High-Resolution Profiling of Novel Transcribed Regions During Rat Spermatogenesis 1

Authors: Chalmel, Frédéric, Lardenois, Aurélie, Evrard, Bertrand, Rolland, Antoine D., Sallou, Olivier, et al.

Source: Biology of Reproduction, 91(1)

Published By: Society for the Study of Reproduction

URL: https://doi.org/10.1095/biolreprod.114.118166

BIOLOGY OF REPRODUCTION (2014) **91**(1):5, 1–13
Published online before print 16 April 2014.
DOI 10.1095/biolreprod.114.118166

High-Resolution Profiling of Novel Transcribed Regions During Rat Spermatogenesis[1]

Frédéric Chalmel,[2,6] Aurélie Lardenois,[3,6] Bertrand Evrard,[6] Antoine D. Rolland,[6] Olivier Sallou,[7] Marie-Charlotte Dumargne,[4,6] Isabelle Coiffec,[6] Olivier Collin,[7] Michael Primig,[5,6] and Bernard Jégou[5,6,8]

[6]*Inserm U1085-IRSET, Université de Rennes 1, Rennes, France*
[7]*Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA/INRIA)–GenOuest platform, Rennes, France*
[8]*Ecole des Hautes Études en Santé Publique, Rennes, France*

ABSTRACT

Mammalian spermatogenesis is a complex and highly orchestrated combination of processes in which male germline proliferation and differentiation result in the production of mature spermatozoa. If recent genome-wide studies have contributed to the in-depth analysis of the male germline protein-encoding transcriptome, little effort has yet been devoted to the systematic identification of novel unannotated transcribed regions expressed during mammalian spermatogenesis. We report high-resolution expression profiling of male germ cells in rat, using next-generation sequencing technology and highly enriched testicular cell populations. Among 20 424 high-confidence transcripts reconstructed, we defined a stringent set of 1419 long multi-exonic unannotated transcripts expressed in the testis (testis-expressed unannotated transcripts [TUTs]). TUTs were divided into 7 groups with different expression patterns. Most TUTs share many of the characteristics of vertebrate long noncoding RNAs (lncRNAs). We also markedly reinforced the finding that TUTs and known lncRNAs accumulate during the meiotic and postmeiotic stages of spermatogenesis in mammals and that X-linked meiotic TUTs do not escape the silencing effects of meiotic sex chromosome inactivation. Importantly, we discovered that TUTs and known lncRNAs with a peak expression during meiosis define a distinct class of noncoding transcripts that exhibit exons twice as long as those of other transcripts. Our study provides new insights in transcriptional profiling of the male germline and represents a high-quality resource for novel loci expressed during spermatogenesis that significantly contributes to rat genome annotation.

intergenic transcripts, intronic transcripts, lncRNAs, mammalian spermatogenesis, novel transcribed regions, RNA profiling, Sertoli cells

INTRODUCTION

A large number of genes are temporally regulated during spermatogenesis. This process consists of three main steps: male germ cell mitoses, meiotic divisions (meiosis), and spermiogenesis (postmeiosis); the last step leads to formation of spermatozoa. Before the advent of next-generation sequencing, a number of groups, including ours, used various transcriptome technologies (e.g., expressed sequence tag libraries, serial analysis of gene expression, and microarray analyses) to study gene expression during spermatogenesis [1–8]. It was clearly demonstrated that testis is among the organs that expresses the largest number of genes in a tissue-specific manner and that these testis-specific genes are expressed mostly in the germline. Recently, several studies conducted RNA sequencing (RNA-seq) for expression quantification analyses during spermatogenesis by using either whole testes or enriched populations of germ cells in the mouse [9–12]. Those studies provide a global overview of the testicular protein-encoding gene expression program; however, they were somewhat limited in terms of deciphering its noncoding counterpart and genomic characterization of novel unannotated transcribed regions.

Advances in sequencing technologies are making it possible to explore transcriptomes in unprecedented detail and to identify numerous novel transcriptionally active unannotated genomic loci that are likely not translated into proteins [13–16]. The resulting transcriptional products, present in all eukaryotic species, were grouped into a heterogeneous class of uncharacterized transcripts termed noncoding RNAs (ncRNAs) [14, 17, 18]. Long noncoding RNA (lncRNAs) molecules are a recently discovered subclass of ncRNAs [19, 20]; they are by definition longer (mature transcripts $\geq$200 nucleotides in length) than another subclass of ncRNAs called small ncRNAs (sncRNAs; $<$200 nucleotides), which includes micro-RNAs [21, 22]. Many genomic characteristics are commonly shared by lncRNAs in vertebrates, 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 [23–30]. It has been suggested that the lower GC content may partly explain the lower expression level of lncRNAs than that of mRNAs [20, 31–34]. Like mRNAs, lncRNAs commonly consist of several exons that are combined after splicing of introns into mature transcripts [35]. However, Cabili et al. [23]

[1]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); INERIS-STORM grant N 10028NN to B.J.; an INSERM Young Investigator postdoctoral fellowship grant to A.L.; a Rennes Métropole Défis scientifiques émergents-2011 grant to F.C.; and INSERM Avenir and Région Bretagne CREATE grants R07216NS and R11016NN to M.P. The RNA sequence data files were submitted to National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and to NCBI Gene Expression Omnibus (GEO) under accession numbers SRP026340 and GSE48321, respectively. Selected testicular unannotated transcript fragments were deposited in GenBank Transcriptome Shotgun Assembly sequence database as BioProject no. PRJNA209702.
[2]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
[3]Current address: INRA, UMR703 PAnTher, F-44307 Nantes, France; LUNAM Université, Oniris, École nationale vétérinaire, agro-alimentaire et de l'alimentation Nantes-Atlantique, F-44307 Nantes, France.
[4]Current address: Human Developmental Genetics Unit, Institut Pasteur, Paris, France.
[5]These authors contributed equally to this work.

Received: 27 January 2014.
First decision: 3 March 2014.
Accepted: 26 March 2014.

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

suggested that the transcript length and exon number of lncRNAs may be underestimated because of partial transcript reconstruction due to their low abundance. It has also been reported that lncRNAs are preferentially located next to genes associated with specific biological processes [23, 25, 29, 30]. These observations contributed to the hypothesis that lncRNAs might be involved in mechanisms of tissue-specific/cell-specific regulatory controls via epigenetic modifications over neighboring protein-encoding genes [30, 36, 37]. Despite valuable bioinformatics efforts to predict their biological roles [38], they remain mostly uncharacterized from a functional point of view. However, there have been numerous studies of individual lncRNAs such as HOTAIR [39, 40], Xist [41], MALAT-1 [42], PCAT-1 [43], lincRNA-p21 [44], PANDA [45], and Jpx [46] showing they are involved in diverse cellular and biological processes such as chromatin remodeling, gene expression, post-transcriptional processing, intracellular trafficking, neurogenesis, and embryogenesis [19, 20, 33, 34, 47–52]. Recent reports suggest potential associations between lncRNAs and a number of human disorders [53] including cancers [35, 39, 42, 43]. It is, therefore, becoming clear that lncRNAs can act through a large diversity of mechanisms to regulate many biological processes in eukaryotes.

Apart from the accumulation of known lncRNAs observed in the whole testis during the first wave of spermatogenesis [9] and the most recent study by Soumillon et al. [10] in mouse, no comprehensive survey and characterization of novel unannotated loci expressed during spermatogenesis in mammals has been undertaken. Here, we report the transcriptional profiling and characterization by RNA-seq of novel testis-expressed unannotated transcripts (TUTs) present during spermatogenesis in the rat. We performed paired-end high-throughput sequencing with RNAs from highly enriched preparations of somatic Sertoli cells (SE), spermatogonia (SG; mitosis), spermatocytes (SC; meiosis) and round spermatids (ST; postmeiotic germ cells). After mapping reads, we were able to assemble a large fraction of the annotated transcripts and also novel isoforms for known protein-encoding and noncoding loci. This dataset was compared to All-Exon GeneChip (Affymetrix) data, which were used as an internal control of data quality, which widely confirmed our RNA-seq data. We focused our analysis on systematic identification of long, multi-exonic TUTs that were highly detected during spermatogenesis. A high-confidence set of 1419 TUTs including 435 potential lncRNAs was defined, and subsequent characterizations identified several properties. These unannotated transcripts showed most of the genomic features typically associated with vertebrate lncRNAs (e.g., short length, low exon number, low abundance, low GC content, low sequence conservation, and restricted expression patterns). Importantly, classification of TUTs and known lncRNAs according to their expression pattern during spermatogenesis revealed several specific characteristics, including an exon length of meiosis-induced TUTs and known lncRNAs that was unexpectedly twice as long as that of known lncRNAs with other expression patterns. Transcriptional profiling and subsequent characterization of TUTs dynamically expressed during rat spermatogenesis may lead to identification of novel candidate loci for the regulation of gene expression in either *cis* or *trans* in mammalian testis. Our study provides insight into the lncRNA expression program in mammalian testis and significantly improves annotation of the rat genome. It may also ultimately help elucidate molecular events leading to reproductive disorders. A graphic display of this high-quality dataset is conveniently available to the scientific and medical communities through the ReproGenomics Viewer (RGV) [54].

MATERIALS AND METHODS

Ethics Statement

Experimental research using animals reported here conformed to the principles for the use and care of laboratory animals in compliance with French and European regulations of animal welfare. Furthermore, experimenters were granted authorization from the French Direction des Services Vétérinaires to conduct or supervise experiments with live animals.

