Cell Reports
Resource

Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis

Magali Soumillon,[1,2,9,10,*] Anamaria Necsulea,[1,2,11] Manuela Weier,[1] David Brawand,[1,2] Xiaolan Zhang,[3] Hongcang Gu,[3] Pauline Barthès,[4] Maria Kokkinaki,[5] Serge Nef,[6] Andreas Gnirke,[3] Martin Dym,[5] Bernard de Massy,[4] Tarjei S. Mikkelsen,[3,7,8] and Henrik Kaessmann[1,2,*]

[1]Center for Integrative Genomics, University of Lausanne
[2]Swiss Institute of Bioinformatics
Génopode, CH-1015, Lausanne, Switzerland
[3]Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA
[4]Institut de Génétique Humaine, UPR1142, CNRS, 34396 Montpellier Cedex 5, France
[5]Department of Biochemistry and Molecular and Cellular Biology, Georgetown University Medical Center, 3900 Reservoir Road NW, Washington, DC 20057, USA
[6]Department of Genetic Medicine and Development, University of Geneva Medical School, University of Geneva, CH-1211 Geneva, Switzerland
[7]Harvard Stem Cell Institute
[8]Harvard Department of Stem Cell & Regenerative Biology
Harvard University, Cambridge, MA 02138, USA
[9]Present address: Harvard Department of Stem Cell & Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
[10]Present address: Broad Institute, Cambridge, MA 02142, USA
[11]Present address: Laboratory of Developmental Genomics, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland
*Correspondence: magalisoumillon@fas.harvard.edu (M.S.), henrik.kaessmann@unil.ch (H.K.)
http://dx.doi.org/10.1016/j.celrep.2013.05.031

SUMMARY

Understanding the extent of genomic transcription and its functional relevance is a central goal in genomics research. However, detailed genome-wide investigations of transcriptome complexity in major mammalian organs have been scarce. Here, using extensive RNA-seq data, we show that transcription of the genome is substantially more widespread in the testis than in other organs across representative mammals. Furthermore, we reveal that meiotic spermatocytes and especially postmeiotic round spermatids have remarkably diverse transcriptomes, which explains the high transcriptome complexity of the testis as a whole. The widespread transcriptional activity in spermatocytes and spermatids encompasses protein-coding and long noncoding RNA genes but also poorly conserves intergenic sequences, suggesting that it may not be of immediate functional relevance. Rather, our analyses of genome-wide epigenetic data suggest that this prevalent transcription, which most likely promoted the birth of new genes during evolution, is facilitated by an overall permissive chromatin in these germ cells that results from extensive chromatin remodeling.

INTRODUCTION

The transcriptome (i.e., the full set of RNA molecules in a tissue or constituent cells) represents a key connection between genomic information and phenotype. Consequently, mammalian transcriptomes have been extensively studied with the use of hybridization or sequencing technologies, which provided insights into the number of transcribed protein-coding genes and spatial expression patterns (Su et al., 2004; Velculescu et al., 1995; Zhang et al., 2004). However, a deeper understanding of mammalian transcriptome complexity has only recently begun to emerge thanks to the advent of new technologies, in particular high-throughput RNA sequencing (RNA-seq) (Wang et al., 2009).

A large proportion of the mammalian genome is transcribed into messenger RNAs (mRNAs) and various types of noncoding RNA molecules (Brawand et al., 2011; Djebali et al., 2012), but the precise extent of genomic transcription and its functional relevance (in particular that of noncoding transcripts) remain unclear (Ponting and Belgard, 2010). Also, transcriptomes have mainly been characterized for cell lines (Djebali et al., 2012), leaving differences in transcriptome complexity among mammalian organs largely unexplored. However, extensive RNA-seq data sets for several organs from different mammals have recently become available (Brawand et al., 2011), and an initial analysis of human RNA-seq data showed that certain organs, such as brain and especially testis, express more protein-coding genes than others (Ramsköld et al., 2009).

Here we report a detailed genome-wide assessment of transcriptome complexity in major mammalian organs. Using

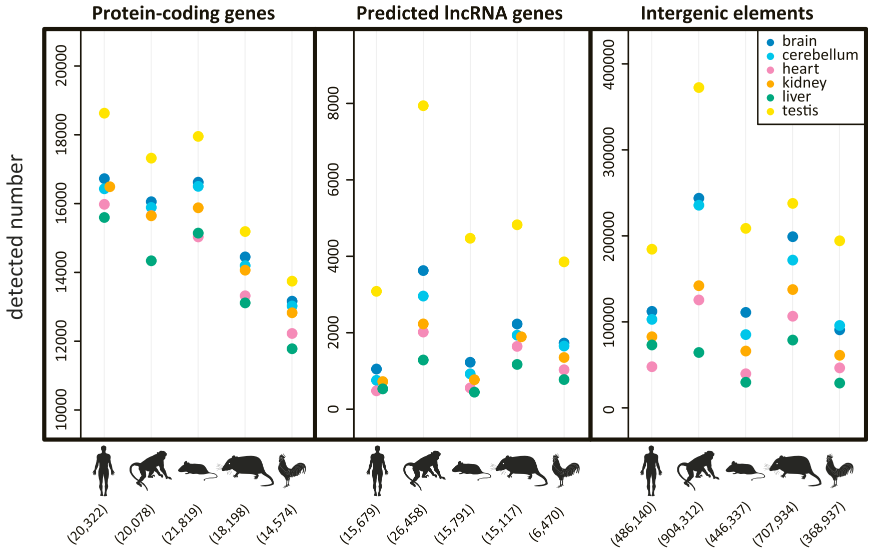

Figure 1. Transcriptome Complexity of the Mammalian and Avian Testis

Number of autosomal transcribed protein-coding genes, predicted lncRNA genes, and transcribed intergenic elements (pseudogenes, transposable elements, and other intergenic sequences) in six organs from five species (from left to right: human, rhesus macaque, mouse, opossum, and chicken), based on 8 million randomly selected RNA-seq reads per sample (total number of elements of each type indicated in brackets).

