RNA profiling of human testicular cells identifies syntenic lncRNAs associated with spermatogenesis

Antoine D. Rolland, B Evrard, T A Darde, C Le Béguec, Y Le Bras, K Bensalah, S Lavoué, B Jost, M Primig, Nathalie Dejucq-Rainsford, et al.

RNA profiling of human testicular cells identifies syntenic lncRNAs associated with spermatogenesis

Running title: Syntenic lncRNAs expressed during spermatogenesis

AD. Rolland[1], B. Evrard[1], TA. Darde[1,2], C. Le Béguec[1], Y. Le Bras[2], K. Bensalah[3], S. Lavoué[4], B. Jost[5], M. Primig[1], N. Dejucq-Rainsford[1], F. Chalmel[1,†,¤],B. Jégou[1,†]

[1] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S1085, F-35000 Rennes, France

[2] Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

[3] Urology department, University of Rennes, Rennes, France

[4] Unité de coordination hospitalière des prélèvements d'organes et de tissus, Centre Hospitalier Universitaire de Rennes, Rennes, France

[5] Plateforme GenomEast - Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), INSERM U964, CNRS UMR 7104, Université de Strasbourg, 67404 Illkirch, France

† The authors consider that the last two authors should be regarded as joint Last Authors

¤ To whom correspondence should be addressed

Correspondence: frederic.chalmel@inserm.fr

Abstract

STUDY QUESTION: Is the noncoding transcriptional landscape during spermatogenesis conserved between human and rodents?

SUMMARY ANSWER: We identified a core group of 113 long noncoding RNAs (lncRNAs) and 20 novel genes dynamically and syntenically transcribed during spermatogenesis.

WHAT IS KNOWN ALREADY: Spermatogenesis is a complex differentiation process driven by a tightly regulated and highly specific gene expression program. Recently, several studies in various species have established that a large proportion of known lncRNAs are preferentially expressed during meiosis and spermiogenesis in a testis-specific manner.

STUDY DESIGN, SIZE, DURATION: To further investigate lncRNA expression in human spermatogenesis, we carried out a cross-species RNA profiling study using isolated testicular cells.

PARTICIPANTS/MATERIALS, SETTING, METHODS: Human testes were obtained from post-mortem donors (N=8, 51 years old on average) or from prostate cancer patients with no hormonal treatment (N=9, 80 years old on average) and only patients with full spermatogenesis were used to prepare enriched populations of spermatocytes, spermatids, Leydig cells, peritubular cells and Sertoli cells. To minimize potential biases linked to inter-patient variations, RNAs from two or three donors were pooled prior to RNA-sequencing (paired-end, strand-specific). Resulting reads were mapped to the human genome, allowing for assembly and quantification of corresponding transcripts.

MAIN RESULTS AND THE ROLE OF CHANCE: Our RNA-sequencing analysis of pools of isolated human testicular cells enabled us to reconstruct over 25,000 transcripts. Among them we identified thousands of lncRNAs, as well as many previously unidentified genes (novel unannotated transcripts) that share many properties of lncRNAs. Of note is that although

noncoding genes showed much lower synteny than protein-coding ones, a significant fraction of syntenic lncRNAs displayed conserved expression during spermatogenesis.

LARGE SCALE DATA:

Raw data files (fastq) and a searchable table (.xlss) containing information on genomic features and expression data for all refined transcripts have been submitted to the NCBI GEO under accession number GSE74896.

LIMITATIONS, REASONS FOR CAUTION: Isolation procedures may alter the physiological state of testicular cells, especially for somatic cells, leading to substantial changes at the transcriptome level. We therefore cross-validated our findings with three previously published transcriptomic analyses of human spermatogenesis. Despite the use of stringent filtration criteria, i.e. expression cut-off of at least three fragments per kilobase of exon model per million reads mapped, fold-change of at least three and false discovery rate adjusted p-values of less than < 1%, the possibility of assembly artefacts and false-positive transcripts cannot be fully ruled out.

WIDER IMPLICATIONS OF THE FINDINGS: For the first time, this study has led to the identification of a large number of conserved germline-associated lncRNAs that are potentially important for spermatogenesis and sexual reproduction. In addition to further substantiating the basis of the human testicular physiology, our study provides new candidate genes for male infertility of genetic origin. This is likely to be relevant for identifying interesting diagnostic and prognostic biomarkers and also potential novel therapeutic targets for male contraception.

STUDY FUNDING/COMPETING INTEREST(S): This work was 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 to B.J. [N 10028NN]; Rennes

68 Métropole “Défis scientifiques émergents” to F.C (2011) and A.D.R (2013). The authors have
69 no competing financial interests.

70

71 **Keywords:** Human spermatogenesis; RNA profiling; novel unannotated transcripts; lncRNAs;
72 expression conservation, synteny

73

Introduction

Over the last two decades genome-wide association studies (GWAS) have found very few significant hits that could explain male infertility (Tüttelmann *et al.*, 2007; Krausz *et al.*, 2015). Indeed causal genetic diagnoses in infertile males can be established in less than 30% of men in infertile couples and the etiology of altered spermatogenesis thus remains largely unclear (Tüttelmann *et al.*, 2018). In this context the noncoding counterpart of the genome is usually ignored in GWAS whereas accumulated evidence emphasizes that genetic variants in these regions can be a cause for missing heritability (Zhang and Lupski, 2015).

Because of the highly dynamic and complex expression program underlying spermatogenesis, which includes a great number of genes expressed in a testis-specific manner, this process has been the focus of numerous genomics studies (for review, see (Chalmel and Rolland, 2015)). Several RNA sequencing for expression quantification (RNA-seq) analyses in rodents and human have thus contributed to the identification of thousands of long noncoding RNAs (lncRNAs) expressed in the testis (Cabili *et al.*, 2011; Laiho *et al.*, 2013; Soumillon *et al.*, 2013; Chalmel *et al.*, 2014; Chocu *et al.*, 2014). However, such approaches have not been applied frequently in male infertility (Tüttelmann *et al.*, 2018) especially when considering the noncoding counterpart of the genome for genetic screening.

By definition, lncRNAs are a large class of noncoding RNAs more than 200 nucleotides in length. They are broadly classified as intergenic, intronic, or overlapping in the sense or antisense orientation according to their position relative to known protein-coding genes (Derrien *et al.*, 2012). Most lncRNAs are RNA polymerase II-transcribed and thus likely to be capped, polyadenylated, and spliced, like mRNAs (Cabili *et al.*, 2011; Guttman and Rinn, 2012). Unlike mRNAs, however, lncRNAs exhibit unique cellular localization patterns highly correlated with the functions they perform in the cell. Known roles include the regulation of DNA methylation, histone modification, chromatin remodeling, and control of gene expression,

either in *cis* in the nucleus or in *trans* in the nucleus or the cytoplasm (for review, see Mercer and Mattick, 2013; Chen, 2016; Schmitz *et al.*, 2016). Common genomic features shared by lncRNAs include relatively short lengths, low exon numbers, low GC content, low sequence conservation (comparable to that of introns), low abundance, and high expression specificity. lncRNAs expressed within the testis share additional features, including peak expression at particular phases of germ cell differentiation, specifically during late steps of meiosis and spermiogenesis (Cabili *et al.*, 2011; Laiho *et al.*, 2013; Chalmel *et al.*, 2014). Furthermore, in rats, lncRNAs that are expressed most highly in meiotic spermatocytes have exons twice as long as those of lncRNAs or mRNAs expressed in other cell types or organs (Chalmel *et al.*, 2014). While the relevance of these observations remains unknown, two recent studies assessed the functionality of testicular lncRNAs and revealed their critical roles during spermatogenesis in fly and mouse (Wen *et al.*, 2016; Hosono *et al.*, 2017; Wichman *et al.*, 2017).

In this study, we performed an in-depth characterization of the human testicular transcriptome. Our analysis of enriched adult testicular cells enabled us to reconstruct more than 25,000 "high-confidence" transcripts, including 1,368 lncRNAs and 511 novel unannotated transcribed regions (NUTs), whose low conservation, low protein-encoding potential, and absence of evidence at the protein level strongly suggest that they are novel lncRNAs. Finally, we investigated gene expression correlation between human and rodent spermatogenesis. Our finding that a core group of lncRNAs is transcribed syntenically during germ cell differentiation suggests they might play important roles in this process, thus deepening and broadening the molecular basis for understanding spermatogenesis (Chalmel *et al.*, 2007). Since those conserved lncRNAs also exhibit cell-specific expression, they represent excellent candidates for identifying diagnostic and prognostic biomarkers in azoospermic men and designing novel therapeutic options in male contraception. A graphical display of the

123 corresponding dataset is available *via* the ReproGenomics Viewer (http://rgv.genouest.org)
124 (Darde *et al.*, 2015, 2019).

125

Materials and Methods

Ethical considerations

Human adult testes were collected from multiorgan donors (N=8, 51 years old on average) and from prostate cancer patients (with no hormonal treatment) undergoing orchidectomy (N=9, 80 years old on average), as detailed in Supplementary Table SI. The presence of full spermatogenesis was assessed by transillumination of freshly dissected seminiferous tubules (Nikkanen *et al.*, 1978), observation of spermatozoa and differentiated spermatids following tissue dissociation, as well as ploidy characterization of dissociated cells by cytometry and histology analysis of paraffin-embedded biopsies. All the required authorizations were obtained from the "Agence de la Biomédecine" (PFS09-015) (for deceased donors) and the "Comité de Protection des Personnes" (#02/31-407) (for cancer patients).

Prior to sequencing, RNAs from distinct donors were eventually pooled in an equimolar manner (Supplementary Table SI).

Sample isolation

Leydig cells. Human Leydig cells were isolated as previously described (Simpson *et al.*, 1987) with minor modifications (Willey *et al.*, 2003). Briefly, interstitial cells were recovered after collagenase digestion of the testicular parenchyma (0.5 mg/mL, 32°C, 45 minutes) and loaded onto a discontinuous density percoll gradient. After centrifugation at 600 g for 30 minutes, the cells were carefully recovered, washed with PBS, and allowed to plate for 2 days at 32°C in DMEM-F12 containing 10% fetal calf serum (FCS), fungizone (2.5 µg/mL), penicillin (50 UI/mL), streptomycin (50 µg/mL), vitamin C (0.1 mM), vitamin E (10 µg/Ml), insulin (10 µg/mL), transferrin (10 µg/mL) and hCG (100 mIU/mL). The cells were then washed with PBS to remove nonadhering germ cells and were cultured for 3 more days in the same medium

without FCS. Cells were then harvested with a cell scraper, pelleted down, snap-frozen, and stored at -80°C until RNA extraction.