Sample Isolation

Male Sprague-Dawley rats were purchased from Elevage Janvier. Sertoli cells, spermatogonia, pachytene spermatocytes, and round spermatids were highly enriched as previously described [2, 6, 55]. Briefly, pachytene spermatocytes and round spermatids were prepared from 90-day-old rats: testes were trypsinized and the resulting cell suspension was fractionated by centrifugal elution. Spermatogonia were purified from 9-day-old rat testes that were sequentially dissociated with various enzymes and then sedimented at 1 *g* in a 2%–4% bovine serum albumin gradient. Sertoli cells were isolated from 20-day-old rats: testes were sequentially dissociated with various enzymes, and the resulting cell suspension was differentially sedimented; Sertoli cells were then plated and cultured for 4–7 days. Testicular cells were enriched from 2 pools (duplicate samples for the RNA-seq experiment) and 3 pools (triplicate samples for the All-Exon array [Affymetrix] experiment of 8 Sertoli cells, 20 spermatogonia, and 4 rats [spermatocytes and round spermatids]).

For in situ expression analysis, 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 (μM) for 20 min. Testes were isolated and immersed in the same fixative for 6 h. Specimens were dehydrated in a graded series of ethanol concentrations of butanol and then embedded in paraffin. Five-micrometer-thick sections were cut and mounted onto poly-L-lysine-coated slides.

RNA Isolation and RNA-Seq Library Preparation

RNA isolation. Total RNA was isolated using TRIzol reagent (Invitrogen) according to the manufacturer's instructions. RNA integrity was measured using a model 2100 Bioanalyzer (Agilent), and only samples with an RNA integrity number (RIN) score >7.0 were further processed.

RNA-seq library preparation. RNA-seq libraries were prepared using an mRNA-seq sample prep kit (product no. RS-100-0801; Illumina Inc.) according to the manufacturer's protocol, with some modifications. Aliquots of 10 μg of total RNA were hybridized with eukaryote rRNA sequence-specific 5′ biotin-labeled oligonucleotide probes to deplete selectively abundant ribosomal RNA. The rRNA/5′ biotin-labeled probe hybrid was removed from the sample with streptavidin-coated magnetic beads (Ribominus eukaryote kit for RNA-seq; product no.A10837-08; Invitrogen). Then, 250 ng of rRNA-depleted RNA was fragmented with divalent cations at 95°C for 5 min. The cleaved RNA fragments were reverse transcribed to cDNA, using random primers and SuperScript II reverse transcriptase (product no. 18064-014; Invitrogen). Second-strand cDNA was then synthesized using polymerase I and RNase H. Double-stranded cDNA fragments were end-repaired using T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase (PNK). Klenow fragment (3′–5′ exonuclease minus) was used to add a single adenosine to the 3′ ends of the blunt DNA fragments. The ends of the DNA fragments were ligated to double-stranded adapters by using T4 DNA ligase. Ligated products were separated by 2% agarose gel electrophoresis, and ~200- to 220-base pair (bp) fragments were excised, purified using QIAquick gel extraction kit (Qiagen), and amplified by PCR (30 sec at 98°C; then 10 sec at 98°C, 30 sec at 65°C, 30 sec at 72°C for 13 cycles; and a final step for 5 min at 72°C). Surplus PCR primers were then eliminated by purification (Agencourt AMPure XP beads; product no.A63881; Beckman). The resulting DNA libraries were checked for quality and quantified (2100 BioAnalyzer; Agilent). Each library was loaded into two lanes of the Illumina flow cell at a concentration of 5 pM, and clusters were generated using the cluster station and sequenced on the genome Analyzer II (Illumina) as unstranded, paired-end 2 × 60 base reads at depths of ~16.2–18.0 million paired-end reads per library (for statistics of read counts see Supplemental Table S1; all supplemental data are available online at www.biolreprod.org). Pipeline version 1.6 software (Illumina) was used for image analysis and base calling.

Mapping Reads, Transcriptome Assembly, and Quantification with the Tuxedo Suite

Comprehensive database of known transcripts. Transcript annotations from public databases (Ensembl [56], National Center for Biotechnology Information [NCBI; release RGSC3.4] [57], and AceView [58] and mRNAs from University of California Santa Cruz [UCSC] rn4 [59]) were merged into a combined set of nonredundant, known transcript annotations using Cuffcompare software [60, 61].

Mapping reads. RNA-seq-derived reads from each sample replicate were aligned independently with the *Rattus norvegicus* genome (rn4, downloaded from the UCSC genome browser website [59]) with TopHat (version 1.4.1) [62] using published approaches [29, 61]. The database of known transcripts and expressed sequence tag 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 software was run again for each sample, using the resulting AJS. The output of this second run comprised the final alignment. Finally, individual sample alignments for each testicular cell type were pooled.

Ab initio transcriptome assembly. The transcriptome of each individual cell type was assembled with Cufflinks (version 1.2.0) by finding a parsimonious allocation of reads to the transcripts within a locus, using default settings [60, 61]. The Cufflinks assembly step yielded a set of ~23 000–47 000 transcript fragments (transfrags) for each testicular cell type.

Merging and classification of transcript fragments. Cuffcompare software [60, 61] was used to merge the individual transfrags into a combined set (nonredundant union of all transcript fragments that share all introns and exons) and to classify the 122 262 resulting transcripts according to the known transcript annotation database. All transcripts that were not automatically annotated as complete match (Cuffcompare class ''='') , potentially novel isoform (class j), unknown intronic (class i; i.e., loci falling entirely within a reference intron and without exon–exon overlap with another known locus), and intergenic (class u) isoforms were discarded, yielding 77 490 transfrags that were retained in the analysis.

Transcriptome quantification and preprocessing. The isoform abundance (expression) levels were assessed using Cuffdiff [60, 61] for each sample with upper quantile normalization. Abundance was measured in fragments per kilobase of exons model per million reads mapped (FPKM). A matrix of FPKM values was then prepared from the results of transcriptome quantification. These expression data were subsequently log2-transformed after adding 0.05 to all FPKM values. Data were quantile-normalized to reduce systematic effects and to allow direct comparisons among individual samples.

GeneChip Hybridization and Preprocessing

A parallel expression profiling of the same testicular cells was performed in triplicate (rat exon 1.0 ST GeneChip; Affymetrix). Total RNA was purified using an RNA cleanup kit (Zymo Research). One microgram of each RNA sample was processed as prescribed by the GeneChip whole-transcript sense target labeling assay (Affymetrix). Briefly, GeneChip wild-type cDNA synthesis kit, cDNA amplification kit, and terminal labeling kit (Affymetrix) were used for target preparation. Fragmented second cycle cDNA were verified with RNA Nano 6000 chips run with the BioAnalyzer (Agilent), and end-labeled cocktail hybridization was applied to GeneChip rat exon 1.0 ST arrays. Arrays were hybridized for 16 h. A wash-and-stain script (precommercial FS450_0001 script) was applied (Station 450; Fluidics). Raw data files (in .DAT and .CEL data formats) were produced using the GCS 3000 TG system and ExpressionConsole 1.0 (Affymetrix) with the appropriate library file. GeneChip data were normalized using the robust multiarray average method (RMA) [63].

Refinement of Transcript Fragment Selection

As observed in the study by Prensner et al. [43], manual inspection of the resulting 77 490 transfrags revealed that almost all predicted loci probably corresponded to stochastic transcriptional noise, genomic DNA present in the sample, or artifacts due to errors in read mapping and transcript assembly. To eliminate poor-quality quantifications and identify the most robust transcripts from background signal, we applied three additional filtering steps. First, we defined a background expression cutoff value of 3.72 FPKM, calculated as the overall median of unlogged intensities for the assembled transcripts that completely matched (using Cuffcompare class ''='') NCBI RefSeq-curated mRNAs (Natural mRNA category, ''NM'') [57]; this allowed selection of 28 992 detectable or expressed transfrags (37.4%), defined as those for which expression levels (FPKM) were above the background expression cutoff value in both replicates of at least one cell type. Second, we selected 69 725 transfrags with a total length $\geq$200 nucleotides (nt; 90.0%). Finally we selected 39 885 transfrags (51.5%) that harbored at least two exons (multi-exonic). A set of 20 424 transfrags (26.4%) fulfilled all three conditions (intersection of the three additional filtering steps) and were thus identified as a minimal set of long multi-exonic RNA molecules expressed in rat Sertoli and/or germ cells (Fig. 1A).

Statistical Filtration and Cluster Analysis

The transfrags differentially expressed in four testicular cell types were statistically filtered using the annotation, mapping, expression, and network (AMEN) suite of tools [64] . We first isolated 19 116 transfrags that exhibited a $\geq$3-fold difference in expression between averaged cellular conditions (pairwise comparisons). A LIMMA (linear models for microarray data) statistical test [65] was then used to identify 14 856 significantly differentially expressed transfrags (F value was adjusted using the false discovery rate method: $P \leq 0.01$). The resulting transfrags were then grouped into six expression patterns (P1–P6) using the partitioning around medoids (PAM) algorithm. The ability of the patterns to discriminate between transcripts was verified using a silhouette plot. The six resulting patterns were ordered according to peak expression levels in the different cell types. The 5568 remaining transfrags for which no significant differential expression was observed (<3-fold change or $P > 0.01$) were placed in a 7th group named P0.

Coding Potential Analysis of Novel Transcribed Regions

Before analyzing the protein-encoding potential of transfrags corresponding to intronic or intergenic (Cuffcompare classes i and u, respectively) TUT regions, we extracted their DNA sequences and the corresponding open reading frames. We also aligned whole-genome DNA sequences from four mammalian species including human (hg18), mouse (mm8), dog (canFam2), and cow (bosTau2) that were generated by Multiz program [66] and downloaded from the UCSC genome browser [67]. Transfrags were classified as either coding or noncoding by an empirical integrative approach based on four distinct predictive tools: phylogenetic coding substitution frequency (PhyloCSF), HMMER, CPC, and txCdsPredict [67–70]. These four tools aimed to predicting the coding potential of a given amino acid or nucleic acid sequence based on: (i) phylogenetic alignments (PhyloCSF); (ii) similarities to known protein domains (HMMER); (iii) a support vector machine-based classifier using several sequence features including similarities to known proteins (CPC); and (iv) a weighting scheme producing a score corresponding to the protein-encoding capacity (txCdsPredict). Transcripts were considered protein-encoding candidates if they had a PhyloCSF score >20, an E value $<10^{-4}$ in HMMER (versus Pfam-A and -B), if they were classified as coding by CPC, or if they showed a txCdsPredict score of >800 (~90% predictive of protein-encoding genes). By combining the results, we were able to organize TUTs into five classes possessing very high (4 of 4 tools predicting protein-encoding potential), high (3 of 4 tools), medium (2 of 4 tools), low (1 of 4 tools), or no (0 of 4 tools) coding potential according to whether transcripts were considered protein-encoding by 4, 3, 2, 1, or none of the four predictive tools, respectively.