See also Figure S1.

an extensive set of mammalian RNA-seq data (Brawand et al., 2011), we show that substantially more genic and intergenic regions are transcribed in the testis compared with other organs. Further in-depth transcriptome analyses of all major testicular cell types in the mouse reveal that the high complexity of the testis transcriptome stems mainly from widespread transcription of both functional (genic) and potentially nonfunctional (poorly conserved) portions of the genome (i.e., pseudogenes, transposable elements, and other intergenic sequences) during and after meiosis. Genome-wide epigenetic data relate these patterns to a transcriptionally permissive chromatin state that presumably is associated with the continuous repackaging of the DNA during these spermatogenic stages. Thus, our study reveals that the maturation of germ cells during spermatogenesis leads to promiscuous transcription of the genome, with important functional and evolutionary consequences.

RESULTS

Widespread Genomic Transcription in the Mammalian and Avian Testis

To assess global patterns of transcriptome complexity (i.e., the diversity of transcripts of different types), we exploited a set of RNA-seq data that we generated for six organs from species that represent all main mammalian lineages and birds (Brawand et al., 2011). We found that autosomal protein-coding genes are more frequently transcribed in testis than in other organs (Benjamini-Hochberg corrected $p < 10^{-10}$, chi-square test; Figure 1 and Figure S1) and that many genes are also transcribed in the nervous tissues and kidney, whereas fewer genes are transcribed in heart and liver, consistent with recent estimates for humans (Ramsköld et al., 2009). Remarkably, the skew toward testis was even more pronounced for predicted long noncoding RNA (lncRNA) genes and the intergenic remainder of the genome, which includes pseudogenes, transposable elements, and other intergenic sequences (Figures 1 and S1).

High Transcriptome Complexity during and after Male Meiosis in the Mouse

To explore the cellular source of the high testis transcriptome complexity, we used mouse as a model system. We isolated cell populations for the five major cell types involved in spermatogenesis (see Experimental Procedures). These include Sertoli cells, which are somatic cells that nurture the germ cells during spermatogenesis and constitute <1% of all cells in the testis, and four cell types that represent the main stages of germ cell differentiation (Grabske et al., 1975; Figure 2A): type A spermatogonia (mitotic precursor germ cells, a mix of spermatogonial stem cells and differentiated spermatogonia, representing ~1% of all testis cells), pachytene spermatocytes (which represent the developmental stage of spermatocytes in which meiotic crossover occurs and are derived from differentiated spermatogonia via intermediate steps; ~5% of testis cells), round spermatids (haploid germ cells derived from spermatocytes; ~45% of testis cells), and spermatozoa (mature sperm cells derived from elongated spermatids; ~31% of testis cells). We produced deep-coverage, strand-specific RNA-seq data for the polyadenylated RNA fraction of each cell type, resulting in >60 million mapped reads of 76 bp per sample. For comparison, we collected similar deep-coverage data for representative mouse organs spanning various degrees of transcriptome complexity: testis (high complexity), brain (intermediate to high), and liver (low complexity).

Our analyses of these deep-coverage mouse data confirm the cross-mammalian observation in that a significantly larger number of autosomal protein-coding genes are transcribed in testis (~18,700) than in brain (~18,000) and liver (~15,500; Figure 2B; corrected $p < 10^{-10}$, chi-square test). We detected transcripts for a large number of genes (~18,000) in spermatogonia, spermatocytes, and spermatids, whereas fewer genes (~15,500) are transcribed in Sertoli cells. We detected transcripts for a relatively large number of genes (~16,900) in spermatozoa, which may seem surprising given that the chromatin in mature sperm is generally thought to be transcriptionally inert (Johnson et al., 2011). However, most transcripts detected in spermatozoa may represent residual transcripts from their precursor cells (i.e., elongated spermatids; Johnson et al., 2011). Consistently, a read-coverage analysis revealed a decreased 3′ representation of spermatozoa transcripts (Figure S2A), which might