Peritubular cells. Human peritubular cells were isolated and cultured as described elsewhere (Albrecht *et al.*, 2006) with minor modifications. First, the testicular parenchyma underwent collagenase digestion (0.5 mg/mL, 32°C, 45 minutes). Individualized seminiferous tubule fragments (1-2 cm long) were selected, placed onto the surface of a plastic cell culture dish, and covered with 20-30 µL FCS. The FCS drops were allowed to start evaporating for 30 minutes, and then DMEM-F12 containing 10% FCS, fungizone (2.5 µg/mL), penicillin (50 UI/mL), and streptomycin (50 µg/mL) was slowly added to the culture dish. After 2 to 3 weeks of culture at 32°C, when peritubular cells had started to grow out of the seminiferous tubules, fragments were removed and the cells were allowed to grow for 2 more weeks. Cells were then harvested with a cell scraper, pelleted down, snap-frozen, and stored at -80°C until RNA extraction.

Sertoli cells. Isolated human primary Sertoli cells (Chui *et al.*, 2011) were purchased from Lonza (Walkersville, MD, USA) and cultured for 3 weeks at 32°C in Sertoli Cell Basal Medium (SeBM™, Lonza) containing FBS, fungizone, and gentamicin (SeGM™, Lonza). Cells were then harvested with a cell scraper, pelleted down, snap-frozen, and stored at -80°C until RNA extraction.

Germ cells. Pachytene spermatocytes and round spermatids were isolated according to previously described procedures (Guillaudeux *et al.*, 1996). Briefly, after collagenase digestion of testicular parenchyma (0.5 mg/mL, 32°C, 45 minutes), the seminiferous tubules were further

dilacerated with scalpels and finally redigested by trypsin (0.3 mg/mL, 32°C, 30 minutes). Cell suspensions were next washed with PBS, filtered on sheets of nylon gauze (300-, 100-, and 20-µm pore size) and on nylon wool before separation by an elutriation rotor (JE5 Beckman Instruments, Inc., Fullerton, CA, USA). Cells were then pelleted down, snap-frozen, and stored at -80°C until RNA extraction.

RNA extraction, library construction, and RNA-Seq

Total RNA was extracted from tissues and cell pellets with the RNeasy mini Kit (Qiagen, Hilden, Germany), quantified with a NanoDrop™ 8000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), and quality controlled with a 2100 Electrophoresis Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Isolated cells and tissue samples from a total of 16 patients were investigated in duplicate and, in some cases, were pooled according to the availability of the material. Libraries of template molecules suitable for strand-specific high-throughput DNA sequencing were created by using "TruSeq Stranded Total RNA with Ribo-Zero Gold Prep Kit" (catalog # RS-122-2301; Illumina Inc., San Diego, CA, USA). Briefly, cytoplasmic and mitochondrial rRNA were removed from 500 ng of total RNA with biotinylated, target-specific oligos combined with Ribo-Zero rRNA removal beads. After purification, the RNA was fragmented by using divalent cations at a high temperature. The cleaved RNA fragments were copied into first-strand cDNA with reverse transcriptase and random primers; second-strand cDNA synthesis followed, with DNA Polymerase I and RNase H. The double-stranded cDNA fragments were blunted with T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase. A single 'A' nucleotide was added to the 3' ends of the blunt DNA fragments by using a Klenow fragment (3' to 5'exo minus) enzyme. The cDNA fragments were ligated to double-stranded adapters with T4 DNA Ligase. The ligated products were enriched by PCR amplification (30 sec at 98°C; [10 sec at 98°C, 30 sec at 60°C,

30 sec at 72°C] × 12 cycles; 5 min at 72°C). Excess PCR primers were removed by purification with AMPure XP beads (Beckman Coulter, Brea, CA, USA). Final cDNA libraries were quality-checked and quantified with a 2100 Electrophoresis Bioanalyzer (Agilent Technologies). The libraries were loaded in the flow cell at 7 pM concentration, and clusters were generated in the Cbot and sequenced in the Illumina Hiseq 2500 as paired-end 2×50 base reads following Illumina's instructions. Image analysis and base calling were performed with RTA 1.17.20 and CASAVA 1.8.2. Raw data files (fastq) and a searchable table (.xlss) containing information on genomic features and expression data for all 25,161 refined transcripts have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession number GSE74896. All data are also conveniently accessible through the ReproGenomics Viewer (Darde *et al.*, 2015, 2019).

RNA-seq analysis

Assembly of a unique set of human reference transcripts. Ensembl (Yates *et al.*, 2016) and RefSeq (Pruitt *et al.*, 2014; Brown *et al.*, 2015) transcript annotations of the hg19 release of the human genome were downloaded from the University of California Santa Cruz (UCSC) genome browser website (Rosenbloom *et al.*, 2014) on November 4, 2014. Both transcript annotation files (GTF format) were subsequently merged into a combined set of nonredundant human reference transcripts (HRT), with Cuffcompare (Trapnell *et al.*, 2012). We also defined a nonredundant dataset of human splice junctions (HSJ) extracted from alignments of human transcripts and expressed sequence tags (ESTs) provided by UCSC.

Mapping reads. RNA-seq-derived reads from each sample duplicate were aligned independently with the hg19 release of the human genome sequence by TopHat (version 2.0.10)

(Trapnell *et al.*, 2009) using previously published approaches (Pauli *et al.*, 2012; Trapnell *et al.*, 2012; Chalmel *et al.*, 2014; Zimmermann *et al.*, 2015). Briefly, the TopHat program was run a first time for each RNA-seq fastq file with the HRT and HSJ datasets to improve read mapping. The resulting junction outputs produced by all TopHat runs were pooled and added to the HSJ dataset. TopHat was rerun a second time for each sample with the new HSJ dataset. The output of this second run comprised the final alignment (BAM format). Finally, BAM files corresponding to sample duplicates were subsequently merged and sorted with the samtools suite (Li *et al.*, 2009).

To compare RNA-seq data from different testicular cell populations across species appropriately, previously published rat and mouse RNA-seq datasets (Gan *et al.*, 2013; Soumillon *et al.*, 2013; Chalmel *et al.*, 2014) were reanalyzed with the same mapping protocol on the rn6 and mm9 releases of the rat and mouse genomes, respectively.

Transcriptome assembly and quantification. The transcriptome of each human testicular cell population was subsequently assembled, compared to known transcript annotations and quantified with the Cufflinks suite (version 2.2.1), with default settings applied (Trapnell *et al.*, 2012). Briefly, the assembly step performed by Cufflinks using the merged alignment files yielded a set of ~51,000-380,000 transcript fragments (transfrags) for each testicular cell type (Supplementary Table SII). The Cuffcompare program was then used: first to define a nonredundant set of 778,012 assembled transcripts by tracking Cufflinks transfrags from all experimental conditions; and second to compare the resulting transcripts to the HRT dataset (i.e., known transcript annotations). Finally, the abundance of each transcript in each sample was assessed by Cuffdiff (within Cufflinks), expressed as fragments per kilobase of exon model per million reads mapped (FPKM). Abundance values were quantile-normalized to reduce systematic effects and to allow direct comparison between the individual samples.

Refinement of assembled transcripts. As suggested by Prensner *et al.* (2011) and Chalmel *et al.* (2014), we sequentially applied four filtering steps to discriminate the most robust transfrags from background noise (Supplementary Fig. S1). First, we selected 37,310 "detectable" or "expressed" transfrags, defined as those for which abundance levels exceeded 3 FPKM in at least one experimental condition (average value of sample duplicates). We next selected 33,562 transcripts with a cumulative exon length ≥200 nt. Third, all transfrags that were not automatically annotated by Cuffcompare as a complete match (Cuffcompare class "="), potentially novel isoform ("j"), unknown intronic ("i", i.e. loci falling entirely within a reference intron and without exon-exon overlap with another known locus), intergenic ("u"), or antisense ("x") isoforms were discarded, thereby leaving 28,253 transfrags for analysis. Finally, all transcript fragments that were annotated as either novel isoforms or novel genes (class codes "j", "i", "u" or "x") and that did not have at least two exons (multiexon) were filtered out. Altogether, this strategy produced a high-confidence set of 25,161 transcripts meeting these refinement conditions and supporting total RNA molecules expressed in human testicular cells during spermatogenesis (Supplementary Fig. S1).

Microarray data normalization

To monitor the expression level of the assembled mRNAs and lncRNAs in infertile men with non-obstructive azoospermia we integrated microarray data published by Malcher and colleagues (Malcher *et al.*, 2013). Raw data were downloaded from the NCBI GEO (Barrett *et al.*, 2012) repository under the accession number GSE45885. The gene array data were normalized with the Robust Multi-Array Average method (Irizarry *et al.*, 2003) implemented in the statistical software R (version 3.5.1) using the Brainarray custom Chip Description Files

(CDF, version 23.0.0) so that intensity values are not summarized for each probe set but directly for each Entrez Gene ID (Dai *et al.*, 2005).

Analysis of coding potential of the assembled transcripts

Before analyzing the protein-coding potential of the 25,161 high-confidence transcripts assembled in human testicular cell populations, we used TopHat's gffread tool to extract their DNA sequences. As already described (Chalmel *et al.*, 2014; Zimmermann *et al.*, 2015), the resulting nucleic sequences were classified as either coding or noncoding, according to an empirical integrative approach using four distinct predictive tools: Coding-Potential Assessment Tool (CPAT), HMMER, Coding Potential Calculator (CPC) and txCdsPredict (Kong *et al.*, 2007; Finn *et al.*, 2011; Kuhn *et al.*, 2013; Wang *et al.*, 2013a). Transcripts were considered protein-coding candidates if they had a coding probability >0364 in CPAT, 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 >800 (~90% predictive of protein-coding genes). Finally, transcripts were organized into five groups considered to have "Very High" (4/4 tools predict protein-coding potential), "High" (3/4), "Medium" (2/4), "Low" (1/4), or "No" (0/4) protein-coding potential according to whether their nucleic sequences were considered as protein-coding by four, three, two, one or none of the four predictive tools, respectively.

Proteomics Informed by Transcriptomics strategy

To provide evidence at the protein level for assembled transcripts, we applied a Proteomics Informed by Transcriptomics (PIT) approach, as recently described by Evans and co-workers (Evans *et al.*, 2012). This approach relies on the query of tandem mass spectrometry (MS/MS)

proteomics spectra against a customized protein database derived from RNA-seq data of the same or similar samples.

Assembly of a unique human reference proteome database. The nucleic sequences of the 25,161 high-confidence assembled transcript isoforms were translated into the three-first open reading frames with the EMBOSS's Transeq program (Rice *et al.,* 2000). Deduced amino acid sequences of at least 20 residues between two stop codons were defined as potential protein sequences. Finally, a nonredundant human reference proteome database was assembled by merging the UniProt (89,033 canonical and isoform sequences; release 2014_08) (Pundir *et al.*, 2015) and Ensembl (99,459 known and 50,117 predicted protein sequences; release-76) (Yates *et al.*, 2016) proteome databases with the set of predicted protein sequences.