Nearest-Neighbor Analysis

For each TUT, the nearest known protein-encoding genes located upstream and downstream were identified without distance restriction. This resulted in a list of associations between TUTs and protein-encoding genes that was exploited for expression-based relationship analysis using the Pearson correlation coefficient as previously described [23, 29] and Gene Ontology (GO) [71] term enrichment analysis using AMEN. The correlation coefficient was calculated by considering pairs of neighboring genes for which both loci were detectable (expressed at levels above background expression cutoff value) in at least one testicular cell type. For GO term analysis, enrichments were estimated with the Fisher exact probability, using a Gaussian hypergeometric test implemented in AMEN [64]. A term was considered significantly enriched in a group of genes if the false discovery rate-corrected P value was $\leq$0.01 and the number of genes bearing this annotation was $\geq$5. Given the small numbers of TUTs and known annotated lncRNAs in the somatic and mitotic expression clusters (P1–P3), no GO term enrichment could be calculated.

Data Access

The RNA-seq data files were submitted to NCBI Sequence Read Archive (SRA) and to NCBI Gene Expression Omnibus (GEO) under accession numbers SRP026340 and GSE48321, respectively. All data are also accessible through RGV [54]. Selected TUTs were deposited with the GenBank

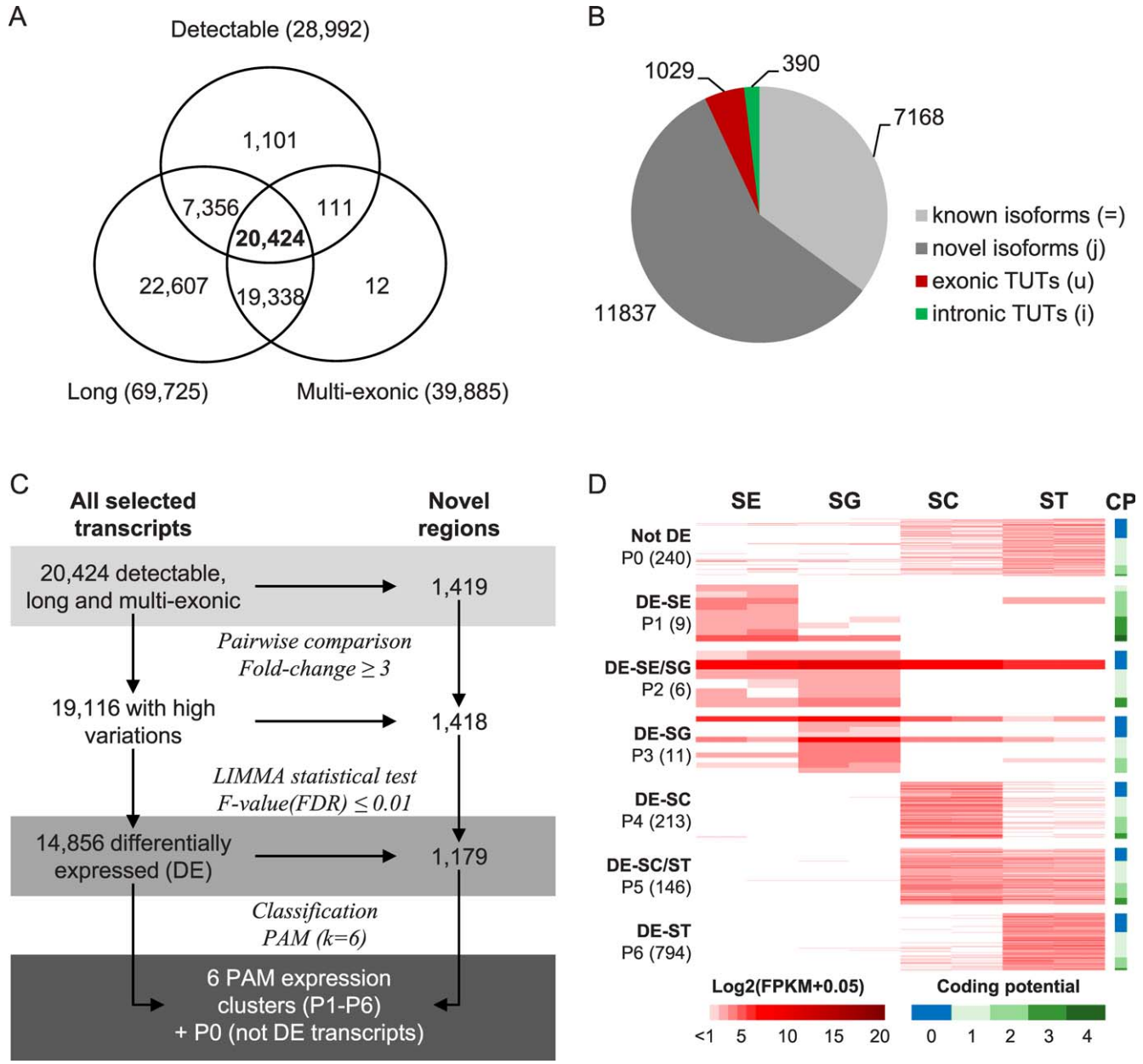

FIG. 1. Profiling the testis-expressed unannotated transcripts (TUTs). **A**) Venn diagram illustrates the three-step refinement strategy used to select a high-confidence set of 20 424 long (≥200 bp), multi-exonic (≥2 exons), and detectable transfrags. **B**) Pie chart shows the proportion of known isoforms of annotated loci (class code =), novel isoforms of annotated genes (j), intronic (i), and intergenic (u) TUTs selected after the refinement strategy. **C**) Flow chart summarizing the filtration steps and clustering strategy used to select significantly differentially expressed high-confidence transfrags (left), including TUTs (right). The total number of selected transcripts is given for each filtration step. **D**) A false-color heatmap summarizes the 7 expression patterns defined according to the abundance of TUTs in the four testicular cell types (columns): Sertoli cells (SE), spermatogonia (SG), spermatocytes (SC), and spermatids (ST). Each line corresponds to a transfrag. For each expression pattern, the number of TUTs is indicated on the left. Log2 FPKM values are displayed according to the color code (bottom left). The last column indicates the degrees of protein-encoding potential (CP) on a color scale (bottom right).

Transcriptome Shotgun Assembly sequence database as BioProject no. PRJNA209702.

Experimental Validation

Reverse transcription PCR. Complementary DNA was obtained from 4-μg aliquots of DNase-treated RNA (DNase I; Promega) using random hexamers and Moloney murine leukemia virus reverse transcriptase (Invitrogen). Conventional PCR was performed using *Taq* polymerase (Qiagen), a Peltier thermocycler (Labgene), and the following oligonucleotide primers (Eurogentec) for the given transfrags: TCONS_00074622 (exons 4–5), forward primer 5′-GAG-CTC-CTA-AAG-GCC-GAG-TT-3′, and reverse primer 5-'GTC-TGC-ACC-CTG-CCA-TAT-TT-3′; and TCONS_00083977 (exons 1–4), forward primer 5′-CAG-GCG-AGT-GGT-CCA-GTA-AT-3′, and reverse primer 5′-AGG-CAG-CGT-CTG-GAG-ATA-AG-3′. PCR products were then resolved on 1.5% agarose gel.

In situ expression analysis. RT-PCR products corresponding to TCONS_00074622 and TCONS_00083977 were gel-purified using Qiaquick Gek extraction kit (Qiagen), cloned into the pCR II-TOPO vector, and used to transform Mach1-T1 *Escherichia coli* (Topo cloning kit for sequencing; Invitrogen). Clones were screened by PCR and sequenced. Sense and antisense riboprobes were generated from SP6 or T7 promoters and labeled with digoxigenin-UTP (Boehringer Mannheim).

Expression levels of TCONS_00074622 and TCONS_00083977 in rat testis were analyzed by in situ hybridization (ISH) using antisense or sense riboprobes at 0.8 ng/μl as previously described [1]. Bound probe was detected

with an alkaline-phosphatase-conjugated anti-digoxigenin antibody at 1:500 dilution (Boehringer Mannheim) and *S*-bromo-4-chloro-3-indolyl phosphate (50 mg/ml) and nitro blue tetrazolium (75 mg/ml) as substrates (Boehringer Mannheim) for 16 h at room temperature.

RESULTS

Assembly of High-Confidence Testicular Transcriptome in Rat Revealed Hundreds of Novel Unannotated Transcribed Regions

To identify, quantify, and characterize novel TUTs with potential functions during male germline differentiation, we performed large-scale, paired-end RNA sequencing experiments at various stages of rat spermatogenesis. Similar to our previous study [2], RNAs were extracted in duplicate from three male germ cell populations that marked three important developmental steps: (i) mitotic spermatogonia; (ii) meiotic pachytene spermatocytes; and (iii) early spermatids undergoing spermiogenesis. RNA was also extracted from somatic testicular cells (i.e., Sertoli cells).

On average, ~35 million reads per sample (~279 million reads in total) were generated (Supplemental Table S1). We assembled and quantified transcripts using the Tuxedo suite [61] (see *Materials and Methods*). Approximately 80% of reads (~224 million reads) were correctly paired and aligned to the rat genome sequence, which notably covered 911 339 splice junctions for use in transcriptome assembly. Aligned reads were then assembled into cell-specific transcriptomes and subsequently combined into a unique set of 122 262 nonredundant transfrags from 53 409 loci across all cell types. The resulting transfrags were classified by comparison with a comprehensive list of annotated coding and noncoding transcripts. We finally selected those corresponding to novel intronic (loci falling entirely within a reference intron and without exon–exon overlap with another known locus) or intergenic TUTs and compared them to known annotated protein-encoding and noncoding transcripts (see Supplemental Table S2).