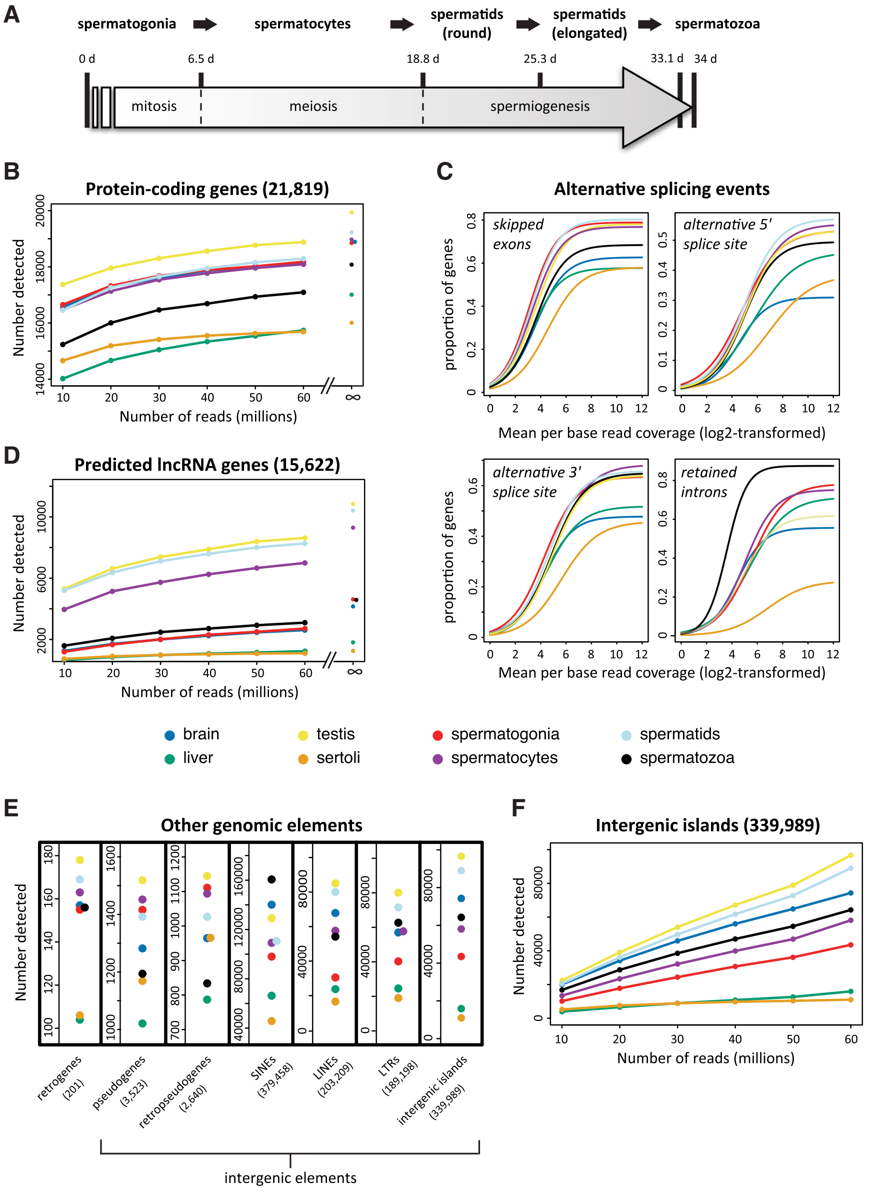

Figure 2. Transcriptome Complexity during Mouse Spermatogenesis

(A) Overview of germ cell differentiation during mouse spermatogenesis.

(B) Number of autosomal protein-coding genes for which transcripts were detected in two somatic tissues, total testis, and five testicular cell types, on the basis of 10–60 million randomly selected RNA-seq reads for each tissue. Right: theoretical number of transcribed genes when assuming no read limitation (see also Experimental Procedures).

(C) Mathematical modeling of the proportion of protein-coding genes affected by alternative splicing events (skipped exons, alternative 5′/3′ splice sites and retained introns) as a function of gene-expression levels (see also Experimental Procedures) for the different tissues. The modeling is based on observed frequencies of alternative splicing events (see Figure S2 for details).

(D) Number of autosomal predicted lncRNA genes for which transcripts were detected. Right: theoretical number of transcribed genes when assuming no read limitation.

(E) Number of other autosomal transcribed genomic elements detected based on 60 million randomly sampled reads per tissue.

(F) Number of autosomal intergenic islands for which transcripts were detected.

suggest that a large proportion of transcripts in mature sperm are degraded. Importantly, read resampling confirmed that the available 60 million reads were sufficient to detect most of the theoretically predicted transcripts in a given sample (Figures 2B and S2B). It is also noteworthy that out of the ~18,700 autosomal protein-coding genes that were detected in samples from whole testis, only 190 (~1%) were not detected in any of the individual spermatogenic cell types.

Germ cell transcriptomes are also characterized by complex alternative splicing patterns. We find that larger proportions of protein-coding genes have splice variants in germ cells (especially in spermatocytes/spermatids, as well as spermatozoa with respect to retained introns) and total testis than in somatic tissues and Sertoli cells (Figures 2C and S2C; corrected $p < 0.01$, randomization test; Experimental Procedures). These results agree with the frequent and specific alternative transcript isoforms identified for individual protein-coding genes expressed in meiotic and postmeiotic cells (Kleene, 2001).

The difference between testis and other organs in terms of transcript diversity is pronounced for noncoding transcripts. Out of 15,622 lncRNA candidate genes, many more were detected in testis (~8,600) than in brain (~2,600) or liver (~1,000), consistent with a recent study in humans (Cabili et al., 2011; Figure 2D). Especially large numbers of lncRNAs are transcribed in spermatids and spermatocytes, whereas much fewer lncRNA transcripts were detected in the other germ cells (Figure 2D).

The testis also transcribes larger numbers of intergenic elements compared with other tissues (Figures 2E and 2F). Thus, a much larger portion of intergenic sequence was covered by RNA-seq reads in testis (~31.7 Mb) than in brain (~20.2 Mb) or liver (~7.2 Mb). Transcripts cover particularly large intergenic regions in spermatids (~29.0 Mb), spermatocytes (~20.8 Mb), and spermatozoa (~21.6 Mb), whereas substantially less intergenic sequence is covered by RNA-seq reads in spermatogonia (~12.7 Mb) and Sertoli cells (~5 Mb).

The expression levels of the different genomic elements vary among tissues (Figures 3A, 3B, and S3A). Spermatocytes and spermatids show a distinct distribution of protein-coding gene expression levels (Figure 3B). For example, they show a larger