Protein identification. For this experiment, we used a human adult testis MS/MS proteomics dataset available from the Human Proteome Map (Kim *et al.*, 2014). All analyses were performed with PeptideShaker (release 0.31.4) (Vaudel *et al.*, 2011), implemented in the Galaxy web-based genome analysis environment (Blankenberg *et al.*, 2010) (release 1.19.5.0), based on SearchGUI (release 1.19.5) (Vaudel *et al.*, 2011). First, the 46 raw data files (.raw) were downloaded from the PRIDE database under accession number PXD000561 and converted to mgf format with PeptideShaker. A concatenated target/decoy database was created by reversing the sequences from a target database with SearchGUI. Cross-linked peptide identification was thus carried out with X!Tandem, Open Mass Spectrometry Search Algorithm (OMSSA) and MS-GF+ (Vaudel *et al.*, 2011). We applied the parameters used by Pinto and coworkers (Pinto *et al.*, 2014): precursor ion tolerance units set at 10 ppm; fragment tolerance set at 0.05 Da; carbamidomethylation of cysteine defined as a fixed modification; oxidation of methionine defined as a variable modification; and only tryptic peptides with up to two missed cleavages were considered. All peptides with at least one validated peptide-

spectrum match (PSM) and a confidence interval greater than 80% were kept for further analyses. Finally, only identifications with a false discovery rate (FDR) < 1% indicated by the PeptideShaker validation method were considered.

Statistical filtration and clustering analysis

Statistical analysis. The statistical filtration of the transfrags showing a differential expression (DE) across experimental samples was performed using AMEN (Chalmel and Primig, 2008) (Supplementary Fig. S3). First, we performed every pairwise comparison between experimental conditions and selected 23,687 transcript fragments yielding at least one fold-change greater or equal to 3.0 (average values of sample duplicates). A Linear Models for Microarray Data (LIMMA) statistical test (was finally used to identify 21,264 transcripts with significant abundance variations across samples (F-value adjusted with the FDR method: $P \leq 0.05$) (Smyth, 2004).

Cluster analysis. The 21,264 DE transfrags were next clustered into 11 expression patterns (P1-P11) by the k-means algorithm (Supplementary Fig. S3). The quality of the resulting k-means clusters was verified with Silhouette plots. The 11 resulting patterns were ordered according to peak expression levels in the different cell types. The 3,897 transfrags for which no significant differential expression was observed (fold-change <3 or $P > 0.05$) were placed in a group term P0.

Testis specificity analysis. To filter transcripts expressed testis-specifically, we downloaded a tissue profiling dataset from the NCBI GEO under the accession number GSE45326 (Nielsen *et al.*, 2014). This experiment comprises 12 normal tissues including ovary, bladder, brain,

breast, colon, heart, kidney, liver, lung, muscle, prostate, and skin. To allow proper comparison with our own data, this dataset was re-analyzed with TopHat as described above. Next, the abundance of each transcript assembled in our dataset was assessed with Cuffquant and normalized with Cuffnorm (Trapnell *et al.*, 2012).

An empirical filtration approach based on abundance was applied to select transcripts reliably detected only in human testis. Assembled transcripts were considered testis-specific if their abundance was > 1 FPKM in the testis and 1 < FPKM in the 12 other tissues. To avoid selection of candidates with values close to the threshold, only those with an abundance at least three-fold higher in testis than in other tissues were retained.

Quantification of syntenic transcripts in rodents

Conversion of genome co-ordinates from humans to rodents. Since potential orthologous loci in rodents of most human lncRNAs and novel loci identified in our study are probably unknown, co-ordinates of the 25,161 high-confidence assembled transcripts (GenePred format) were mapped to syntenic mouse (mm9) and rat (rn6) regions with UCSC's liftOver tool (parameters: −minMatch=0.1 −minBlocks=0.5, as recommended by UCSC for cross-species conversion).

Syntenic transcript quantification. The abundance of each syntenic transcript in each rodent RNA-seq dataset (Gan *et al.*, 2013; Soumillon *et al.*, 2013; Chalmel *et al.*, 2014) was assessed with the Cufflinks suite (Pollier *et al.*, 2013). Abundance values (FPKM) were quantile-normalized to reduce systematic effects and to allow direct comparison between the individual samples.

Identification of conserved and correlated loci. *Rattus norvegicus* was selected as the reference rodent species since the human and rat datasets were produced with similar library preparation protocols (including an rRNA-depletion method) and comprise five similar testicular cell types (Leydig, peritubular, and Sertoli cells, spermatocytes, and round spermatids). A syntenic transcript was considered as "detectable" when its abundance value ≥ 1 FPKM in the rat. Finally, syntenic and detectable transcripts showing similar patterns in the rat (correlation coefficient ≥0.8 between the five common testicular cell types) defined the set of loci with a conserved and correlated expression profile between humans and rodents.

Statistical tests

Enrichment calculation. AMEN (Chalmel and Primig, 2008) was used to calculate the Fisher exact probability, and the Gaussian hypergeometric test to identify significantly enriched terms from the gene ontology in the 11 expression patterns (P1-P11). A specific annotation term was considered enriched in a group of coexpressed genes if the *P* value was <0.001 and the number of genes in this cluster showing this annotation was >3.

Quantitative PCR experiments

Total RNA (2.5 µg) was first submitted to DNAse treatment by the TURBO DNA-free™ Kit (Thermo Fisher Scientific). cDNA synthesis was performed on 800 ng of DNAse-treated RNA, with the iScript™ Reverse Transcription Supermix for RT-qPCR (Bio-Rad Laboratories, Hercules, CA, USA), according to the manufacturer's instructions. Quantitative PCR (qPCR) experiments were next performed on cDNA from isolated human Leydig cells (n=5), peritubular cells (n=2), Sertoli cells (n=2), pachytene spermatocytes (n=4), and round spermatids (n=4). Each gene was assessed in each sample in technical duplicates on 2 ng of

cDNA, by using iTaq™ Universal SYBR® Green Supermix (Bio-Rad Laboratories) and CFX384 Touch™ Real-Time PCR Detection System (Bio-Rad Laboratories), according to the default program (95°C for 3 min followed by 40 cycles of 95°C for 10 s, 55°C for 30 s) including a melting curve step (65°C to 95°C with a 0.5°C increment and a hold time of 5 s before reading plate). Results were analyzed with the Bio-Rad CFX Manager. Previously published primers for GAPDH and RPLP0 were used for normalization purposes (Svingen *et al.*, 2014). Specific primers for candidate transcripts were designed with the Universal ProbeLibrary System Assay Design (https://lifescience.roche.com/shop/products/universal-probelibrary-system-assay-design). When intron-spanning primers could not be found, the Primer3 software (v4.0.0; http://primer3.ut.ee/) was used. Only primer pairs showing both a single peak in the melting curve analysis and an amplification efficiency between 95 and 105% were used in subsequent experiments (Supplementary Table SIII).

Results

Transcript assembly in human adult testicular cells identifies almost 500 new genes

We performed here an RNA-seq analysis of human adult testis, of meiotic and postmeiotic germ cells, as well as of Leydig, peritubular, and Sertoli cells, all originating from donors with apparent normal spermatogenesis (Supplementary Table SI). To cover their transcriptomes as broadly as possible, we sequenced ribosomal RNA-depleted total RNA rather than polyA RNAs that tend to be biased towards mRNA (Guttman *et al.*, 2013). We further reconstructed both known and unknown transcript isoforms according to an approach previously described (Chalmel *et al.*, 2014; Zimmermann *et al.*, 2015). In total, 77.3% of the reads that mapped to the human genome were further used to assemble, quantify, and refine a set of 25,161 “high-confidence” transcripts corresponding to 10,703 loci (Supplementary Table SII, and

Supplementary Fig. S1). Comparison to the human genome annotation (Ensembl and RefSeq combined, 204,222 nonredundant transcripts) showed that most of them correspond to known (8,777 transcripts, 34.9%) and novel (13,882, 55.2%) isoforms of annotated protein-coding loci, as well as to known (562, 2.2%) and novel (806, 3.2%) isoforms of annotated lncRNAs (Supplementary Fig. S2). Importantly, 511 transcripts (452 loci) appeared to be novel intronic (148, 0.59%), intergenic (193, 0.77%), and antisense (170, 0.68%) unannotated loci, referred to as NUTs (Fig. 1). An integrative approach combining a coding-potential prediction analysis and a PIT strategy (Evans *et al.*, 2012) further suggested that NUTs actually correspond to as-yet unidentified human lncRNAs (Supplementary Fig. S2). Use of isolated cells was critical for identifying these new loci, for which 83.6% would not have met the expression threshold criteria in total testis samples. Additionally, 623 remaining transcripts (2.5%) corresponded to other RNA types, such as pseudogenes and microRNAs, which were not further analyzed.

The dynamics of the testicular transcriptional landscape highlight the accumulation of lncRNAs during human spermiogenesis

To identify the transcripts preferentially expressed in each testicular cell population, we next performed a differential expression analysis. We found that 21,264 transcripts (84.5% of those assembled) showed significant differential expression, which we classified into 11 patterns (P1-P11) (Supplementary Fig. S3, and Fig. 2A): P1, P2, P3, and P4 comprise transcripts highly expressed, respectively, in all somatic cells, in Leydig cells, in peritubular cells and in Sertoli cells. P5 contains transcripts expressed in every cell type but spermatids. Transcripts in patterns P6 to P10 show gradual peak expression in differentiating germ cells from spermatocytes to spermatids. Finally, P11 is composed of transcripts with peak expression in the total testis. Importantly, the consistency of expression patterns was confirmed by different analyses (Supplementary Fig. S4-S10). These showed: the overall appropriate expression of known

testicular cell markers (Supplementary Fig. S4); a highly significant overlap with data from microarray analyses of testicular biopsies from infertile patients (Chalmel *et al.*, 2012; Malcher *et al.*, 2013) (Supplementary Fig. S5 and Supplementary Fig. S6) or from single-cell RNA-seq analysis of human adult spermatogenesis (Wang *et al.*, 2018) (Supplementary Fig. S7); the overrepresentation of relevant biological processes for each cluster (Supplementary Fig. S8); and the underrepresentation of X chromosome-derived genes in germ cell patterns P5 to P10 as a reflection of meiotic sex chromosome inactivation (Turner, 2007) (Supplementary Fig. S9).

Next, we focused on known lncRNAs, and found that most (64.7%) were preferentially expressed in spermatids (P9 and P10, $P < 10^{-177}$) (Fig. 2B). Similarly, the finding that most NUTs (69.6%) were also preferentially expressed during spermiogenesis (P9 and P10, $P < 10^{-81}$) (Fig. 2B) supports the idea that many of the NUTs are *bona fide* lncRNAs. The preferential expression of eight NUTs in spermatocytes and/or spermatids (Supplementary Fig. S10) was further validated with qPCR (Supplementary Table SIII).