The intrinsic properties of read alignment processes and potential contamination by unspliced pre-mRNA and genomic DNA can lead to erroneously assembled transcripts [43]. To reduce the number of such artifacts, we developed a highly stringent filtering strategy based on transcript abundance, transcript length, and number of exons (Fig. 1A; see *Materials and Methods*). These selection criteria resulted in a final set of 20 424 high-confidence, long, multi-exonic, nonredundant transcript isoforms expressed during rat spermatogenesis, corresponding to 11 116 loci, including: (i) 7168 known isoforms of annotated loci, including 6915 coding for protein; (ii) 11 837 novel isoforms of annotated loci, including 11 294 coding for protein; and (iii) 390 intronic and 1029 intergenic TUTs (Fig. 1B). All subsequent analyses were conducted using this final set of high-confidence transfrags (see Supplemental Table S3).

TUTs and lncRNAs Accumulate During Meiotic and Spermiogenic Stages

We next applied several statistical filtration steps to study global expression dynamics of the reconstructed transcripts during spermatogenesis (Fig. 1C). Among the 20 424 high-confidence, nonredundant transcripts, 14 856 (72.7%) were identified as being differentially expressed (DE), including 13 677 known or novel isoforms of 8560 annotated loci and 327 intronic and 852 intergenic TUTs.

The DE transfrags were further divided into six expression patterns: those with their highest expression in Sertoli cells (pattern P1 named DE-SE, 1780 transfrags, including 9 TUTs), in both Sertoli cells and spermatogonia (P2 named DE-SE/SG, 1737 transfrags, including 6 TUTs), in spermatogonia (P3 named DE-SG, 1341 transfrags, including 11 TUTs), in spermatocytes (named DE-SC, 3327 transfrags, including 213 TUTs), in both spermatocytes and spermatids (P5 named DE-SC/ST, 1898 transfrags, including 146 TUTs), or in spermatids (P6 named DE-ST, 4773 transfrags, including 794 TUTs) (Fig. 1D; Supplemental Fig. S1, A-D; and Table 1). The 5568 unpatterned candidate transcripts (including 240 TUTs) showing no significant differential expression between cell types, were grouped in P0. Notably, the proportion of transfrags with strongest expression in meiotic and postmeiotic germline samples (P4–P6) for both TUTs (~81.3%) and known annotated lncRNAs (~73.9%) was significantly higher than that of known annotated protein-encoding genes (~45.5%, $P < 10^{-40}$) (Fig. 1D; Supplemental Fig. S1, A–D; and Table 1).

A parallel experiment using rat exon 1.0 ST GeneChips was conducted to evaluate the robustness of the RNA-seq data: the expression profiles found for both known coding and known noncoding loci were widely confirmed. The majority (89%) of genes identified in the RNA-seq analysis as being differentially expressed indeed exhibited an expression correlation of $\geq$0.5 with profiles obtained in the exon array experiment (see Supplemental Fig. S1).

Genomic Characterization of TUTs Confirmed lncRNA Features and Revealed Unusually Longer Exon Length for Meiosis-Induced lncRNAs

To determine whether TUTs expressed during spermatogenesis in the rat displayed features similar to those of known lncRNAs (e.g., short length, low conservation, low expression level, or very low coding potential [23, 27, 29]), we annotated each transcript isoform with a comprehensive list of traits.

Size and compositional characteristics. We found that both TUTs (first quartile [q1] = 391 bp, median (med) = 570 bp, third quartile (q3) = 862 bp) and known lncRNAs ([q1 = 449, med = 683, q3 = 1142]) expressed during spermatogenesis were less than half the size (cumulative exon size) of known mRNAs (q1 = 1068, med = 1811, q3 = 2,904; $P < 10^{-100}$, Wilcoxon signed-rank test) (Fig. 2A). Moreover, both TUTs (q1 = 2, med = 2, q3 = 3) and known lncRNAs (q1 = 2, med = 3, q3 = 4) had approximately 3–4 times fewer exons than known mRNAs (q1 = 5, med = 8, q3 = 13; $P < 10^{-200}$) (Fig. 2B). The total gene sizes (cumulative exon and intron size) of TUTs (q1 = 1,871 bp, med = 4305, q3 = 10 357) and known lncRNAs (q1 = 3130, med = 7086, q3 = 17 251) were also significantly less than half that of known mRNAs (q1 = 8054, med = 19 431, q3 = 42 857; $P < 10^{-60}$) (Fig. 2C). Because of space constraints, the total gene size for intronic TUTs (q1 = 1538, med = 3234, q3 = 6149) was two-thirds that of intergenic TUTs (q1 = 2022, med = 5126, q3 = 12 416; $P < 10^{-10}$), although transcript sizes were not significantly different (Fig. 2E). Analysis of the sequence features of TUTs and known lncRNAs indicated that they have a significantly lower GC content than known mRNAs (median GC content of 48.3% for TUTs and 48.8% for known lncRNAs versus 50.6% for known mRNAs; $P < 10^{-17}$) (Fig. 2F).

For both meiotic (DE-SC or P4) TUTs (q1 = 524, med = 979, q3 = 2073) and meiotic known lncRNAs (q1 = 571, med = 991, q3 = 2063), the transcript sizes were twice as large as those showing other expression patterns (P1–P3 and P5 and P6;

TABLE 1. Classification of 20 424 high-confidence transcripts according to their expression pattern.[a]

		Expression patterns[b]							
Classification	Total	Not DE P0	DE-SE P1	DE-SE/SG P2	DE-SG P3	DE-SC P4	DE-SC/ST P5	DE-ST P6	% P4–P6
Total no. of transcript fragments	20 424	5568	1780	1737	1341	3327	1898	4773	49.0%
Known isoforms (=)	7168	1670	1036	1170	708	1160	488	936	36.0%
Coding	6915	1623	1028	1161	693	1113	462	835	34.9%
Noncoding	215	40	4	4	11	37	22	97	72.6%
Other	38	7	4	5	4	10	4	4	47.4%
Novel isoforms (j)	11 837	3658	735	561	622	1954	1264	3043	52.9%
Coding	11 294	3529	726	555	615	1873	1184	2812	52.0%
Noncoding	459	106	3	2	6	65	65	212	74.5%
Other	84	23	6	4	1	16	15	19	59.5%
TUTs (i and u)	1419	240	9	6	11	213	146	794	81.3%
Intronic (i)	390	63	5	5	4	39	35	239	80.3%
CP = 0	125	21	0	2	2	15	12	73	80.0%
Low, CP = 1	166	29	0	2	0	7	12	116	81.3%
Medium, CP = 2	80	13	2	0	2	14	8	41	78.8%
High, CP = 3	16	0	2	1	0	3	3	7	81.3%
Very high, CP = 4	3	0	1	0	0	0	0	2	66.7%
Intergenic (u)	1029	177	4	1	7	174	111	555	81.6%
CP = 0	310	60	0	0	2	39	22	187	80.0%
Low, CP = 1	447	84	1	1	4	69	45	243	79.9%
Medium, CP = 2	208	27	2	0	1	49	30	99	85.6%
High, CP = 3	62	6	1	0	0	16	13	26	88.7%
Very high, CP = 4	2	0	0	0	0	1	1	0	100.0%

[a] For each of the seven expression patterns, the number of transcripts, the annotation provided by Cuffcompare, and the protein-encoding status are given.
[b] CP, coding potential; DE, differentially expressed; j, annotated genes; i, intronic TUTs; SC, spermatocytes; SE, Sertoli cells; SG, spermatogonia; ST, spermatids; u, intergenic TUTs.

$P < 0.001$) (Fig. 2B). This difference was not due to a greater number of exons but to a significantly longer exon length (median length of 371 bp and 297 bp for meiotic TUTs and known lncRNAs, respectively, versus ~200 bp for all other transcriptional events, $P < 0.01$) (Fig. 2, B and D). Importantly, this difference was not found for known meiotic protein-encoding transfrags (q1 = 149, med = 203, q3 = 320; $P < 0.001$).

Sequence conservation. To assess the evolutionary sequence conservation of the transcript isoforms identified, we computed an empirical score by averaging the base-by-base phastCons conservation scores calculated among nine vertebrates as provided by the UCSC genome browser [59]. In agreement with previous observations [25, 27, 29, 30, 37, 72], most TUTs and known lncRNAs expressed during spermatogenesis showed substantially less exon conservation than known mRNAs (median conservation scores of 0.024 for TUTs and 0.049 for known lncRNAs versus 0.609 for known mRNAs; $P < 10^{-250}$) (Fig. 2G).

Abundance and specificity. Although we focused on higher abundance transfrags because the associated data are more reliable, we still observed a lower expression level in testicular cells of TUTs (median of the highest FPKM for all testicular cell samples of 9.9) than that of known lncRNAs (median FPKM of 12.6; $P < 10^{-12}$) and, in an even more pronounced manner, than that of known mRNAs (median FPKM of 14.8; $P < 10^{-77}$) (Fig. 2H). These observations are consistent with the weak expression of lncRNAs in several vertebrates and biological systems [23, 27, 29, 73]. We calculated an expression specificity score based on the Shannon (information theoretic measure) entropy Q value to estimate the abundance specificity in the various testicular cell types [74] as previously suggested [23, 29]. TUTs showed a significantly higher cell type specificity (median Shannon entropy-based specificity score = 0.632) than known lncRNAs (median score = 0.842; $P < 10^{-16}$) and a much higher specificity than known mRNAs (median score = 1.296; $P < 10^{-200}$) (Fig. 2I). Overall, expression levels of TUTs and known lncRNAs were thus significantly more restricted than that of transcripts corresponding to known protein-encoding loci ($P < 10^{-100}$). This is consistent with previous observations of lncRNA specificity in other vertebrate systems [29].