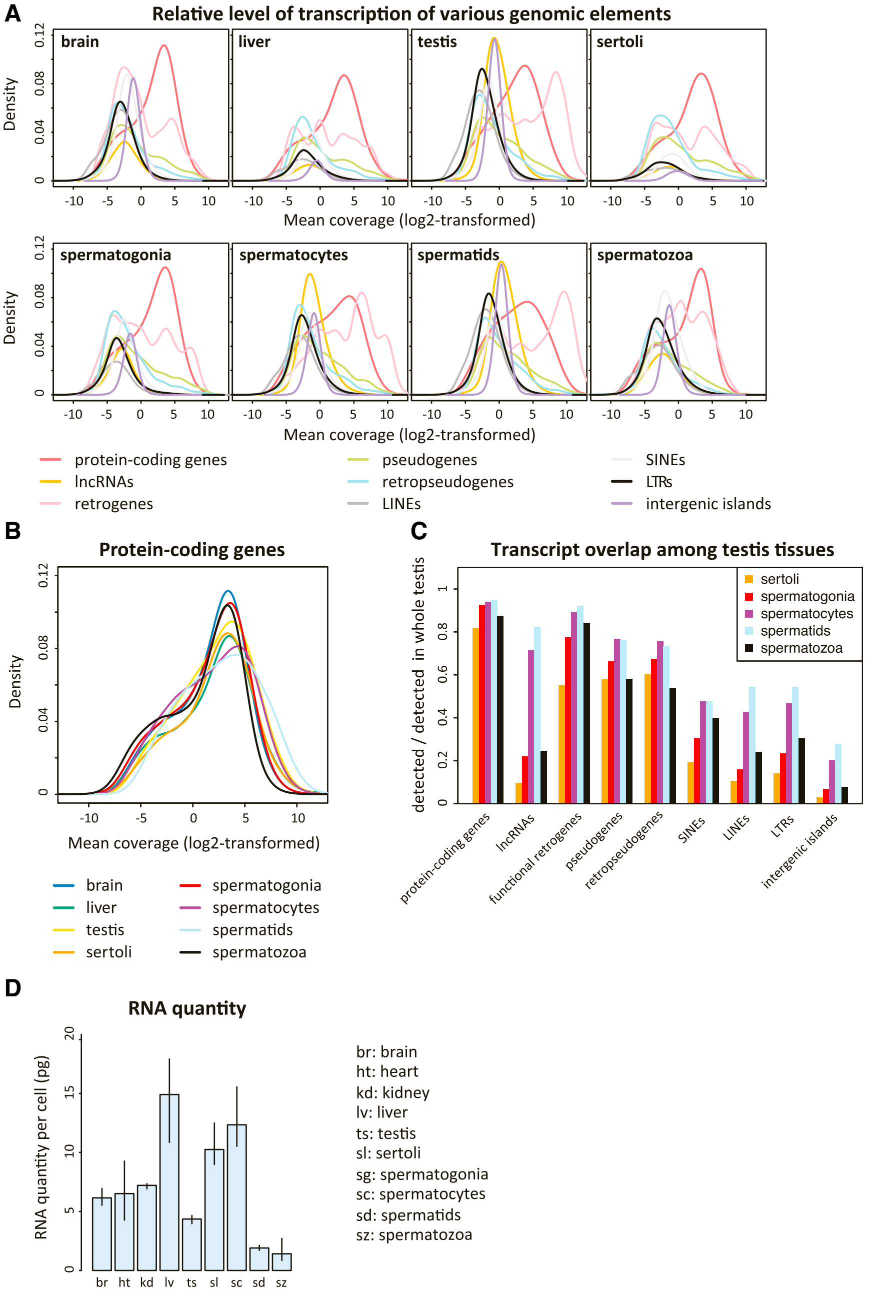

Figure 3. Relative Transcript Levels and Cellular RNA Quantities

(A) Density plots of the total per-base read coverage for the different types of autosomal genomic elements in the different tissues (see also Figure S3).

(B) Density plot of the total per-base read coverage for autosomal protein-coding genes in the different tissues.

(C) Overlap of transcribed elements of the testis and the five testicular cell types, respectively.

(D) Average amounts of total RNA produced by single cells (pg) in the different tissues. Estimates are based on simultaneous DNA/RNA extractions (see also Experimental Procedures and main text). Error bars: range between minimum and maximum values among biological replicates ($n \geq 3$); 21 out of 36 pairwise comparisons of RNA amounts reveal significant differences between tissues ($p < 0.05$, Tukey's post hoc test). In particular, the RNA content of liver, spermatocytes, and Sertoli cells was significantly higher than that of other types of cells.

proportion of genes with high expression levels ($\log_2$ mean coverage > 7) compared with other germ cells or somatic tissues, potentially reflecting specific functional requirements. All types of intergenic elements tend to be more highly expressed in spermatids (Figure S3), further supporting the notion of widespread promiscuous transcription in this cell type.

Expression-level patterns are strikingly similar between spermatocytes and testis, and between spermatids and testis (Figures 3A and S3B), which indicates that spermatocytes and especially spermatids contribute significantly to the transcriptome of the testis as a whole. In support of this notion, we found that these two cell types show the most substantial overlap of transcribed elements with testis (Figure 3C). The presumably substantial contribution of spermatids to the testis transcriptome is consistent with the high transcriptome complexity observed in both spermatids and testis, and with the fact that round spermatids constitute a large proportion (~45%) of all testis cells. In contrast, pachytene spermatocytes represent only a small proportion (~5%) of all testis cells and might therefore be expected to contribute less to the testis transcriptome. However, in addition to the number of cells, the amount of RNA produced per cell determines the contribution of a cell type to the transcriptome of the whole organ. We thus estimated the amount of RNA per cell for each tissue based on quantifications of simultaneously extracted RNA and DNA, and given the amount of DNA per genome equivalent (Experimental Procedures). These analyses revealed significant cellular RNA quantity differences among tissues (one-way ANOVA, $p < 10^{-9}$). Consistent with previous estimates (Schmidt and Schibler, 1995), we found that the average liver cell produces larger quantities of RNA (~15 pg RNA per cell) than cells from other tissues (e.g., brain: ~6 pg RNA per cell; Figure 3D). Interestingly, spermatocytes also produce large amounts of RNA (~12 pg) and contain ~6 times more RNA than round spermatids ($p < 0.05$, Tukey's post hoc test). Thus, although pachytene spermatocytes are ~9 times less abundant than round spermatids, they nevertheless are likely to contribute substantially to the testis transcriptome.