Expansion of the testis-specific repertoire of lncRNAs

Tissue-specific expression is often considered an indication that the genes involved play unique and important functions in a narrow range of biological processes. To ascertain in more detail the fraction of the testicular transcriptional landscape expressed only in this organ, we analyzed a tissue-profiling dataset including 12 types of normal human tissues (Nielsen *et al.*, 2014) and found that 16.5% (3515 / 21264) of the differentially expressed transcripts were detected only in testis. As expected, 91.4% were predominantly expressed in the germline, showing peak transcriptional induction in early spermatids (P9-P10, 2162/3515, 61.5%) and, to a lesser extent, in spermatocytes (P5-P8, 1051/3515, 29.9%). These RNA-seq data thus confirm and extend

earlier observations by showing that twice as many testis-specific transcripts are expressed during the haploid phase as during meiosis (Chalmel *et al.*, 2012). Consistent with their preponderant expression in meiotic and postmeiotic germ cells, 71% of the newly identified NUTs were also detected specifically in the testis. As already observed in the rat (Chalmel *et al.*, 2014), exons of a specific subset of lncRNAs expressed in meiotic and postmeiotic germ cells are longer than those of other coding and noncoding transcripts (Supplementary Fig. S11).

Conserved testicular expression of syntenic mammalian lncRNAs

The core testicular transcriptome in mammals includes 18,847 “high-confidence” transcripts (~88.6%) located in genomic regions hypothesized to be homologous due to their shared synteny between humans and rodents (mice and rats) (Table I). The great majority of these syntenic transcripts (15,119 transcripts, 80.2%) were expressed in at least one type of testicular cell in rats. Furthermore, approximately half of the syntenic transcripts (8,457 transcripts, 44.9%) displayed very similar expression patterns (correlation > 0.5) in human and rat, and therefore constitute a core set of evolutionary conserved loci involved in spermatogenesis (Fig. 3). The present RNA-seq study broadens our insight into the conserved testicular expression program, since our previously published microarray analysis identified only 12.6% of coding genes in humans and rodents as being correlated (Chalmel *et al.*, 2007). Importantly, we were able to determine that the noncoding component of the core testicular transcriptome includes at least 113 lncRNAs and 20 NUTs (Table I). The expression pattern of syntenic mRNAs are far more often correlated than those of lncRNAs (45.8% versus 20.2%).

Discussion

Gaining insight into the testicular transcriptional landscape is essential for a better understanding of genetic causes underlying infertility in men. It may also facilitate unraveling what is behind negative trends in several components of male reproductive health, including testicular cancer; and the genomic mechanisms involved in the genetic introgression that occurred during ancestral human admixtures and in which meiosis appears central (Jégou *et al.*, 2017). The testis is undeniably the organ that expresses the highest number of genes in a tissue-specific manner, due primarily to the complex processes and associated factors that are required for male germ cell development (for reviews, see (Kleene, 2001; Eddy, 2002; Kimmins *et al.*, 2004)). Some researchers consider, however, that the specificities of the spermatogenic cell expression program may also result from promiscuous or leaky transcription during and after meiosis, which would lead to the adventitious synthesis of nonfunctional transcripts (for review, see (Ivell, 1992)). Interestingly, the atypical patterns of gene expression in the testis have also been hypothesized to result from sexual selection, a distinct form of natural selection based on mate choice and competition for mating that acts at the level of the organism, cell, and molecule (for review, see (Kleene, 2005)). Sexual selection is notably responsible for the rapid evolution of often exaggerated reproductive traits in the competing sex, i.e., males in most species (Darwin, 1871; Hosken and House, 2011). This intense selection pressure to which male reproduction is subjected also triggers the rapid divergence of testicular genes, at both the expression and sequence levels (for review, see (Grath and Parsch, 2016)). Finally, together with potential leaky transcription in meiotic and postmeiotic germ cells, sexual selection is also responsible for making the testis a fertile ground for the birth of new genes, coding as well as noncoding, either from scratch or from pre-existing genes ((Xie *et al.*, 2012; Ruiz-Orera *et al.*, 2015); for review, see (Kaessmann, 2010)). In this context, lncRNAs constitute a class of genes with a high origination rate and rapid turnover; most of them are consequently species- or lineage-specific, and their expression is highly enriched in the testis (Cabili *et al.*, 2011; Derrien

512 *et al.*, 2012; Necsulea *et al.*, 2014; Hezroni *et al.*, 2015); for review (Kapusta and Feschotte,
513 2014)).

514 Several RNA-seq analyses have investigated the expression of lncRNAs during
515 spermatogenesis in various animal species or used the testis as a model organ to identify new
516 lncRNAs (Soumillon *et al.*, 2013; Necsulea *et al.*, 2014). Recently, two RNA-seq studies were
517 performed using human male germ cells (Zhu *et al.*, 2016; Jan *et al.*, 2017). In those works,
518 the expression of both mRNAs and lncRNAs was assessed in human spermatogonia,
519 spermatocytes, and spermatids. However, since the analysis pipelines did not include transcript
520 assembly, both studies failed to identify novel genes.

521 The study reported here is original because it aimed at identifying new transcript
522 isoforms and unknown genes expressed during spermatogenesis in human testicular cells. Thus,
523 the stringent quality criteria allowed us to assemble and quantify over 25,000 high-confidence
524 transcripts. These included 11,627 and 766 new transcript isoforms for coding and for
525 noncoding genes, respectively, as well as 511 completely novel unannotated multi-exon
526 transcripts. Of particular interest is that more than 85% of these unknown genes showed
527 preferential expression in spermatocytes and spermatids; this finding clearly illustrates that
528 adult male germ cells constitute an important reservoir for gene discovery purposes (Chalmel
529 *et al.*, 2014; Chocu *et al.*, 2014). It is also noteworthy that fewer than 15% of these transcripts
530 would have met the expression cutoff we used if our analysis had included total testis only:
531 most (~70%) were indeed identified because of their high expression in spermatids and, to a
532 lower extent (~16%), in spermatocytes. This demonstrates the striking advantage of using
533 isolated cells, in terms of sensitivity.

534 The 511 NUTs we identified actually correspond to 451 new human genes. Their virtual
535 absence of protein-encoding potential strongly suggests that most, if not all, encode new
536 lncRNAs. In agreement with previous reports in different species (Laiho *et al.*, 2013;

Soumillon *et al.*, 2013; Chalmel *et al.*, 2014), most lncRNAs showed preferential expression in human postmeiotic spermatids. Furthermore, a subset of lncRNAs that are expressed during meiosis have exons twice as long as other lncRNAs or mRNAs. These longer exons, of unknown functional relevance, appear to be a conserved phenomenon, since the same observation has been made for meiotic lncRNAs in rats (Chalmel *et al.*, 2014). As suggested previously (Naro *et al.*, 2017), one possible explanation could be that lncRNAs critical for spermatogenesis are stabilized for several days after their synthesis thanks to an intron retention program in meiotic spermatocytes.

Although many essential functions of lncRNAs in various biological processes have been demonstrated, the biological relevance of the massive expression of such RNAs during the metamorphosis of haploid spermatids into mature spermatozoa remains unknown. We cannot rule out the possibility that some of them are junk products of leaky transcription during meiosis. Nonetheless, their promoters' high degree of conservation — at least as conserved as protein-coding gene promoters — suggests strong selective constraints at the transcriptional level and important functions for these molecules (Carninci *et al.*, 2005; Necsulea *et al.*, 2014). Possible roles for certain spermatozoal lncRNAs during early embryonic development can also be proposed, given that sperm contain a complex population of transcripts that are delivered to the embryo upon fertilization (Ostermeier *et al.*, 2004; Jodar *et al.*, 2013; Sendler *et al.*, 2013). The preferential localization of lncRNAs in the vicinity of protein-coding genes involved in developmental processes points to a potential role for sperm lncRNAs in regulating expression of these genes (Ponjavic *et al.*, 2009; Cabili *et al.*, 2011; Chalmel *et al.*, 2014).

Finally, some lncRNAs may play fundamental roles during germ cell development itself (Wen *et al.*, 2016; Hosono *et al.*, 2017; Wichman *et al.*, 2017), even though predicting functions of lncRNAs on the basis of their sequence remains challenging. Their conservation at both the sequence and expression levels does, however, hint at important roles for some of

them (for review, see (Ulitsky, 2016)). In contrast to mRNAs, and as expected because of their intrinsic low sequence conservation, we identified syntenic regions for only 37% of human testicular lncRNAs. Among these, we defined a core group of 131 lncRNAs with syntenic transcription during spermatogenesis, which suggests they play key roles in germ cell development. The absence of conservation of the rest, at either the sequence or expression level, does not necessarily imply they are not functional: the vast majority of lncRNAs indeed have no homologs in species that diverged more than 50 million years ago (Necsulea *et al.*, 2014; Hezroni *et al.*, 2015). Additional work relying on expression data obtained in other hominins will be needed to clarify the functional contribution of such lineage-specific lncRNAs in human spermatogenesis.

That lncRNAs play important biological functions is now supported by many independent studies reporting their functional implications in almost all the investigated physiological and pathophysiological biological systems (Tao *et al.*, 2016), including spermatogenesis (Wen *et al.*, 2016; Hosono *et al.*, 2017; Wichman *et al.*, 2017). It is therefore important to note that lncRNAs are now presented as a novel class of diagnostic biomarkers and therapeutic targets for several disorders and pathologies (Lavorgna *et al.*, 2016; Arun *et al.*, 2018). On one hand, due to their cell-type specific expression pattern and their diversity of functions, GWAS of patients with non-obstructive azoospermia would obviously benefit from systematically screening for causal genetic variants in evolutionary-conserved lncRNAs expressed in the germline. On the other hand, those molecules could also be used as interesting diagnostic biomarkers, probing the presence or absence of specific testicular cell populations in infertile men with distinct spermatogenic arrests. They might also represent possible biomarkers that could help to determine the prognosis of hormonal therapy with hCG/recombinant FSH for infertile men with idiopathic nonobstructive azoospermia. Last but not least, advances in nucleic acid-based therapies are evolving at a steady rate and have already

shown success in several preclinical studies (Arun *et al.*, 2018). Such promising therapeutic approaches targeting lncRNAs critical for the male germ cell differentiation could pave the way towards exploring novel male contraceptive options that might be clinically relevant in the decade ahead (Lavorgna *et al.*, 2016). These examples are just some of the many clinical applications that could be made in our scientific field. Although many challenges remain to be addressed, especially regarding functional annotations of lncRNAs, the systematic characterization of the noncoding transcriptional landscape at play during the human spermatogenesis process is an indispensable prerequisite to such future directions. In this context, the novel and abundant data provided by the present study substantiates further the general basis without which deciphering the extremely complex mechanisms of normal and failed spermatogenesis in men will remain an utopic challenge.