Chromosomal localization. Protein-encoding loci on the X chromosome are silenced by a phenomenon called meiotic sex chromosome inactivation (MSCI, for review see ref. [75]). On the other hand, the X chromosome is enriched for genes expressed in testicular somatic cells, spermatogonia, and postmeiotic cells [2, 76–78]. We analyzed the chromosomal localization of the selected high-confidence transfrags and found that not a single X-linked annotated protein-encoding locus escaped the MSCI silencing effect in spermatocytes (0 genes in the meiotic expression pattern P4/DE-SC were on the X chromosome, although 66 would be expected by chance; $P < 10^{-31}$), whereas somatic, spermatogonial, and postmeiotic expression patterns (P1–P3 and P6) were found to be enriched for such X-linked transfrags (see Supplemental Fig. S2, A–F). This validates the meiotic expression pattern that contains transcripts showing peak induction in pachytene spermatocytes. Importantly, we found the same result for TUTs: not a single X-linked TUT belonged to the meiotic expression pattern P4, although five would be expected by chance ($P = 0.008$; see Supplemental Fig. S2, G–L).

Protein-encoding potential analysis. The high-confidence set of unannotated transcribed regions we identified are likely to correspond to either coding or noncoding genes. To assess the protein-encoding potential of our set of intronic and intergenic TUTs, we developed a pipeline based on the results of four predictive tools. We found that nearly three-quarters of the TUTs exhibited no (all four tools predicting no coding potential (CP = 0; 435 TUTs) or very low (3 of 4 tools predicting no CP (1613 TUTs) (Table 1). The coding potential of a significant proportion of the remaining TUTs is also

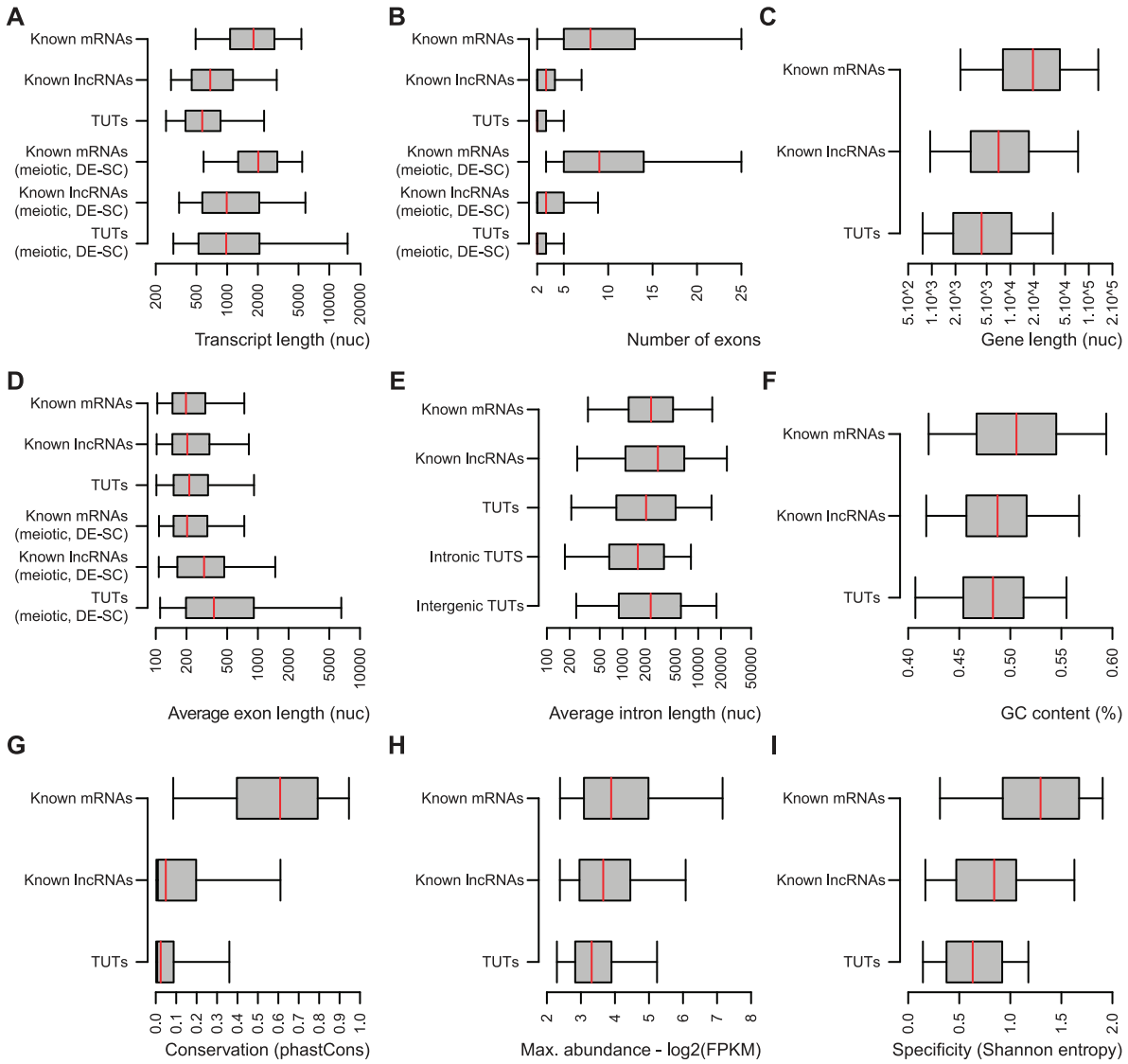

FIG. 2. Genomic features of testicular transcripts. Known protein-encoding transcripts (mRNAs) and known lncRNAs were compared to TUTs. Transcripts belonging to the meiotic cluster are indicated with the corresponding expression pattern number (DE-SC, P4). Box plots summarize the distributions of: transcript length (**A**), number of exons (**B**), gene length (**C**), mean exon length (**D**), mean intron length (**E**), percentage of GC-content (**F**), sequence conservation (phastCons score) among vertebrates (**G**), maximum abundance in samples in log2 (FPKM + 0.05) (**H**), and cell specificity measures based on Shannon entropy (**I**). Note that the lower the value of Shannon entropy, the more the expression is restricted to one cell type. **A**, **C**, **D** and **E**) Lengths are shown in nucleotides (nuc) on a logarithmic scale (x-axis).

dubious given the relatively low scores (just above the thresholds) obtained with the different tools.

Altogether, the genomic characterization of the unannotated transcribed regions we identified suggests that most TUTs are therefore likely to correspond to newly identified lncRNAs.

Intergenic TUTs Are Highly Distant from Their Neighboring Protein-Encoding Genes

Distance to neighboring protein-encoding genes. To study how TUTs and known lncRNAs are related to their protein-encoding neighbors, we identified their nearest upstream and downstream known protein-encoding genes, without distance restriction. We found that 44.4% of the intronic TUTs (173 of 390), 74.5% of the intergenic TUTs (767 of 1029), and 51.3% of the known lncRNAs (346 of 674) were mapped in genomic regions >10 kb away from any known protein-encoding loci. Intergenic TUTs were found to be at least three to five times more distant from any coding loci (q1 = 9699 bp, med = 38 439, q3 = 122 755) than intronic TUTs (q1 = 1400, med = 7845, q3 = 27 030; $P < 10^{-35}$) or known lncRNAs (q1 = 593, med = 11 138, q3 = 55 577; $P < 10^{-26}$). Therefore, *cis* regulation of nearby protein-encoding genes is unlikely, and possibly, many of these stand-alone intergenic TUTs may instead act by mechanisms of *trans* regulation.

Transcriptional correlation with neighboring protein-encoding loci. To test for functional links among TUTs or known lncRNAs and their neighboring protein-encoding loci, we analyzed correlations among their abundance levels (Pearson correlation coefficient, r) [26, 29]. Consistent with the analysis by Cabili et al. [23], but not with those of other studies [26, 29], we observed that TUTs and known lncRNAs tended to correlate more positively with their neighbor protein-encoding loci ($r = 0.629$ and 0.549, respectively) than pairs of known protein-encoding genes did ($r = 0.360$; $P < 0.003$). Moreover, we did not detect greater correlation between

intergenic TUTs and their neighbors ($r = 0.597$) than for intronic TUTs and their neighbors ($r = 0.653$; $P = 0.425$). Possibly, many of the TUTs and known lncRNAs are positive regulators of neighboring protein-encoding genes or vice versa. They might also be under the control of the same enhancer elements.

Association with biological processes of neighboring protein-encoding genes. Long noncoding RNAs are preferentially located next to genes associated with specific processes [23, 25, 29, 30]. We therefore analyzed the GO terms of genes that were neighbors of TUTs and known annotated lncRNAs expressed during spermatogenesis. We found significant enrichment of broad annotation terms among protein-encoding neighbors of TUTs and known lncRNAs with a peak expression in postmeiotic germ cell types (DE-ST, P6) but not among the neighbors of those belonging to the other expression patterns (P0–P5) (Fig. 3). Genes next to postmeiotic known lncRNAs and intergenic TUTs were significantly associated with embryonic development (hypergeometric P value adjusted with the false discovery rate method = 0.008) and morphogenesis ($P = 0.001$). We also observed significant enrichment of cell junction ($P = 0.003$) and synaptic ($P = 0.0005$) subcellular components and system development ($P = 0.003$) among neighbors of intronic post-meiotic TUTs. This analysis identified several groups of postmeiotic intergenic TUTs neighboring protein-encoding loci of distinct functional categories such as: (i) embryonic developmental processes including organ ($P = 4 \times 10^{-5}$), tube ($P = 4 \times 10^{-6}$), lung ($P = 0.007$), and nervous system ($P = 4 \times 10^{-5}$) development; (ii) cellular differentiation ($P = 5 \times 10^{-6}$) processes including regulation of neuron ($P = 7 \times 10^{-5}$), glial cell ($P = 0.004$), and myoblast ($P = 0.008$) differentiation; (iii) regulation of cell migration ($P = 0.005$) and proliferation ($P = 4 \times 10^{-6}$); (iv) cellular communication processes such as cell communication ($P = 3 \times 10^{-5}$), cell-cell adhesion ($P = 0.01$), signal transduction ($P = 0.0005$), regulation of signaling ($P = 4 \times 10^{-5}$), regulation of cell communication ($P = 6 \times 10^{-5}$), and regulation of kinase activity ($P = 0.004$); (iv) transcriptional circuitry such as regulation of transcription from RNA polymerase II promoter ($P = 0.0003$) and sequence-specific DNA-binding transcription factor activity ($P = 0.0001$); and (v) phosphoregulation terms such as regulation of phosphorylation ($P = 0.003$) and more specifically regulation of protein phosphorylation ($P = 0.001$).