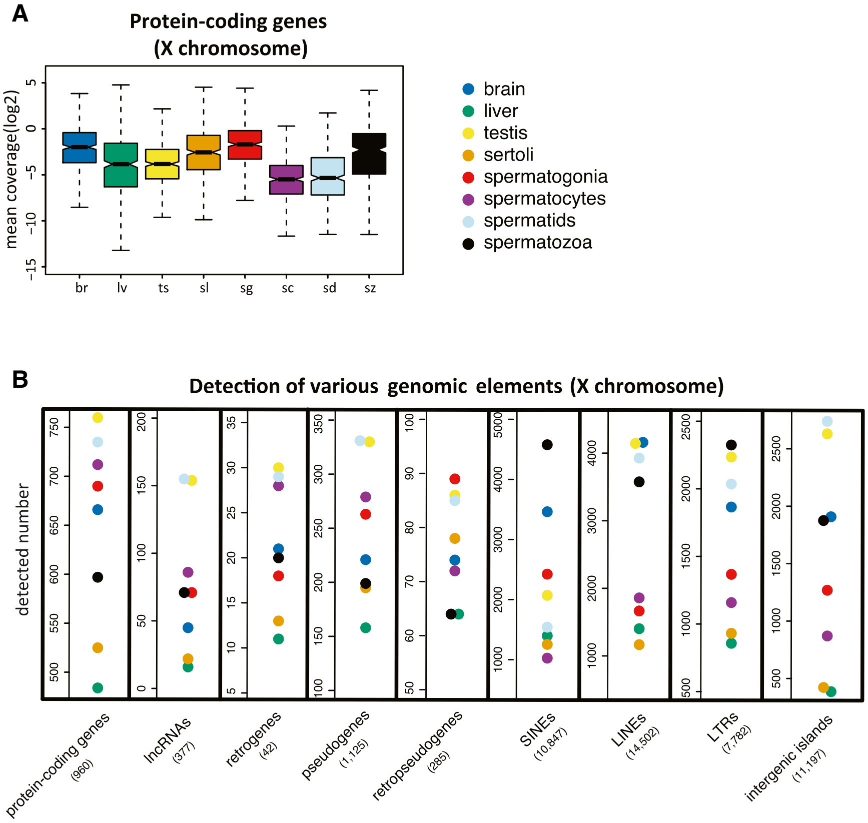

Figure 4. MSCI and Testis Transcriptome Complexity on the X Chromosome

(A) Distribution of gene-expression levels ($\log_2$ mean per-base read coverage) of protein-coding genes on the X chromosome.

(B) Number of transcribed genomic elements on the X chromosome detected based on 60 million randomly sampled reads per tissue.

Facilitated Transcription of Duplicate Gene Copies in Meiotic and Postmeiotic Cells

It has been hypothesized that the testis promotes the birth of new mammalian genes, given that many new genes are often transcribed in this tissue (Kaessmann, 2010; Kaessmann et al., 2009). To evaluate this hypothesis, we analyzed the transcriptional activity of duplicate genes, the major raw material underlying new gene origination (Kaessmann, 2010). We found that protein-coding genes located in chromosomal regions that recently experienced segmental duplications in the mouse (She et al., 2008) were more highly expressed in spermatids and spermatocytes (and total testis) relative to the other tissues, when compared with genes in the rest of the genome (Table S1). Our results suggest that transcription of duplicated genes, which often lack their ancestral regulatory elements (Kaessmann, 2010), is facilitated in spermatocytes and spermatids, which may explain why transcripts that arose from recent segmental duplicons in primates show a strong tendency to be specifically transcribed in testis (She et al., 2004).

We also analyzed transcription of intronless gene copies (so-called retrocopies) that originated through the reverse transcription of mRNAs of parental source genes (Kaessmann et al., 2009). Retrocopies are particularly informative regarding the mechanisms for new gene transcription, as they usually lack regulatory elements. Our analysis showed that retrocopies with truncated open reading frames (retropseudogenes) are transcribed at significantly higher levels in spermatids than in the other tissues (Table S1; corrected $p < 10^{-10}$; Mann-Whitney *U* test). Functional retrocopies (retrogenes) show a striking upregulation in both spermatids and spermatocytes (corrected $p < 10^{-10}$; Mann-Whitney *U* test; Figures 3A and S3A; Table S1), consistent with previous observations (Potrzebowski et al., 2008). Thus, our results suggest that transcription of newly emerged retrocopies is facilitated in these germ cells.

Transcription Patterns on the X Chromosome in Germ Cells

The transcriptional activity of sex chromosomes is suppressed during and, to a lesser extent, after male meiosis due to epigenetic chromatin modifications, a process termed meiotic sex chromosome inactivation (MSCI) (Turner, 2007). Our analyses revealed reduced expression levels of protein-coding genes on the X chromosome in spermatocytes and spermatids (Figure 4A; corrected $p < 10^{-9}$ and $p < 0.05$, respectively; Mann-Whitney *U* test), consistent with the action of MSCI. Surprisingly, we nevertheless detected large numbers of transcripts for most X-linked genomic elements in spermatocytes and especially in spermatids (Figure 4B), indicating that in particular postmeiotic chromatin silencing is not complete and allows for widespread low-level transcription on the X.

We also used our data to evaluate previous suggestions pertaining to the enrichment of genes with male-biased expression (function) on the X chromosome. To do this, we first separated genes into a set that contained all genes with 1:1 orthologs in human (i.e., genes presumably present on the ancestral X chromosome) and a set with the remaining genes (i.e., a set enriched for recently emerged genes), following our previously described procedure (Julien et al., 2012). An analysis of spatial expression specificity (i.e., 3-fold higher expression in a tissue compared with the other tissues; minimum expression level: $\log_2$ mean coverage > 1) shows that brain-specific genes are enriched on the X for the ancient gene set (X: 5%, autosomes [A]: 3.3%, $p < 0.01$; chi-square test), consistent with previous studies (Julien et al., 2012; Zhang et al., 2010). Although liver genes show no (ancient genes) or even a paucity of (recent genes; X: 0.2%, A: 4.2%, $p < 10^{-4}$) X enrichment, spermatogenic cells show interesting patterns. First, spermatogonia-specific genes are significantly enriched on the X for the set of ancient genes (X: 4.2%, A: 2.3%, $p < 10^{-3}$), consistent with earlier findings (Wang et al., 2001), but not for recent genes. Second,