Acknowledgments

We thank all members of the SEQanswers forums for helpful advice; Steven Salzberg and Cole Trapnell for continuous support with the "Tuxedo" suite; and the UCSC Genome team members. Sequencing was performed by the GenomEast platform, a member of the 'France Génomique' consortium (ANR-10-INBS-0009).

Authors' roles

FC, ADR, and BJ designed the study and wrote the manuscript. FC and ADR supervised the research. FC prepared, analyzed, and interpreted RNA sequencing data. ADR and BJ prepared the testicular samples and interpreted sequencing data. BE prepared the testicular samples and validated expression data. YLB performed the PIT analysis. TAD contributed to the analysis

of the common genomic features shared by the assembled transcripts. CLB contributed to the cross-species data comparison. MP, NDR, BE, and YLB contributed to the manuscript. All authors approved the final version of the manuscript, and declare that they have no competing interests.

Funding

This work was 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 to B.J. [N 10028NN]; Rennes Métropole "Défis scientifiques émergents" to F.C (2011) and A.D.R (2013). The authors have no competing financial interests.

Conflict of interest

There are no competing interests related to this study.

References

Albrecht M, Rämsch R, Köhn FM, Schwarzer JU, Mayerhofer A. Isolation and cultivation of human testicular peritubular cells: a new model for the investigation of fibrotic processes in the human testis and male infertility. *J Clin Endocrinol Metab* [Internet] 2006;**91**:1956–1960.

Arun G, Diermeier SD, Spector DL. Therapeutic Targeting of Long Non-Coding RNAs in Cancer. *Trends Mol Med* [Internet] 2018;**24**:257–277.

632 Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA,
633 Phillippy KH, Sherman PM, Holko M, *et al.* NCBI GEO: archive for functional genomics
634 data sets—update. *Nucleic Acids Res* [Internet] 2012;**41**:D991–D995.
635 Blankenberg D, Kuster G Von, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A,
636 Taylor J. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc*
637 *Mol Biol* [Internet] 2010;**Chapter 19**:Unit 19.10.1-21.
638 Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, Tolstoy I, Tatusova T,
639 Pruitt KD, Maglott DR, *et al.* Gene: a gene-centered information resource at NCBI.
640 *Nucleic Acids Res* [Internet] 2015;**43**:D36–D42.
641 Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative
642 annotation of human large intergenic noncoding RNAs reveals global properties and
643 specific subclasses. *Genes Dev* [Internet] 2011;**25**:1915–1927.
644 Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T,
645 Lenhard B, Wells C, *et al.* The Transcriptional Landscape of the Mammalian Genome.
646 *Science (80-)* [Internet] 2005;**309**:1559–1563.
647 Chalmel F, Lardenois A, Evrard B, Mathieu R, Feig C, Demougin P, Gattiker A, Schulze W,
648 Jégou B, Kirchhoff C, *et al.* Global human tissue profiling and protein network analysis
649 reveals distinct levels of transcriptional germline-specificity and identifies target genes
650 for male infertility. *Hum Reprod* [Internet] 2012;**27**:3233–3248.
651 Chalmel F, Lardenois A, Evrard B, Rolland AD, Sallou O, Dumargne M-C, Coiffec I, Collin
652 O, Primig M, Jégou B. High-resolution profiling of novel transcribed regions during rat
653 spermatogenesis. *Biol Reprod* [Internet] 2014;**91**:5.
654 Chalmel F, Primig M. The Annotation, Mapping, Expression and Network (AMEN) suite of
655 tools for molecular systems biology. *BMC Bioinformatics* [Internet] 2008;**9**:86.
656 Chalmel F, Rolland AD. Linking transcriptomics and proteomics in spermatogenesis.

Reproduction [Internet] 2015;**150**:R149-57.

Chalmel F, Rolland ADAD, Niederhauser-Wiederkehr C, Chung SSWSSW, Demougin P, Gattiker A, Moore J, Patard JJ-J, Wolgemuth DJDJ, Jégou B, *et al.* The conserved transcriptome in human and rodent male gametogenesis. *Proc Natl Acad Sci U S A* [Internet] 2007;**104**:8346–8351. National Academy of Sciences.

Chen L-L. Linking Long Noncoding RNA Localization and Function. *Trends Biochem Sci* [Internet] 2016;**41**:761–772.

Chocu S, Evrard B, Lavigne R, Rolland AD, Aubry F, Jégou B, Chalmel F, Pineau C. Forty-four novel protein-coding loci discovered using a proteomics informed by transcriptomics (PIT) approach in rat male germ cells. *Biol Reprod* [Internet] 2014;**91**:123.

Chui K, Trivedi A, Cheng CY, Cherbavaz DB, Dazin PF, Huynh ALT, Mitchell JB, Rabinovich GA, Noble-Haeusslein LJ, John CM. Characterization and functionality of proliferative human Sertoli cells. *Cell Transplant* [Internet] 2011;**20**:619–635.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* [Internet] 2005;**33**:e175.

Darde TA, Lecluze E, Lardenois A, Stévant I, Alary N, Tüttelmann F, Collin O, Nef S, Jégou B, Rolland AD, *et al.* The ReproGenomics Viewer: a multi-omics and cross-species resource compatible with single-cell studies for the reproductive science community. In Kelso J, editor. *Bioinformatics* [Internet] 2019;Available from: http://www.ncbi.nlm.nih.gov/pubmed/30668675.

Darde TA, Sallou O, Becker E, Evrard B, Monjeaud C, Bras Y Le, Jégou B, Collin O, Rolland AD, Chalmel F. The ReproGenomics Viewer: An integrative cross-species toolbox for the reproductive science community. *Nucleic Acids Res* [Internet] 2015;**43**:W109–W116.

Darwin CR. *The Descent of Man and Selection in Relation to Sex*. In Murray J, editor. 1871;

682 London.
683 Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D,
684 Merkel A, Knowles DG, *et al.* The GENCODE v7 catalog of human long noncoding
685 RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* [Internet]
686 2012;**22**:1775–1789.
687 Eddy EM. Male germ cell gene expression. *Recent Prog Horm Res* [Internet] 2002;**57**:103–
688 128.
689 Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of
690 proteomes from transcriptomes for transcript and protein identification. *Nat Methods*
691 [Internet] 2012;**9**:1207–1211.
692 Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching.
693 *Nucleic Acids Res* [Internet] 2011;**39**:W29-37.
694 Gan H, Wen L, Liao S, Lin X, Ma T, Liu J, Song C-X, Wang M, He C, Han C, *et al.* Dynamics
695 of 5-hydroxymethylcytosine during mouse spermatogenesis. *Nat Commun* [Internet]
696 2013;**4**:1995.
697 Grath S, Parsch J. Sex-Biased Gene Expression. *Annu Rev Genet* [Internet] 2016;**50**:29–44.
698 Guillaudeux T, Gomez E, Onno M, Drénou B, Segretain D, Alberti S, Lejeune H, Fauchet R,
699 Jégou B, Bouteiller P Le. Expression of HLA class I genes in meiotic and post-meiotic
700 human spermatogenic cells. *Biol Reprod* [Internet] 1996;**55**:99–110.
701 Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature*
702 [Internet] 2012;**482**:339–346. NIH Public Access.
703 Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides
704 evidence that large noncoding RNAs do not encode proteins. *Cell* [Internet]
705 2013;**154**:240–251.
706 Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of Long

Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep* [Internet] 2015;**11**:1110–1122.

Hosken DJ, House CM. Sexual selection. *Curr Biol* [Internet] 2011;**21**:R62–R65.

Hosono Y, Niknafs YS, Prensner JR, Iyer MK, Dhanasekaran SM, Mehra R, Pitchiaya S, Tien J, Escara-Wilke J, Poliakov A, *et al.* Oncogenic Role of THOR, a Conserved Cancer/Testis Long Non-coding RNA. *Cell* [Internet] 2017;**171**:1559–1572.e20.

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* [Internet] 2003;**4**:249–264.

Ivell R. 'All that glisters is not gold'--common testis gene transcripts are not always what they seem. *Int J Androl* [Internet] 1992;**15**:85–92.

Jan SZ, Vormer TL, Jongejan A, Röling MD, Silber SJ, Rooij DG de, Hamer G, Repping S, Pelt AMM van. Unraveling transcriptome dynamics in human spermatogenesis. *Development* [Internet] 2017;**144**:3659–3673.

Jégou B, Sankararaman S, Rolland AD, Reich D, Chalmel F. Meiotic Genes Are Enriched in Regions of Reduced Archaic Ancestry. *Mol Biol Evol* [Internet] 2017;**34**:1974–1980.

Jodar M, Selvaraju S, Sendler E, Diamond MP, Krawetz SA, Reproductive Medicine Network. The presence, role and clinical use of spermatozoal RNAs. *Hum Reprod Update* [Internet] 2013;**19**:604–624.

Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res* [Internet] 2010;**20**:1313–1326.

Kaiser GRRF, Monteiro SC, Gelain DP, Souza LF, Perry MLS, Bernard EA. Metabolism of amino acids by cultured rat Sertoli cells. *Metabolism* [Internet] 2005;**54**:515–521.

Kapusta A, Feschotte C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet* [Internet] 2014;**30**:439–452.