TUTs Show Distinct Subcellular Localization Patterns in Germ Cells

We further investigated the cell type specificity of the TUTs identified in our transcriptome analysis by studying selected candidates, using RT-PCR and RNA ISH. The first candidate we investigated, TCONS_00074622, maps to chromosome 3 (positions 115 212 490–115 230 020), is composed of 5 exons with a total transcript length of 4951 bp, and belongs to the meiotic expression pattern (DE-SC, P4). A coding potential was predicted for this TUT by three tools but each time with a score just above the specified threshold. It is poorly conserved across vertebrates (Fig. 4, A and B). Its expression pattern in 4 testicular cell types and 7 normal tissues was investigated by RT-PCR (Fig. 4C). The strong and specific expression of TCONS_00074622 in spermatocytes and total testis, relative to those in the 3 other cell types and 6 other normal tissues, was confirmed. ISH further confirmed the meiotic expression pattern of TCONS_00074622 and revealed that it was preferentially localized in the nuclei of spermatocytes where it seemed to be associated with chromatin (Fig. 4D). The second TUT investigated, TCONS_00083977, maps to chromosome 4 (from 81 546 232–81 568 496), is composed of 8 exons with a total transcript length of 980 bp, and displays a postmeiotic peak expression (DE-ST, P6). It has no apparent protein-encoding potential and shows slightly greater sequence conservation than TCONS_00074622 among vertebrates (Fig. 4, E and F). Both RT-PCR and ISH analyses confirmed the spermatid-specific expression pattern (Fig. 4, G and H). TCONS_00083977 appeared to accumulate in perinuclear, cytoplasmic structures of postmeiotic haploid round spermatids that presumably correspond to the germline chromatoid bodies (Fig. 4H). This cytoplasmic organelle is a germ-cell-specific RNA-processing granule that plays an important role in post-transcriptional and translation regulation during the late steps of spermatogenesis [79]. These two experimental validations confirm that, like known lncRNAs, TUTs also can be localized in particular subcellular domains in specific germ cell types.

DISCUSSION

We report the first outcome of high-resolution transcriptional profiling of three different germ cell populations as well as somatic Sertoli cells in the rat. Large-scale RNA-seq experiments of these four testicular cell types allowed us to reconstruct 20 424 high-confidence coding and noncoding transcript isoforms from 11 116 loci. Notably, we recovered 7168 transcripts already present in RefSeq, Ensembl, AceView, and UCSC databases and identified 11 837 novel isoforms of known loci. We also reconstructed 1419 high-confidence TUTs, with no previous annotations in the databases mentioned above. Finally, we captured some of the dynamic changes in abundance levels of each of these transcripts as spermatogenesis proceeds.

In addition, we exploited another transcriptomic dataset (exon array technology; Affymetrix) that we used to validate RNA-seq data. We found that expression patterns of transcripts as reconstructed from the RNA-seq analysis correlated well with those obtained using microarray analysis (see Supplemental Fig. S1, A–D). Although GeneChip technology is by definition not designed to detect unannotated loci, the fact that it widely validates the expression profiles of annotated transcripts obtained from our sequencing data (including both coding and noncoding annotated genes) is clear evidence of the reliability of our data and the robustness of the TUTs we identified. These data, which complete and extend those of recent publications [9–12, 80], provide the most comprehensive annotation of the mammalian germ cell transcriptome currently available and contribute to the discovery of novel unannotated loci expressed during spermatogenesis in mammals.

Recently, RNA-seq analysis of the first wave of spermatogenesis in the mouse testis was conducted [9]. In that study, 1953 differentially expressed genes were highlighted, of which 1766 (90.4%) were conserved in rat, and 953 genes (53.5%) significantly overlapped the 6271 loci displaying a significant abundance variation we identified (hypergeometric P value $< 10^{-105}$). Almost 1000 known noncoding genes were also identified as being differentially expressed. However, no or little attention was given to potential new unannotated transcribed regions and to the identification of unknown genes.

In this study we focused on a stringent set of 1419 long and multi-exonic TUTs and thoroughly characterized the novel potential loci. We found that most TUTs were likely to consist of novel lncRNAs. When we assessed their protein-encoding potential, most of them were indeed predicted to possess no or very low coding potential. Additionally, among the remaining

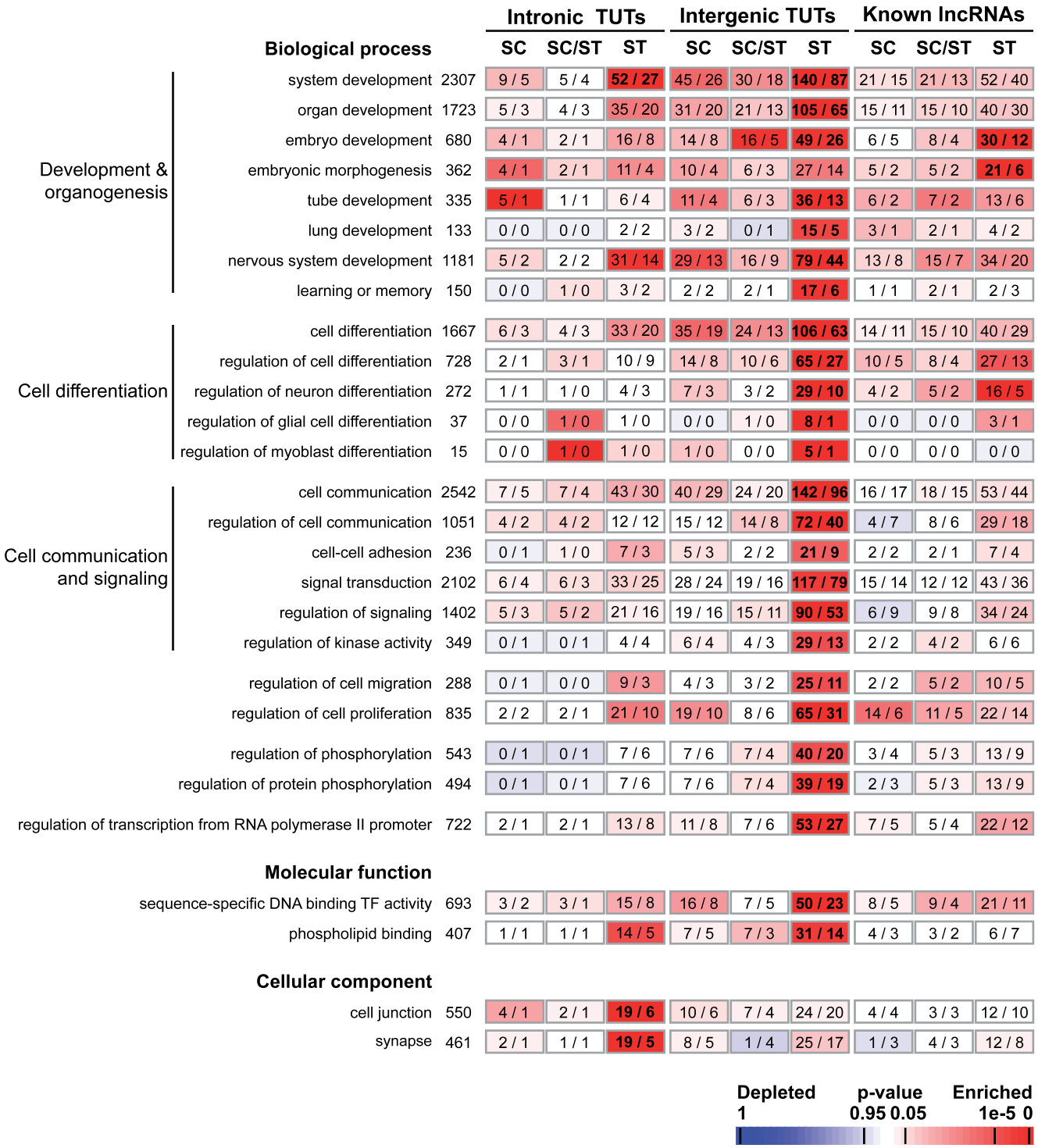

FIG. 3. Functional analysis of known protein-encoding genes neighboring the testis-expressed unannotated transcripts (TUTs). Significantly enriched Gene Ontology (GO) terms among the annotations of the nearest upstream and downstream known protein-encoding genes (in an orientation-independent manner) of differentially expressed TUTs are shown from the meiotic and post-meiotic expression patterns (SC, SC/ST and ST). Total numbers of known protein-encoding genes (NCBI Entrez gene identifiers) neighboring TUTs and lncRNAs are given within rectangles as observed (left) and as expected by chance (right). A color scale illustrating *P* values is provided for enriched (red) and depleted (blue) terms. Numbers in bold indicate a significant over- or under-representation for a given GO term. Note that numbers of transcript isoforms within the somatic and mitotic expression patterns (DE-SE, DE-SE/SG and DE-SG) were too small for such statistical analysis.