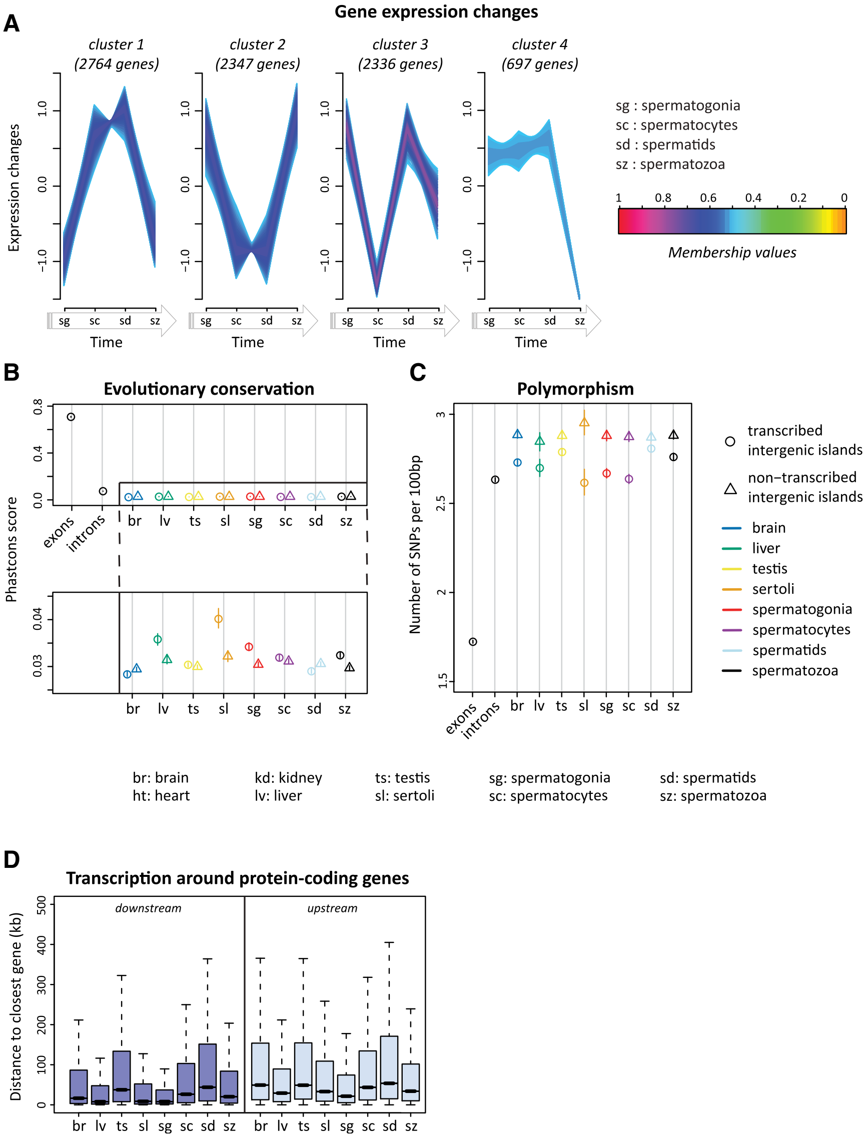

Figure 5. Functional Relevance and Potential Origin of Complex Spermatogenic Transcription Patterns

(A) Clusters of autosomal protein-coding genes with common expression patterns during spermatogenesis.

(B) Evolutionary conservation (mean of mammalian PhastCons score) of transcribed and nontranscribed intergenic islands. Exon/intron conservation levels are shown for comparison. Error bars are based on bootstrapping analysis (1,000 replicates).

(C) Number of SNPs detected (per 100 bp) for transcribed intergenic islands and nontranscribed intergenic islands. Error bars are based on bootstrapping analysis (1,000 replicates).

(D) Distance (upstream and downstream) of transcribed intergenic islands from the nearest protein-coding gene in the different tissues.

See also Extended Discussion.

spermatid-specific genes are highly enriched among recent X-linked genes (X: 45%, A: 11%, $p < 10^{-4}$), but not for the ancient gene set, consistent with previous observations (Julien et al., 2012; Mueller et al., 2008). Spermatozoa-specific genes show no significant difference between the X and autosomes, but, interestingly, genes with Sertoli-cell-specific expression are also significantly enriched on the X for the ancient gene set (X: 3.8%, A: 1.9%; $p < 0.001$), thus revealing yet another facet of the prominent role of the X chromosome in male germ cell development.

Functional Relevance of Spermatogenic Transcription Patterns

To assess the functional relevance of the complex transcriptome patterns observed throughout spermatogenesis, we evaluated the expression of protein-coding genes in the different spermatogenic cells. Soft-clustering analysis revealed four clusters of autosomal genes that show distinct expression-level profiles ($p < 0.05$) and are significantly enriched with functional Gene Ontology (GO) categories related to spermatogenesis (Figure 5A; Table S2). See Extended Discussion for a detailed description of these clusters.

In addition, we evaluated the functional relevance of intergenic transcription by assessing the extent to which transcribed sequences are conserved between species and among mouse strains, based on mammalian base conservation scores and SNP densities. As expected, we found that unannotated intergenic sequences are poorly conserved compared with exons of protein-coding genes, even if they are transcribed (Figures 5B and 5C). Nevertheless, transcribed intergenic islands tend to be more conserved than nontranscribed intergenic sequences, in particular during recent evolution, indicating that a fraction of these sequences might be selectively constrained and hence functional (Figures 5B and 5C).

Potential Mechanisms Underlying the Complex Transcriptomes of Spermatocytes and Spermatids

Several mechanisms might underlie the widespread and (presumably) often nonfunctional transcriptional activity of meiotic and postmeiotic cells. For example, if transcription-coupled repair (TCR) helps minimize DNA mutations in the germline, then mechanisms that promote genome-wide transcription might contribute to maintaining DNA integrity during spermatogenesis. However, although TCR is probably occurring in the germline (Arnheim and Calabrese, 2009), it is unclear why it would be particularly important for meiotic and postmeiotic germ cells, and whether the observed intergenic transcription in spermatocytes and spermatids, which still covers only a

