTs and lncRNAs (q1 = 332, med = 2178, q3 = 20701; \( P < 2.10^{-2} \)) (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 \( \geq 22.5\% \); E value \( < 10^{-4} \)). It maps on chromosome 14 (positions 10975584–10984162) and is composed of three exons with a cumulative exon size of 256
nt and a maximum ORF size of 78 aa (Figs. 4A and 5A). This TUT was among the most conserved among vertebrates (phastCons score \( \approx 0.735 \); 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 \( \approx 0.02 \)). 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. 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.
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
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. 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: JNetPRE, 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.
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 were merged with canonical databases. Searching for novel protein-coding genes identified between MS-identified and non-MS-identified transcripts showed meiotic or postmeiotic expression patterns. 

Among 5379 TUTs and IncRNAs (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. MS-identified transcripts showed genomics features differing from those typically shared by lncRNAs: they 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 evidence of protein corresponding to novel protein-coding loci. 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 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 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 protein in adult rat testis sections as revealed with the anti-recombinant 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 spermatids 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.).

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-recombinant 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 spermatids 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 spermatids (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.
mental 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 recXLOC_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 VAMP6 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 cilio genesis [95]; note that 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 applicability 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 (Bioinformatics Platform, IRISA) and Laetitia Cloarec (Inserm U1085, Genouest) for their help. We thank Dominique Mahe Poiron, Bioinformatics Platform, IRISA) and Laetitia Cloarec (Inserm U1085, Genouest) for their help. We thank all members of the SEQuanssors 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

31. Brewis IA, Brennan P. Proteomics technologies for the global identifica-
PIT REVEALS NOVEL GENES EXPRESSED IN GERM CELLS