ones (medium to very high coding potential), even those that were predicted to be coding by the four tools could be dubious. For instance TCONS_00074622, one of the two candidates we investigated further, was predicted to have a coding potential by three different tools. However, each time, the score obtained for this TUT was only just above the specified threshold. Another line of evidence that TUTs might correspond mostly to noncoding transcriptional events is that they share many of the genomics characteristics observed for lncRNAs in other vertebrates (e.g., relatively short length, low exon number,

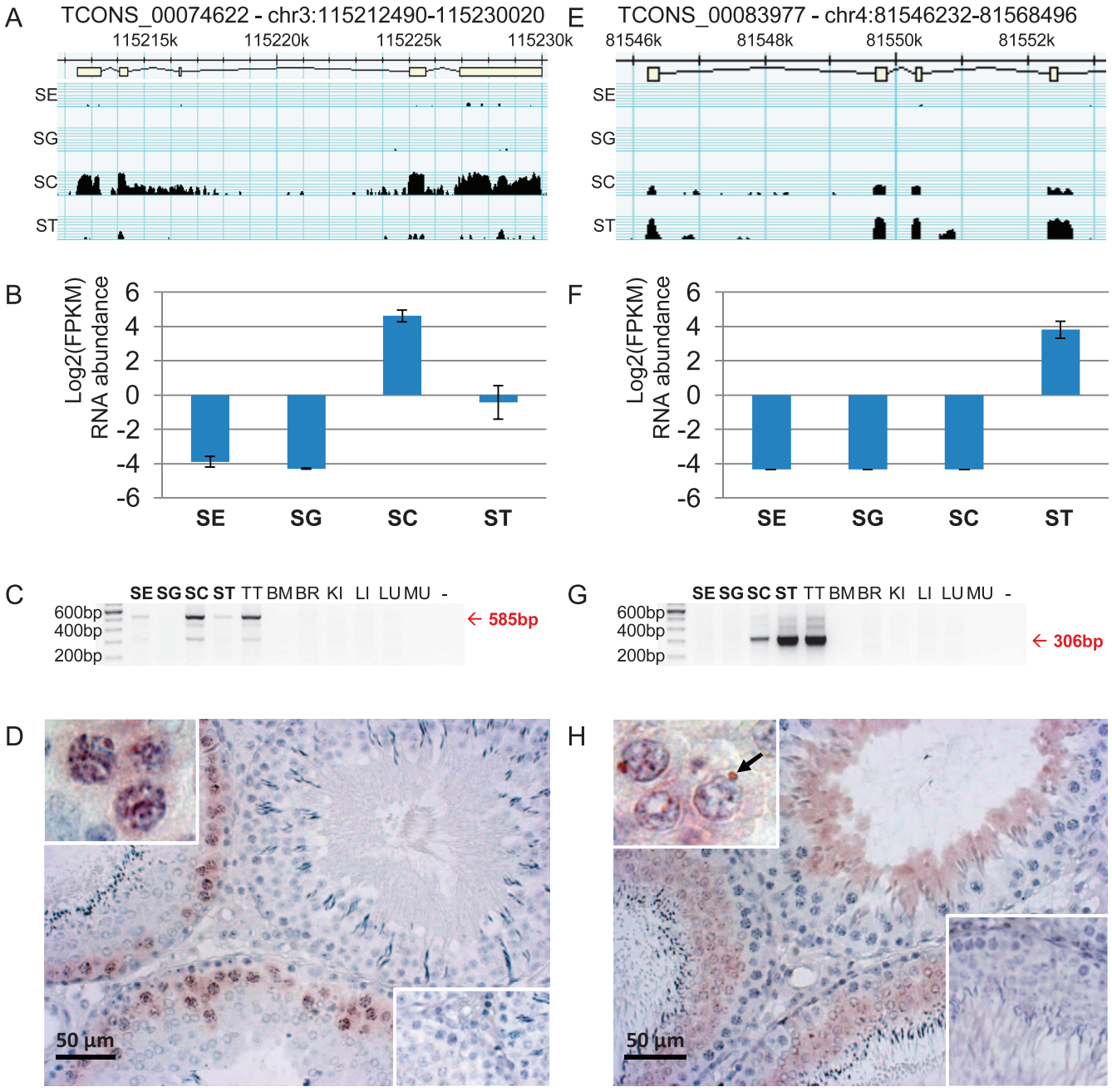

FIG. 4. Cell- and subcell-specific expression patterns of two testicular unannotated transcripts (TUTs) are shown. The expression patterns of two TUTs TCONS_00074622 (**A**–**D**) and TCONS_00083977 (**E**–**H**) were investigated. **A** and **E**) Gene structures are shown for both TUTs and histograms of the numbers of RNA-seq reads that aligned the corresponding genomic locations across the different samples (y-axis ranges from 0 to 8, log2 [FPKM]) (adapted from the RGV; http://rgv.genouest.org). **B** and **F**) RNA-seq abundance levels (y-axis, log2 [FPKM]) of both TUTs are shown in the different testicular cell types (x-axis). **C** and **G**) RT-PCR results are shown for Sertoli cells (SE), spermatogonia (SG), spermatocytes (SC), spermatids (ST), and 7 tissues including bone marrow (BM), brain (BR), kidney (KI), liver (LI), lung (LU), and muscle (MU). − (minus sign) = RT-negative control. **D** and **H**) Testicular ISHimages with probes specific for the selected TUTs are shown. Negative control images (insets) show a lack of signal were obtained by using the sense ribonucleotide probe. RT-PCR and ISH analyses confirmed that the expression of TCONS_00074622 is restricted to pachytene spermatocytes and that of TCONS_00083977 to round spermatids. ISH experiments also showed TCONS_00074622 and TCONS_00083977 to be enriched in the nuclear chromatin of pachytene spermatocytes and in chromatoid bodies of round spermatids, respectively (×5 magnification view shown at the top left corner of each picture). The black arrow in the inset (**H**) shows the accumulation of TCONS_00083977 in perinuclear, cytoplasmic structures that presumably correspond to the germline chromatoid bodies.

low GC content, low sequence conservation [comparable to that of introns], low abundance, and highly temporally and spatially restricted expression patterns). However, like known lncRNAs, the transcript length and exon number of TUTs may be underestimated because of partial transcript reconstruction. We also noticed that, similarly to known lncRNAs, postmeiotic TUTs are preferentially transcribed in the vicinity of genes associated with broad GO annotation terms including transcriptional regulation, embryo development, and cell differentiation [29]. Taken together, these results suggest that our set of TUTs consists essentially of newly identified lncRNAs.

Analysis of the individual expression patterns confirmed or revealed three particular features of both TUTs and lncRNAs. First, we showed that not a single meiosis-induced (P4) TUT or

known lncRNA escapes the silencing effects of MSCI in spermatocytes. Indeed, no locus on the mammalian X chromosome showed a peak expression in meiotic germ cells due to the MSCI phenomenon (for review see ref. [75]). Like those of other genome-wide studies [2, 76], our results confirm this observation for protein-encoding loci associated with a meiotic expression pattern and expand it to X-linked noncoding genes.

Second, most TUTs and known lncRNAs reconstructed in our dataset accumulate in meiotic and post-meiotic germ cells. These data markedly reinforce previous findings showing an enrichment of lncRNAs observed to coincide with the appearance of spermatocytes and spermatids in mouse [9, 10] and human [23] testes. Such accumulation of ncRNAs during gametogenesis has also been observed during sporulation (an analogous biological process) in *Saccharomyces cerevisiae* [81]. Some TUTs accumulated in postmeiotic germ cells may also belong to the poorly characterized set of transcripts paternally provided by the sperm to the early embryo [29, 82].

Third, due to their unusually greater exon length, meiosis-induced TUTs were on average longer than TUTs, showing peak expression in other testicular cell types. Of note, this characteristic was also observed for meiotic known lncRNAs but not for meiotic mRNAs. Exon sizes for protein-encoding genes in vertebrates are usually limited to 200–300 bp. Although longer exons have been shown to be associated with the most recently evolved genes [83], meiotic TUTs do not tend to be less conserved than TUTs with other expression patterns and with a shorter average exon size. The meiosis-induced TUT we validated, TCONS_00074622, is composed of five exons with an average exon size of ~990 bp. ISH indicated that this TUT is likely to be associated with chromatin in pachytene spermatocytes in the rat. Possibly, this feature may be functionally related to a role of meiosis-specific noncoding transcripts in mediating the recognition of homologous chromosomes for pairing during meiosis, as in the fission yeast *Schizosaccharomyces pombe* [84]. This property may thus be indicative of a general functional characteristic of meiotic lncRNAs.

As observed for protein-encoding loci [1, 2], TUTs preferentially expressed in meiotic and spermiogenic stages are likely to act in a tissue- and cell type-specific manner. The significant accumulation of TUTs and annotated lncRNAs during the last phases of spermatogenesis suggests they may be involved in this process. We also showed that two TUTs displayed specific subcellular localization patterns, which may reflect the putative regulatory functions of these transcripts during male germline differentiation in mammals, as suggested by other biological systems and eukaryotic species [85, 86].

In summary, we report high-resolution RNA profiling and high-confidence characterization of many novel unannotated transcribed regions encoding mostly lncRNAs expressed during rat spermatogenesis. In addition to a significant contribution to genome annotation of a major mammalian model organism, this study allowed us to determine that known and novel lncRNAs with a peak expression during meiosis constitute a distinct class of noncoding transcripts with a longer exon length. This extends the most recent publications [9, 10] and provides a useful resource for future genetic and genomic investigations of the roles of testicular lncRNAs in mammalian spermatogenesis, and fertility.

ACKNOWLEDGMENT

We thank all members of the SEQanswers forums for helpful advice; Steven Salzberg and Cole Trapnell for continuous support with the Tuxedo suite; members of the UCSC genome browser; and Bernard Jost, Céline Bordet, Emmanuelle Becker, Séverine Mazaud-Guittot, and Nathalie Dejucq-Rainsford for discussions and helpful comments on the analysis. Sequencing was performed by the IGBMC Microarray and Sequencing platform, member of the France Genomique program.