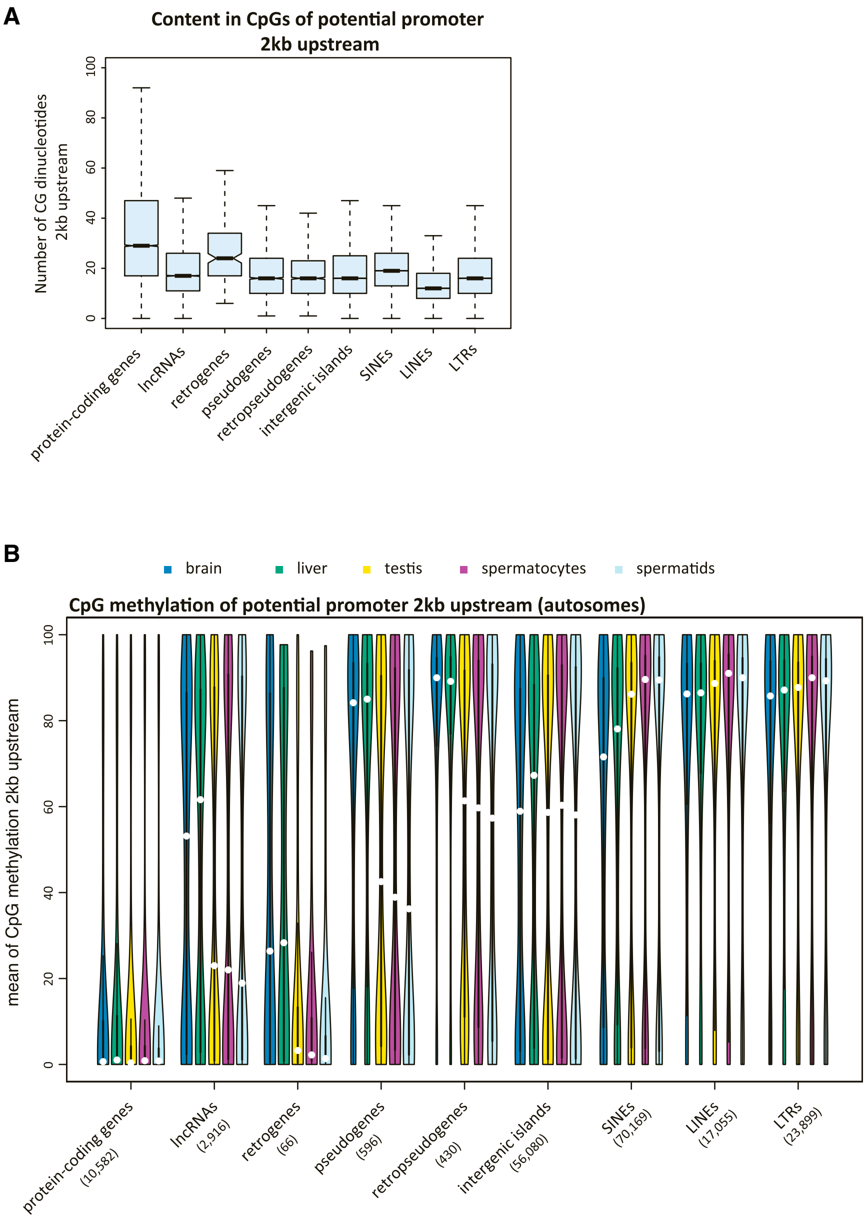

Figure 6. DNA Methylation of Regions Upstream of Transcribed Genomic Elements

(A) Distribution of CpG content upstream of various transcribed autosomal elements.

(B) Violin plot showing the distribution of CpG methylation levels upstream of genomic elements (mean percentages of CpG methylation levels 2 kb upstream). The white point represents the median. See Figure S4 for X chromosome patterns.

limited portion of the genome, would suffice to support pronounced TCR in these cells.

Alternatively, transcription of nonfunctional intergenic sequences could be associated with genic transcription, if extensive transcriptional readthrough leads to transcription of downstream intergenic sequence or facilitates transcription by opening up the chromatin. We thus took advantage of our strand-specific data to evaluate the extent of the transcriptional activity of intergenic DNA around protein-coding genes, which showed that transcribed intergenic sequences tend to be closer to the 3′ end of genes in somatic tissues than in testis and meiotic/postmeiotic germ cells (Figure 5D; $p < 10^{-10}$; Mann-Whitney *U* test). Especially in spermatids, transcribed intergenic sequences are located relatively far downstream of genes (median ~43.9 kb) compared with, for example, the brain (median: ~16.5 kb). Thus, transcriptional readthrough likely explains less of intergenic transcription in germ cells than in somatic tissues. Transcribed intergenic sequences are also significantly more distant from the 5′ end of genes in spermatids than in the somatic tissues (Figure 5D; $p < 10^{-10}$, Mann-Whitney *U* test), suggesting that intergenic transcription in spermatids is not facilitated by a more open chromatin conformation upstream of protein-coding genes. Thus, the excessive intergenic transcription in meiotic/postmeiotic cells does not seem to result from extensive transcription of genes.

DNA Demethylation Contributes to Promiscuous Transcription in Spermatocytes and Spermatids

Next, we explored whether the complex transcriptome patterns in meiotic and postmeiotic cells are associated with epigenetic modification, which potentially is related to the extensive chromatin remodeling events during spermatogenesis (Kimmins and Sassone-Corsi, 2005). Specifically, we assessed whether demethylation of CpG dinucleotides in promoters, which is associated with active (open) chromatin (Deaton and Bird, 2011), contributed to the transcriptional activity of spermatocytes/spermatids.

To do this, we generated representative genome-wide DNA methylation data for spermatocytes, spermatids, and three control tissues (brain, liver, and testis; Experimental Procedures). We assessed the extent of CpG methylation in putative promoter regions (i.e., 2 kb upstream of the transcriptional start site [TSS] or the annotated 5′ end of the element; Experimental Procedures) of all transcribed elements in the different samples. Our analyses confirmed the expectation that promoters of protein-coding genes, which often have a high CpG content compared with other transcribed genomic elements (Figure 6A), show very low CpG methylation levels in all tissues and cells (Figure 6B; median of CpG methylation below 1%). Also as expected, methylation levels are negatively correlated with expression levels (Spearman's *rho* between -0.28 and -0.12 in the different tissues; corrected $p < 10^{-10}$). We also assessed the proportion of promoters that are highly methylated (mean CpG methylation level > 50%) and found fewer protein-coding genes with highly methylated promoters in testis, spermatocytes, and spermatids (~11%) compared with the two somatic tissues (~12% and ~13%, respectively; corrected $p < 10^{-4}$; chi-square test), in agreement with the larger number of protein-coding genes transcribed in the germline tissues.