732 Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar
733 DS, Isserlin R, Jain S, *et al.* A draft map of the human proteome. *Nature* [Internet]
734 2014;**509**:575–581.
735 Kimmins S, Kotaja N, Davidson I, Sassone-Corsi P. Testis-specific transcription mechanisms
736 promoting male germ-cell differentiation. *Reproduction* [Internet] 2004;**128**:5–12.
737 Kleene KC. A possible meiotic function of the peculiar patterns of gene expression in
738 mammalian spermatogenic cells. *Mech Dev* [Internet] 2001;**106**:3–23.
739 Kleene KC. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene
740 expression in spermatogenic cells. *Dev Biol* [Internet] 2005;**277**:16–26.
741 Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, Gao G. CPC: assess the protein-coding
742 potential of transcripts using sequence features and support vector machine. *Nucleic Acids*
743 *Res* [Internet] 2007;**35**:W345-9.
744 Krausz C, Escamilla AR, Chianese C. Genetics of male infertility: from research to clinic.
745 *REPRODUCTION* [Internet] 2015;**150**:R159–R174.
746 Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. *Brief*
747 *Bioinform* [Internet] 2013;**14**:144–161.
748 Laiho A, Kotaja N, Gyenesei A, Sironen A. Transcriptome profiling of the murine testis during
749 the first wave of spermatogenesis. *PLoS One* [Internet] 2013;**8**:e61558.
750 Lavorgna G, Vago R, Sarmini M, Montorsi F, Salonia A, Bellone M. Long non-coding RNAs
751 as novel therapeutic targets in cancer. *Pharmacol Res* [Internet] 2016;**110**:131–138.
752 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R,
753 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format
754 and SAMtools. *Bioinformatics* [Internet] 2009;**25**:2078–2079.
755 Malcher A, Rozwadowska N, Stokowy T, Kolanowski T, Jedrzejczak P, Zietkowiak W,
756 Kurpisz M. Potential biomarkers of nonobstructive azoospermia identified in microarray

757 gene expression analysis. *Fertil Steril* [Internet] 2013;**100**:1686–1694.e7.
758 Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic
759 regulation. *Nat Struct Mol Biol* [Internet] 2013;**20**:300–307.
760 Naro C, Jolly A, Persio S Di, Bielli P, Setterblad N, Alberdi AJ, Vicini E, Geremia R, la Grange
761 P De, Sette C. An Orchestrated Intron Retention Program in Meiosis Controls Timely
762 Usage of Transcripts during Germ Cell Differentiation. *Dev Cell* [Internet] 2017;**41**:82–
763 93.e4. Elsevier.
764 Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F,
765 Kaessmann H. The evolution of lncRNA repertoires and expression patterns in tetrapods.
766 *Nature* [Internet] 2014;**505**:635–640.
767 Nielsen MM, Tehler D, Vang S, Sudzina F, Hedegaard J, Nordentoft I, Orntoft TF, Lund AH,
768 Pedersen JS. Identification of expressed and conserved human noncoding RNAs. *RNA*
769 [Internet] 2014;**20**:236–251.
770 Nikkanen V, Söderström KO, Parvinen M. Identification of the spermatogenic stages in living
771 seminiferous tubules of man. *J Reprod Fertil* [Internet] 1978;**53**:255–257.
772 Ostermeier GC, Miller D, Huntriss JD, Diamond MP, Krawetz SA. Reproductive biology:
773 Delivering spermatozoan RNA to the oocyte. *Nature* [Internet] 2004;**429**:154–154.
774 Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn JL,
775 Regev A, *et al.* Systematic identification of long noncoding RNAs expressed during
776 zebrafish embryogenesis. *Genome Res* [Internet] 2012;**22**:577–591.
777 Pinto SM, Manda SS, Kim M-S, Taylor K, Selvan LDN, Balakrishnan L, Subbannayya T, Yan
778 F, Prasad TSK, Gowda H, *et al.* Functional annotation of proteome encoded by human
779 chromosome 22. *J Proteome Res* [Internet] 2014;**13**:2749–2760.
780 Pollier J, Rombauts S, Goossens A. Analysis of RNA-Seq data with TopHat and Cufflinks for
781 genome-wide expression analysis of jasmonate-treated plants and plant cultures. *Methods*

Mol Biol [Internet] 2013;**1011**:305–315.

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. In Hayashizaki Y, editor. *PLoS Genet* [Internet] 2009;**5**:e1000617.

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, *et al.* Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* [Internet] 2011;**29**:742–749.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM, *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* [Internet] 2014;**42**:D756–D763.

Pundir S, Magrane M, Martin MJ, O'Donovan C, UniProt Consortium. Searching and Navigating UniProt Databases. *Curr Protoc Bioinforma* [Internet] 2015;**50**:1.27.1-10.

Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* [Internet] 2000;**16**:276–277.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* [Internet] 2014;**43**:D670-81.

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-Bonet T, Albà MM. Origins of De Novo Genes in Human and Chimpanzee. In Noonan J, editor. *PLOS Genet* [Internet] 2015;**11**:e1005721.

Schmitz SU, Grote P, Herrmann BG. Mechanisms of long noncoding RNA function in development and disease. *Cell Mol Life Sci* [Internet] 2016;**73**:2491–2509.

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* [Internet]

2013;**41**:4104–4117.

Simpson BJ, Wu FC, Sharpe RM. Isolation of human Leydig cells which are highly responsive to human chorionic gonadotropin. *J Clin Endocrinol Metab* [Internet] 1987;**65**:415–422.

Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Stat Appl Genet Mol Biol* [Internet] 2004;**3**:1–25.

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep* [Internet] 2013;**3**:2179–2190.

Svingen T, Jørgensen A, Rajpert-De Meyts E. Validation of endogenous normalizing genes for expression analyses in adult human testis and germ cell neoplasms. *Mol Hum Reprod* [Internet] 2014;**20**:709–718.

Tao S, Xiu-Lei Z, Xiao-Lin L, Sai-Nan M, Yu-Zhu G, Xiang-Ting W. Recent Progresses of Long Noncoding RNA. *http://www.sciencepublishinggroup.com* [Internet] 2016;**1**:34. Science Publishing Group.

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* [Internet] 2009;**25**:1105–1111.

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* [Internet] 2012;**7**:562–578.

Turner JMA. Meiotic sex chromosome inactivation. *Development* [Internet] 2007;**134**:1823–1831.

Tüttelmann F, Rajpert-De Meyts E, Nieschlag E, Simoni M. Gene polymorphisms and male infertility--a meta-analysis and literature review. *Reprod Biomed Online* [Internet] 2007;**15**:643–658.

Tüttelmann F, Ruckert C, Röpke A. Disorders of spermatogenesis. *medizinische Genet*

[Internet] 2018;**30**:12–20.

Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet* [Internet] 2016;**17**:601–614.

Vaudel M, Barsnes H, Berven FS, Sickmann A, Martens L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* [Internet] 2011;**11**:996–999.

Wang J, Kong L, Gao G, Luo J. A brief introduction to web-based genome browsers. *Brief Bioinform* [Internet] 2013a;**14**:131–143.

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* [Internet] 2013b;**41**:e74.

Wang M, Liu X, Chang G, Chen Y, An G, Yan L, Gao S, Xu Y, Cui Y, Dong J, *et al.* Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis. *Cell Stem Cell* [Internet] 2018;**23**:599–614.e4.

Wen K, Yang L, Xiong T, Di C, Ma D, Wu M, Xue Z, Zhang X, Long L, Zhang W, *et al.* Critical roles of long noncoding RNAs in Drosophila spermatogenesis. *Genome Res* [Internet] 2016;**26**:1233–1244.

Wichman L, Somasundaram S, Breindel C, Valerio DM, McCarrey JR, Hodges CA, Khalil AM. Dynamic expression of long noncoding RNAs reveals their potential roles in spermatogenesis and fertility. *Biol Reprod* [Internet] 2017;**97**:313–323.

Willey S, Roulet V, Reeves JD, Kergadallan M-L, Thomas E, McKnight A, Jégou B, Dejucq-Rainsford N. Human Leydig cells are productively infected by some HIV-2 and SIV strains but not by HIV-1. *AIDS* [Internet] 2003;**17**:183–188.

Xie C, Zhang YE, Chen J-Y, Liu C-J, Zhou W-Z, Li Y, Zhang M, Zhang R, Wei L, Li C-Y. Hominoid-Specific De Novo Protein-Coding Genes Originating from Long Non-Coding

RNAs. In Begun DJ, editor. *PLoS Genet* [Internet] 2012;**8**:e1002942.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, *et al.* Ensembl 2016. *Nucleic Acids Res* [Internet] 2016;**44**:D710–D716.

Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Hum Mol Genet* [Internet] 2015;**24**:R102-10. Oxford University Press.

Zhu Z, Li C, Yang S, Tian R, Wang J, Yuan Q, Dong H, He Z, Wang S, Li Z. Dynamics of the Transcriptome during Human Spermatogenesis: Predicting the Potential Key Genes Regulating Male Gametes Generation. *Sci Rep* [Internet] 2016;**6**:19069.

Zimmermann C, Stévant I, Borel C, Conne B, Pitetti J-L, Calvel P, Kaessmann H, Jégou B, Chalmel F, Nef S. Research resource: the dynamic transcriptional profile of sertoli cells during the progression of spermatogenesis. *Mol Endocrinol* [Internet] 2015;**29**:627–642.

Figure legends

Figure 1 RNA-seq analysis of human testicular cells identifies new genes and transcript isoforms.

Classification of assembled transcripts according to their biotype and their status as known versus novel. Cuffcompare (Trapnell *et al.*, 2012) was used to compare the 25,161 refined transcripts with 362,647 reference transcripts to distinguish between known (Cufflinks class code "=") and novel (class code "j") isoforms of known coding and noncoding genes. This comparison also identified novel unannotated transcripts (NUTs) corresponding to potential new antisense (class code "x"), intronic (class code "i") or intergenic (class code "u") genes.

lncRNA: long noncoding RNA

Figure 2 Expression dynamics across human testicular cells.

A. Expression profiles of differentially expressed (DE) transcripts. After statistical filtration, the 21,264 DE transcripts were clustered into 11 expression patterns (P1-11). The number of transcripts in each expression pattern is given at the top, and their median profile (red line) is plotted as well as the first and third quartiles (Q1 and Q3, gray shading). Samples marked in red and blue correspond to highest and lowest expression values, respectively. Samples marked in orange indicate a slightly lower expression level than maximum abundance, that is, have the second highest expression value. LC = Leydig cells; PC = peritubular cells; SC = Sertoli cells; Spc = spermatocytes; Spt = spermatids; TT = total testis. **B.** Frequency distribution of expression patterns according to transcript biotype. The percentage of transcripts from each of the 11 expression patters (P1-11) is given for known (Cufflinks class code "=") and novel (class code "j") mRNAs and lncRNAs, as well as for intergenic (class code "u"), intronic (class code "i"), and antisense (class code "x") NUTs.

894

895 **Figure 3** Syntenic expression during mammalian spermatogenesis.

896 Heatmap representation of 8,457 transcripts with conserved expression during mammalian
897 spermatogenesis. Each line is a syntenic transcript/region, and each column a
898 sample/experimental condition. The number of transcripts in each expression pattern (P1-P11)
899 is given on the left and their relative expression levels in human (present study), rat (Chalmel
900 *et al.*, 2014), and mouse (Gan *et al.*, 2013; Soumillon *et al.*, 2013) testicular samples are color-
901 coded according to the scale bar (standardized abundance). Samples used for computing
902 expression correlation between humans (present study) and rats (Chalmel et al., 2014) are
903 indicated in red. (p)Spg A/B = (pre)spermatogonia type A/B; (l/p)Spc = (leptotene/pachytene)
904 spermatocytes; r/eSpt = round/elongated spermatids. FPKM: fragments per kilobase of exon
905 model per million reads mapped

906

907 **Supplementary Figure S1** Strategy for refinement assembled transcripts.