REFERENCES

1. Chalmel F, Lardenois A, Evrard B, Mathieu R, Feig C, Demougin P, Gattiker A, Schulze W, Jegou 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.
2. Chalmel F, Rolland AD, Niederhauser-Wiederkehr C, Chung SS, Demougin P, Gattiker A, Moore J, Patard JJ, Wolgemuth DJ, Jegou B, Primig M. The conserved transcriptome in human and rodent male gametogenesis. Proc Natl Acad Sci U S A 2007; 104:8346–8351.
3. Son CG, Bilke S, Davis S, Greer BT, Wei JS, Whiteford CC, Chen QR, Cenacchi N, Khan J. Database of mRNA gene expression profiles of multiple human organs. Genome Res 2005; 15:443–450.
4. Wrobel G, Primig M. Mammalian male germ cells are fertile ground for expression profiling of sexual reproduction. Reproduction 2005; 129:1–7.
5. Eddy EM. Male germ cell gene expression. Recent Prog Horm Res 2002; 57:103–128.
6. Schlecht U, Demougin P, Koch R, Hermida L, Wiederkehr C, Descombes P, Pineau C, Jegou 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.
7. 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.
8. 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.
9. 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.
10. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthes 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.
11. Gan H, Wen L, Liao S, Lin X, Ma T, Liu J, Song CX, Wang M, He C, Han C, Tang F. Dynamics of 5-hydroxymethylcytosine during mouse spermatogenesis. Nat Commun 2013; 4:1995.
12. Margolin G, Khil PP, Kim J, Bellani MA, Camerini-Otero RD. Integrated transcriptome analysis of mouse spermatogenesis. BMC Genomics 2014; 15:39.
13. Encode Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004; 306:636–640.
14. 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.
15. Banfai 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.
16. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011; 470:187–197.
17. Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 2011; 27:2325–2329.
18. Severin AJ, Peiffer GA, Xu WW, Hyten DL, Bucciarelli B, O'Rourke JA, Bolon YT, Grant D, Farmer AD, May GD, Vance CP, Shoemaker RC, et al. An integrative approach to genomic introgression mapping. Plant Physiol 2010; 154:3–12.
19. Hung T, Chang HY. Long noncoding RNA in genome regulation: prospects and mechanisms. RNA Biol 2010; 7:582–585.
20. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet 2009; 10:155–159.
21. Janga SC, Vallabhaneni S. MicroRNAs as post-transcriptional machines and their interplay with cellular networks. Adv Exp Med Biol 2011; 722: 59–74.
22. Zhao Y, He S, Liu C, Ru S, Zhao H, Yang Z, Yang P, Yuan X, Sun S, Bu D, Huang J, Skogerbo G, et al. MicroRNA regulation of messenger-like noncoding RNAs: a network of mutual microRNA control. Trends Genet 2008; 24:323–327.

23. 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.
24. Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Solda 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.
25. 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.
26. 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.
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. 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.
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. 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.
31. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, et al. The transcriptional landscape of the mammalian genome. Science 2005; 309:1559–1563.
32. Niazi F, Valadkhan S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. RNA 2012; 18:825–843.
33. Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. Cell 2009; 136:629–641.
34. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes Dev 2009; 23:1494–1504.
35. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov 2011; 1:391–407.
36. Iglesias-Platas I, Martin-Trujillo A, Cirillo D, Court F, Guillaumet-Adkins A, Camprubi C, Bourc'his D, Hata K, Feil R, Tartaglia G, Arnaud P, Monk D. Characterization of novel paternal ncRNAs at the Plagl1 locus, including Hymai, predicted to interact with regulators of active chromatin. PLoS One 2012; 7:e38907.
37. Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R. Long noncoding RNAs with enhancer-like function in human cells. Cell 2010; 143:46–58.
38. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D, Zhao H, Skogerbo G, Wu Z, et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res 2011; 39:3864–3878.
39. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010; 464:1071–1076.
40. Wan Y, Chang HY. HOTAIR: flight of noncoding RNAs in cancer metastasis. Cell Cycle 2010; 9:3391–3392.
41. Pontier DB, Gribnau J. Xist regulation and function explored. Hum Genet 2011; 130:223–236.
42. Xu C, Yang M, Tian J, Wang X, Li Z. MALAT-1: a long non-coding RNA and its important 3′ end functional motif in colorectal cancer metastasis. Int J Oncol 2011; 39:169–175.
43. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011; 29:742–749.
44. Huarte M, Guttman M, Feldser D, Garber M, Koziol MJ, Kenzelmann-Broz D, Khalil AM, Zuk O, Amit I, Rabani M, Attardi LD, Regev A, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. Cell 2010; 142:409–419.
45. Hung T, Wang Y, Lin MF, Koegel AK, Kotake Y, Grant GD, Horlings HM, Shah N, Umbricht C, Wang P, Kong B, Langerod A, et al. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. Nat Genet 2011; 43:621–629.
46. Tian D, Sun S, Lee JT. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. Cell 2010; 143:390–403.
47. Hannon GJ, Rivas FV, Murchison EP, Steitz JA. The expanding universe of noncoding RNAs. Cold Spring Harb Symp Quant Biol 2006; 71: 551–564.
48. Chen LL, Carmichael GG. Long noncoding RNAs in mammalian cells: what, where, and why? Wiley Interdiscip Rev RNA 2010; 1:2–21.
49. Koziol MJ, Rinn JL. RNA traffic control of chromatin complexes. Curr Opin Genet Dev 2010; 20:142–148.
50. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell 2011; 43:904–914.
51. Valadkhan S, Nilsen TW. Reprogramming of the non-coding transcriptome during brain development. J Biol 2010; 9:5.
52. Pauli A, Rinn JL, Schier AF. Non-coding RNAs as regulators of embryogenesis. Nat Rev Genet 2011; 12:136–149.
53. Bhartiya D, Kapoor S, Jalali S, Sati S, Kaushik K, Sachidanandan C, Sivasubbu S, Scaria V. Conceptual approaches for lncRNA drug discovery and future strategies. Expert Opin Drug Discov 2012; 7:503–513.
54. The Reprogenomics Viewer. http://rgv.genouest.org. Accessed 15 June 2012.
55. Dorval-Coiffec I, Delcros JG, Hakovirta H, Toppari J, Jegou B, Piquet-Pellorce C. Identification of the leukemia inhibitory factor cell targets within the rat testis. Biol Reprod 2005; 72:602–611.
56. 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:D48–55.
57. Pruitt KD, Tatusova T, Brown GR, Maglott DRNCBI. Reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 2012; 40:D130–135.
58. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol 2006; 7(suppl 1): S12 11–14.
59. 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:D64–69.
60. 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.
61. 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.
62. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25:1105–1111.
63. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249–264.
64. Chalmel F, Primig M. The annotation, mapping, expression and network (AMEN) suite of tools for molecular systems biology. BMC Bioinformatics 2008; 9:86.
65. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3. Article3.
66. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 2004; 14:708–715.
67. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform 2013; 14:144–161.
68. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 2007; 35:W345–349.
69. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics 2011; 27:i275–282.
70. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011; 39:W29–37.
71. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25–29.

72. Ponjavic J, Ponting CP, Lunter G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res 2007; 17:556–565.
73. Lee CS, Ungewickell A, Bhaduri A, Qu K, Webster DE, Armstrong R, Weng WK, Aros CJ, Mah A, Chen RO, Lin M, Sundram U, et al. Transcriptome sequencing in Sezary syndrome identifies Sezary cell and mycosis fungoides-associated lncRNAs and novel transcripts. Blood 2012; 120:3288–3297.
74. Schug J, Schuller WP, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ Jr. Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol 2005; 6:R33.
75. Graves JA. Sex chromosome specialization and degeneration in mammals. Cell 2006; 124:901–914.
76. Khil PP, Smirnova NA, Romanienko PJ, Camerini-Otero RD. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. Nat Genet 2004; 36:642–646.
77. Mueller JL, Mahadevaiah SK, Park PJ, Warburton PE, Page DC, Turner JM. The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nat Genet 2008; 40:794–799.
78. Wang PJ, McCarrey JR, Yang F, Page DC. An abundance of X-linked genes expressed in spermatogonia. Nat Genet 2001; 27:422–426.
79. Meikar O, Da Ros M, Korhonen H, Kotaja N. Chromatoid body and small RNAs in male germ cells. Reproduction 2011; 142:195–209.
80. Sendler E, Johnson GD, Mao S, Goodrich RJ, Diamond MP, Hauser R, Krawetz SA. Stability, delivery and functions of human sperm RNAs at fertilization. Nucleic Acids Res 2013; 41:4104–4117.
81. Lardenois A, Liu Y, Walther T, Chalmel F, Evrard B, Granovskaia M, Chu A, Davis RW, Steinmetz LM, Primig M. Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rrp6. Proc Natl Acad Sci U S A 2011; 108:1058–1063.
82. Lalancette C, Miller D, Li Y, Krawetz SA. Paternal contributions: new functional insights for spermatozoal RNA. J Cell Biochem 2008; 104: 1570–1579.
83. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics 2013; 14: 117.
84. Ding DQ, Okamasa K, Yamane M, Tsutsumi C, Haraguchi T, Yamamoto M, Hiraoka Y. Meiosis-specific noncoding RNA mediates robust pairing of homologous chromosomes in meiosis. Science 2012; 336:732–736.
85. Johnstone O, Lasko P. Translational regulation and RNA localization in *Drosophila* oocytes and embryos. Annu Rev Genet 2001; 35:365–406.
86. Long RM, Singer RH, Meng X, Gonzalez I, Nasmyth K, Jansen RP. Mating type switching in yeast controlled by asymmetric localization of ASH1 mRNA. Science 1997; 277:383–387.