908 After transcript reconstruction by Cufflinks, a three-layer filtration strategy was applied: First,
909 only transcripts with an expression of ≥ 3 fragments per kilobase of exon model per million
910 reads mapped (FPKM) in at least one experimental condition (average value of biological
911 replicates) were considered. Second, transcripts with a length less than 200 nucleotides were
912 filtered out. Third, novel transcript isoforms (Cuffcompare class "j") and genes (classes "i", "u"
913 and "x") were required to have at least two exons to be retained.

914

915 **Supplementary Figure S2** Coding potential evaluation of assembled transcripts.

916 First, the protein-encoding potential (PEP) of all 25,161 refined transcripts was evaluated by
917 four distinct bioinformatics tools, i.e. CPAT, HMMER, CPC, and txCdsPredict (Kong *et al.*,

2007; Finn *et al.*, 2011; Kuhn *et al.*, 2013; Wang *et al.*, 2013b). Transcripts that were predicted as protein-coding by two to four tools or by zero or one tool were classified as having high or low PEP, respectively. Second, we used a proteomics informed by transcriptomics (PIT) strategy (Evans et al., 2012) in which a testicular tandem mass spectrometry (MS/MS) proteomics dataset (Kim et al., 2014) was queried against a custom protein database derived from sequences of assembled transcripts. The frequency distribution of transcripts showing high or low PEP and being evidenced (PIT+) or not (PIT-) at the protein level is given for different RNA biotypes. Nearly all mRNAs (97.4%) display a high PEP, and at least one high-confidence peptide was identified in human testis by MS for 67.9% of them. Most long noncoding RNAs (lncRNAs) (76.5%) show low PEP, and were rarely identified by the PIT strategy in the testis proteome. Like lncRNAs, 87.5% of novel unannotated transcripts (NUTs) display low PEP and were almost never (two of 511) identified in the PIT experiment.

Supplementary Figure S3 Differential expression analysis of refined transcripts.

Refined transcripts were considered to be differentially expressed (DE) if they exhibited a fold-change ≥ 3 when all samples were compared to one other, and if they showed a significant expression difference according to a LIMMA statistical test with a false discovery rate-adjusted F-value of ≤ 0.05. Finally, k-means clustering was used to group the 21,264 retained transcripts into 11 expression patterns (P1 to P11).

Supplementary Figure S4 Expression profiles of the human testicular cell markers.

A false-color heatmap summarizes expression profiles of well-known markers for Leydig cells (green), peritubular cells (orange), Sertoli cells (red), spermatogonia (light blue), spermatocytes (blue), spermatids (violet), and germ cells (black). Each line corresponds to a

transcript and each column is a sample. Most Leydig cell markers showed peak expression in these cells (P2) or were either detected in all somatic cells (P1; NR5A1) or in total testis samples (P11; PTGDS and HSD17B3). One peritubular cell marker was preferentially expressed in these cells (P3; ACTA2) while a second was detected in all somatic cells (P1; LMOD1). Two Sertoli cell markers were detected in all somatic cells (P1), whereas expression for most specific markers for mature Sertoli cells peaked in total testis samples (P11). This could suggest that Sertoli cells undergo substantial dedifferentiation when cultured. The latter hypothesis would also explain why a robust marker for immature Sertoli cells showed peak expression in these cells (P4; KRT18). Finally, we found consistent expression profiles for all 22 investigated germ cell markers, including known markers for spermatogonia (P6), spermatocytes (P5-P9), spermatids (P9-10), and germ cells in general (P6-7). LC = Leydig cells; PC = peritubular cells; SC = Sertoli cells; Spc = Spermatocytes; Spt = round spermatids. CCNA1 = cyclin A1; CLU = clusterin; CTSL = cathepsin L; CYP11A1 = cytochrome P450 family 11 subfamily A member 1; DAZL = deleted in azoospermia like; DDX4 = DEAD-box helicase 4; DHCR7 = 7-dehydrocholesterol reductase; DHH = desert hedgehog signaling molecule; FGFR3 = fibroblast growth factor receptor 3; GATA4 = GATA binding protein 4; HSD11B1 = hydroxysteroid 11-beta dehydrogenase 1; HSD17B3 = hydroxysteroid 17-beta dehydrogenase 3; IGF1 = insulin like growth factor 1; INHA = inhibin subunit alpha; INHBB = inhibin subunit beta B; KRT18 = keratin 18; LDHC = lactate dehydrogenase C; LMOD1 = leiomodin 1; MAGEA4 = MAGE family member A4; MEI1 = meiotic double-stranded break formation protein 1; MEIOB = meiosis specific with OB-fold; MNS1 = meiosis specific nuclear structural 1; NR5A1 = nuclear receptor subfamily 5 group A member 1; PHF13 = PHD finger protein 13; PRM1 = protamine 1; PRM2 = protamine 2; PSAP = prosaposin; PTGDS = prostaglandin D2 synthase; SOX9 = SRY-box 9; SPO11 = SPO11 initiator of meiotic double stranded breaks; STAR = steroidogenic acute regulatory protein; SYCP1 = synaptonemal

complex protein 1; SYCP2 = synaptonemal complex protein 2; SYCP3 = synaptonemal complex protein 3; TEX101 = testis expressed 101; TF = transferrin; TNP1 = transition protein 1; TNP2 = transition protein 2; TSPO = translocator protein; TXNDC8 = thioredoxin domain containing 8; VCAM1 = vascular cell adhesion molecule 1; WT1 = WT1 transcription factor.

Supplementary Figure S5 Correlating testicular expression data across technologies.

Over-/Underrepresentation of genes from the 11 expression patterns (RNA-seq data) with those from 13 testicular expression clusters (Microarray data) published by Chalmel and coworkers are shown (Chalmel *et al.*, 2012). The names of expression patterns (P1-P11) and the corresponding numbers of genes are indicated on top of each column, while those for expression clusters (C1-C13) are shown on the left. Each expression cluster is associated with specific testicular cell populations, including prepubertal testicular cells (C1), Leydig and peritubular cells (C2-4), Sertoli cells (C5-7), and germ cells (C8-13). Numbers of loci as observed and expected are given within color-coded rectangles: Red and blue indicate over- and underrepresentation, respectively, according to the scale bar. Numbers in bold indicate significantly over-/underrepresented terms. Genes peaking in somatic cells in the RNA-seq dataset (P1-P4) are significantly overrepresented in clusters C1-C5. P11 (peak expression in total testis samples) shows high enrichment in C5-C7. P5-P10 (progressive peak expression through spermatocytes to spermatids) display a gradual enrichment with clusters C8-C13. Prepub. = prepubertal; LC = Leydig cells; PC = Peritubular cells; SC = Sertoli cells.

Supplementary Figure S6 Correlating RNA-seq data from isolated testicular cells with microarray data from patients with non-obstructive azoospermia.

A heatmap displaying the relative expression levels of known transcripts, including 21,409 mRNAs and 102 lncRNAs, as determined in both our RNA-seq analysis of isolated human testicular cells (left) and a microarray analysis of patients with non-obstructive azoospermia (NOA) (Malcher et al., 2013) (right) is presented. Each line corresponds to a transcript, and each column corresponds to an individual sample (left) or to the average of sample replicates (right). Transcripts are organized according to expression patterns defined in the present study. Relative expression levels are color-coded according to the scale bars. Transcripts showing highest expression in isolated spermatocytes and spermatids show a progressive decreasing signal in biopsies from patients with spermatogenetic arrests at the post-meiotic stage up to Sertoli-cell only syndrome. Conversely, transcripts overexpressed in isolated somatic cells also exhibited strongest expression in biopsies depleted of germ cells. SPC = Spermatocytes; SPT = Spermatids; TT = Total testis; POST = spermatogenesis arrested at the post-meiotic stage; MEI = spermatogenesis arrested at the meiotic stage; PRE = spermatogenesis arrested at the pre-meiotic stage; SCOS = Sertoli cell-only syndrome.

Supplementary Figure S7 Correlating RNA-seq data from isolated testicular cells with single-cell RNA-seq data from patients with normal spermatogenesis or with NOA.

A heatmap displaying the relative expression levels of known transcripts, including 21,696 mRNAs and 455 lncRNAs, as determined in both our RNA-seq analysis of isolated human testicular cells (left) and a single-cell RNA-seq analysis of human testicular cells from patients with normal spermatogenesis or with NOA (Wang *et al.*, 2018) (right) is presented. Each line corresponds to a transcript, and each column corresponds to an individual sample (left) or to the average of several single cells (right). Transcripts are organized according to expression patterns defined in the present study. Relative expression levels are color-coded according to

the scale bars. This comparison show a high consistency of expression profiles obtained from these two different approaches. UMI = Unique Molecular Identifier.

Supplementary Figure S8 Gene ontology term enrichment analysis.

Overrepresented biological processes associated with genes from the 11 expression patterns (P1-P11) are shown. The names of expression patterns are indicated on top of each column. Numbers of loci as observed and expected are given within color-coded rectangles: Red and blue indicate over- and underrepresentation, respectively, according to the scale bar. Numbers in bold indicate significantly overrepresented terms. P1 is enriched for biological processes such as *carbohydrate metabolic process*, *protein glycosylation*, and *vesicle-mediated transport*. P2 is associated with gene ontology (GO) terms related to androgen synthesis (*cholesterol transport*, *steroid biosynthetic process*). P3 is significantly associated with *tube morphogenesis*, *angiogenesis*, and *muscle structure development*. P4 is enriched in terms associated with *neuron projection* and *pyruvate metabolism*; the latter is essential for providing lactate and pyruvate to developing germ cells (Kaiser *et al.*, 2005). P5 is associated with terms related to RNA processing, splicing, and transport. P6 was enriched for *piRNA metabolic process*, *chromatin organization*, and *DNA methylation involved in gamete generation*. P6-7 is associated with *meiotic nuclear division* and *DNA repai*r. P7-P9 are over-represented in GO terms related to flagellum formation such as *cilium assembly* and *cilium movement*. Most germline expression patterns (P6-P10) are enriched in *reproduction* and *spermatogenesis*.

Supplementary Figure S9 Testicular gene expression and sex chromosomal localization. An ideogram of the X (panel A) and Y (panel B) chromosomes as well as the localization of transcripts from the 11 expression patterns P1-P11 are shown. For each expression pattern, the

chromosomal positions of transcripts are displayed as vertical lines that are color-coded according to their corresponding biotype (blue = protein-coding; red = lncRNAs; violet = NUT; gray = other biotypes). Numbers of loci as observed and expected are given within color-coded rectangles: Red and blue indicate over- and underrepresentation, respectively, according to the scale bar. Numbers in bold indicate significant over-/underrepresentation (P value ≤ 0.05). P1 is enriched whereas P5 and P7-P8 are depleted for X-linked genes. A substantial transcriptional reactivation of the sex chromosomes is observed in P9-P10.

Supplementary Figure S10 Quantitative PCR validation of eight NUTs.

Histograms represent expression profiles of candidate genes (+/- SEM) relative to GAPDH mRNA levels. These experiments confirm the expression profiles of selected transcripts and, more importantly, validate the existence of these newly identified genes.

Supplementary Figure S11 A subgroup of meiotic lncRNAs have longer exons.

A. Classification of lncRNAs and NUTs according to their genomic features. All 21,264 DE transcripts underwent multicomponent analysis followed by model-based clustering according to typical genomic features of lncRNAs: expression level (Max. abundance), expression specificity (Shannon entropy), sequence conservation, percentage of GC content, number (N°) of exons, cumulative (Cum.) exon length, average (Av.) exon size, and protein-encoding potential (PEP). Gray dots indicate mRNAs and colored dots lncRNAs and NUTs. This resulted in the classification of lncRNAs and NUTs into eight subgroups (clusters 1-8), including cluster 6, containing transcripts that are much longer because their exons are longer.

B. Violin plot representation of selected genomic features for all differentially expressed mRNAs and lncRNAs as well as for a subgroup (cluster 6) of lncRNAs and NUTs: sequence conservation (phastCons score), number of exons, transcript length (in nucleotides, nt) and

exon length (nt). Transcript size in cluster 6 is significantly larger than that of other lncRNAs ($P < 3.10^{-74}$, Wilcoxon signed-rank test) or of known mRNAs (2275 nt; $P < 2.10^{-11}$). These transcripts have a number of exons similar to that of other lncRNAs, while their exon length is more than five times that of known mRNAs ($P < 7.10^{-89}$) and of other lncRNAs (302 nt; $P < 6.10^{-68}$). Evolutionary sequence conservation is also lower in noncoding transcripts than in mRNAs (P < 9.10-48).

C. Overrepresentation of lncRNAs and NUTs from cluster 6 across expression patterns. The enrichment (-log[p-value], hypergeometric test) is shown and the number of lncRNAs and NUTs from cluster 6 reported in brackets for each expression pattern (P1-P11). Cluster 6 is significantly associated with lncRNAs preferentially transcribed during meiosis (P6).

Figures

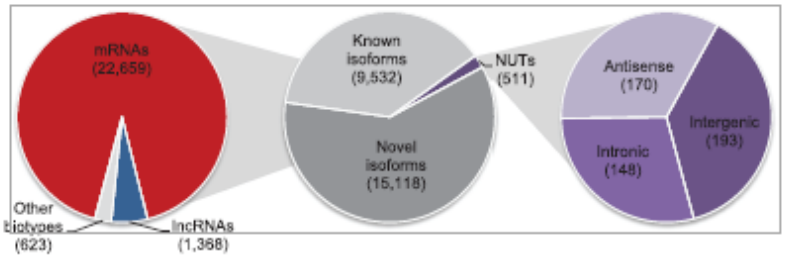

Figure 1 RNA-seq analysis of human testicular cells identifies new genes and transcript isoforms. Classification of assembled transcripts according to their biotype and their status as known versus novel. Cuffcompare (Trapnell *et al.*, 2012) was used to compare the 25 161 refined transcripts with reference transcripts to distinguish between known (Cufflinks class code '=') and novel (class code 'j') isoforms of known coding and noncoding genes. This comparison also identified NUTs corresponding to potential new antisense (class code 'x'), intronic (class code 'i') or intergenic (class code 'u') genes. lncRNA: long noncoding RNA.

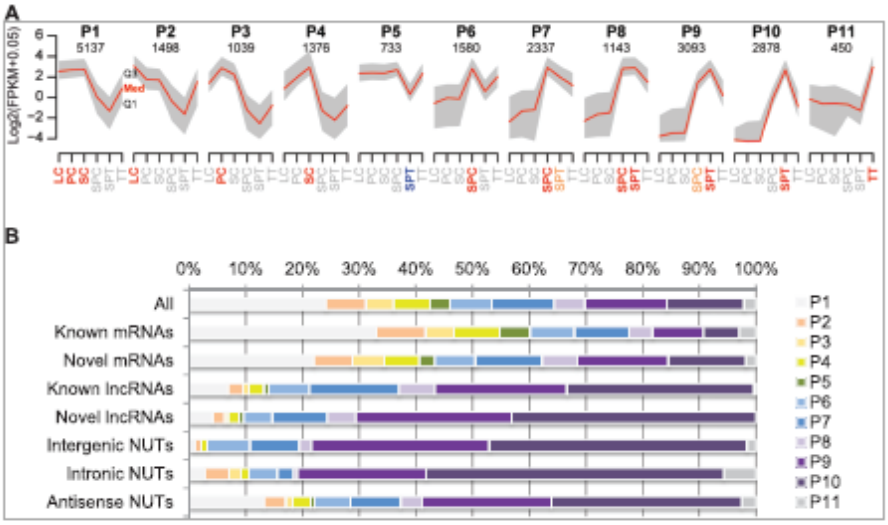

Figure 2 Expression dynamics across human testicular cells. A. Expression profiles of DE transcripts. After statistical filtration, the 21 264 DE transcripts were clustered into 11 expression patterns (P1–P11). The number of transcripts in each expression pattern is given at the top, and their median profile (red line) is plotted as well as the first and third quartiles (Q1 and Q3, gray shading). Samples marked in red and blue correspond to highest and lowest expression values, respectively. Samples marked in orange indicate a slightly lower expression level than maximum abundance, that is, have the second highest expression value. LC = Leydig cells; PC = peritubular cells; SC = Sertoli cells; Spc = spermatocytes; Spt = spermatids; TT = total testis. **B.** Frequency distribution of expression patterns according to transcript biotype. The percentage of transcripts from each of the 11 expression patters (P1–P11) is given for known (Cufflinks class code '=') and novel (class code 'j') mRNAs and lncRNAs, as well as for intergenic (class code 'u'), intronic (class code 'i'), and antisense (class code 'x') NUTs.

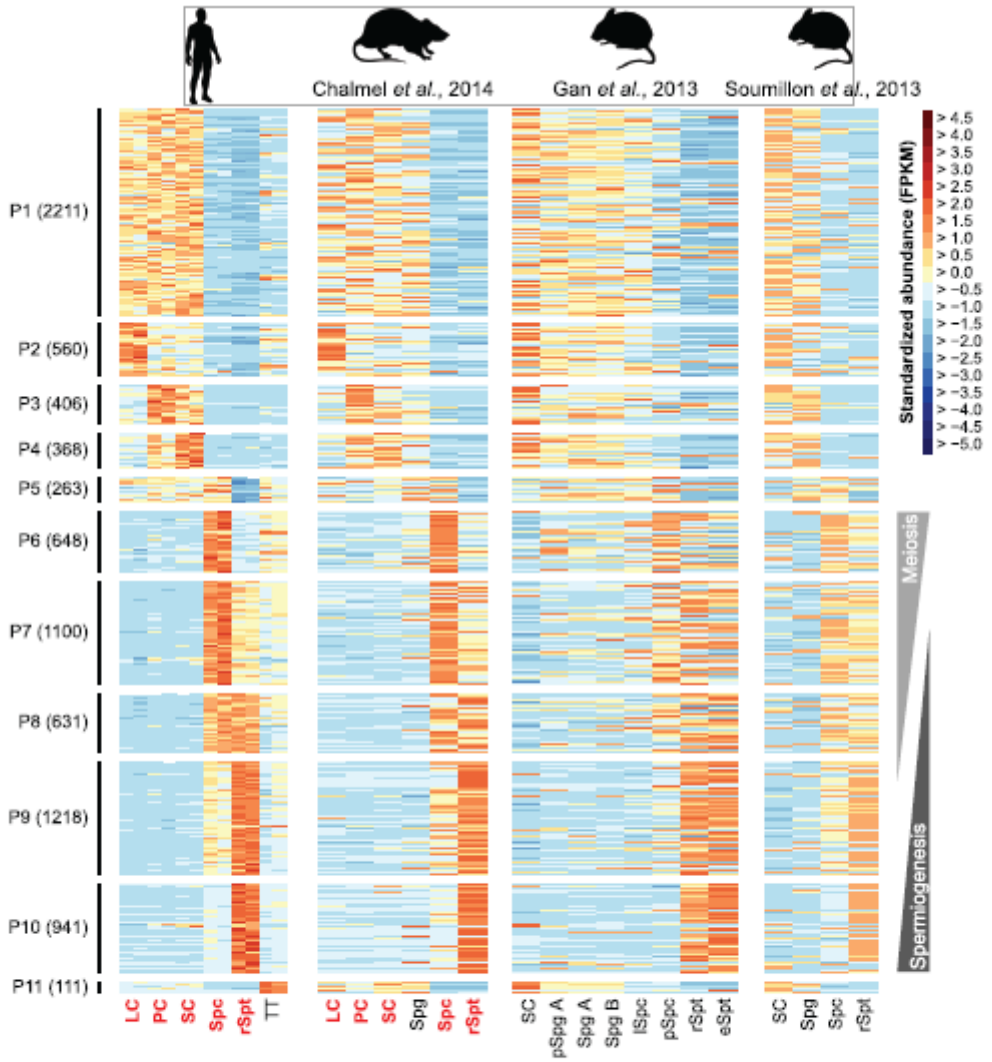

Figure 3 Syntenic expression during mammalian spermatogenesis. Heatmap representation of 8457 transcripts with conserved expression during mammalian spermatogenesis. Each line is a syntenic transcript/region, and each column a sample/experimental condition. The number of transcripts in each expression pattern (P1–P11) is given on the left and their relative expression levels in human (present study), rat (Chalmel *et al.*, 2014) and mouse (Gan *et al.*, 2013; Soumillon *et al.*, 2013) testicular samples are color-coded according to the scale bar (standardized abundance). Samples used for computing expression correlation between humans (present study) and rats (Chalmel et al., 2014) are indicated in red. (p)Spg A/B = (pre)spermatogonia type A/B; (l/p)Spc = (leptotene/pachytene) spermatocytes; r/eSpt = round/elongated spermatids. FPKM: fragments per kilobase of exon model per million reads mapped

Table

Table I Sequence and expression conservation of testicular genes.

	Differentially expressed	Syntenic regions	Syntenic & detected	Syntenic & correlated
Total	21 264	18847	15 119	8457
mRNAs	18915	17848	14 652	8179
lncRNAs	1303	479	181	113
NUTs	484	179	46	20
Other biotypes	562	341	240	145

Statistics of sequence conservation in the rat (Syntenic regions), expression detection in the rat samples (Syntenic and detected), and expression conservation in the rat (Syntenic and correlated) are reported for all DE human transcripts and for distinct transcript biotypes.
lncRNA, long noncoding RNA; NUT, novel unannotated transcribed region.