Potentially functional lncRNAs, retrogenes, and pseudogenes also show significantly lower promoter CpG methylation levels in the germline (in particular in spermatids) than in brain and liver (Figure 6B; corrected $p < 10^{-2}$; Mann-Whitney *U* test). Whereas median methylation levels for intergenic islands are not significantly lower in the germline tissues than in brain, a significantly larger proportion (~48%) of potential promoters show low (<50%) methylation levels for these elements compared with the somatic tissues (~42%–45% of promoters with low methylation; corrected $p < 10^{-4}$; chi-square test).

Regions upstream of annotated transposable elements, which are often not full length due to partial retrotransposition events and/or decay, are generally highly methylated in all tissues or cell types (Figure 6B). Moreover, CpG methylation levels in these regions are higher in the germline than in brain and liver (corrected $p < 10^{-7}$; Mann-Whitney *U* test), consistent with the previously described role of CpG methylation in repressing the expression and mobility of transposable elements (Slotkin and Martienssen, 2007). These results for autosomes generally extend to the X chromosome, with particularly pronounced demethylation patterns in spermatids (Figure S4). This suggests that CpG demethylation may contribute to the widespread transcriptional activity of the X during and especially after meiosis.

High Levels of H3K4me2 Suggest an Overall Open Chromatin State in Spermatocytes and Spermatids

Chromatin reorganization in spermatocytes and spermatids involves histone replacement and modification events that may favor an open chromatin conformation (Kimmins and Sassone-Corsi, 2005). In addition to a transcriptionally active chromatin state at CpG-rich promoters (see above), an overall open chromatin conformation might facilitate transcription throughout the genome in these cells. To test this hypothesis on a genome-wide scale, we generated chromatin immunoprecipitation sequencing (ChIP-seq) data for the dimethylation of histone H3 at lysine 4 (H3K4me2), a marker of transcriptional activity and open chromatin configuration (Barski et al., 2007; Experimental Procedures).

Our analyses revealed significantly higher levels of H3K4me2 around the TSS of protein-coding genes compared with nearby upstream DNA (Figure 7; corrected $p < 10^{-10}$; Mann-Whitney *U* test). Furthermore, H3K4me2 levels are significantly positively correlated with transcription levels of these genes (Figure 7), in particular for the region upstream of the TSS, as the downstream signal is weakened by the competitive H3K4me3 methylation state (Barski et al., 2007). Notably, the H3K4me2 signal is stronger for spermatocytes (fold enrichment of the H3K4me2 signal in promoter region relative to further upstream region: 5.2) than for the other tissues (fold enrichment: 3.3–4.4).

Similar H3K4me2 signatures are present around the predicted TSS of lncRNA genes (Figure 7) and are significantly correlated with expression levels (*rho* > 0.19; Benjamini-Hochberg corrected $p < 10^{-10}$), suggesting that opening of chromatin in the promoter region facilitates the expression of lncRNAs as well. The less distinct methylation signatures for lncRNAs compared with protein-coding genes are likely due to the imprecise definition of TSSs and/or to their overall low transcription levels (Figures 3A and S3). Notably, lncRNA H3K4me2 signals are stronger for the three germline tissues (fold enrichment: 1.5 in total testis, 1.7 in spermatids, 1.9 in spermatocytes) than for brain and liver (fold enrichment: 1.3 and 1.4, respectively).

In contrast to the brain and liver, all germline samples and in particular spermatocytes show significantly elevated levels of H3K4me2 upstream of retrogenes, pseudogenes, and retropseudogenes (Figure 7; fold enrichment from 1.3 [testis pseudogenes] to 2.2 [spermatocytes retrogenes]; corrected $p < 10^{-9}$; Mann-Whitney *U* test), and significant correlations of H3K4me2 levels with the transcription levels of these elements (Figure 7; *rho*: 0.13–0.34; corrected $p < 0.02$). Finally, we detected elevated H3K4me2 levels upstream of unannotated intergenic islands and transposable elements that (except for LINEs) are weakly (*rho*: 0.001–0.05) but significantly (corrected $p < 2 \times 10^{-2}$) correlated with transcript levels and are particularly pronounced for elements of the highest-expression class in spermatocytes and spermatids (fold enrichment: 1.9 and 1.3, respectively, versus 1.1 for both brain and liver; Figure 7).

Overall, our observations suggest a generally more open chromatin state in spermatocytes and spermatids that facilitates transcription of both genic and intergenic genomic elements and thus causes extensive transcriptional background noise in these cells.

DISCUSSION

Since the 1990s, studies of individual protein-coding genes have reported peculiar patterns of transcription in spermatogenic cells, such as overexpression of several mRNAs, frequent generation of specific transcript isoforms through alternative promoters, and truncation of transcripts due to upstream polyadenylation sites (Kleene, 2001). Here we assessed transcriptional patterns in the testis and its constituent cells on a genome-wide scale. Our analyses revealed substantially more widespread transcription of the genome in the testis than in other organs in representative mammals and a bird. Thus, promiscuous transcriptional activity in the testis is common to amniotes.

Our deep transcriptome sequencing analyses of all major testis cell types in the mouse show that postmeiotic spermatids and to a lesser extent meiotic spermatocytes have highly complex transcriptomes. The widespread transcriptional activity of the genome in these cells covers protein-coding genes, consistent with an increased expression activity during and after meiosis of this type of gene detected in previous microarray-based studies (Schultz et al., 2003; Shima et al., 2004). However, the widespread transcription pattern is particularly pronounced for lncRNA candidate genes and also includes

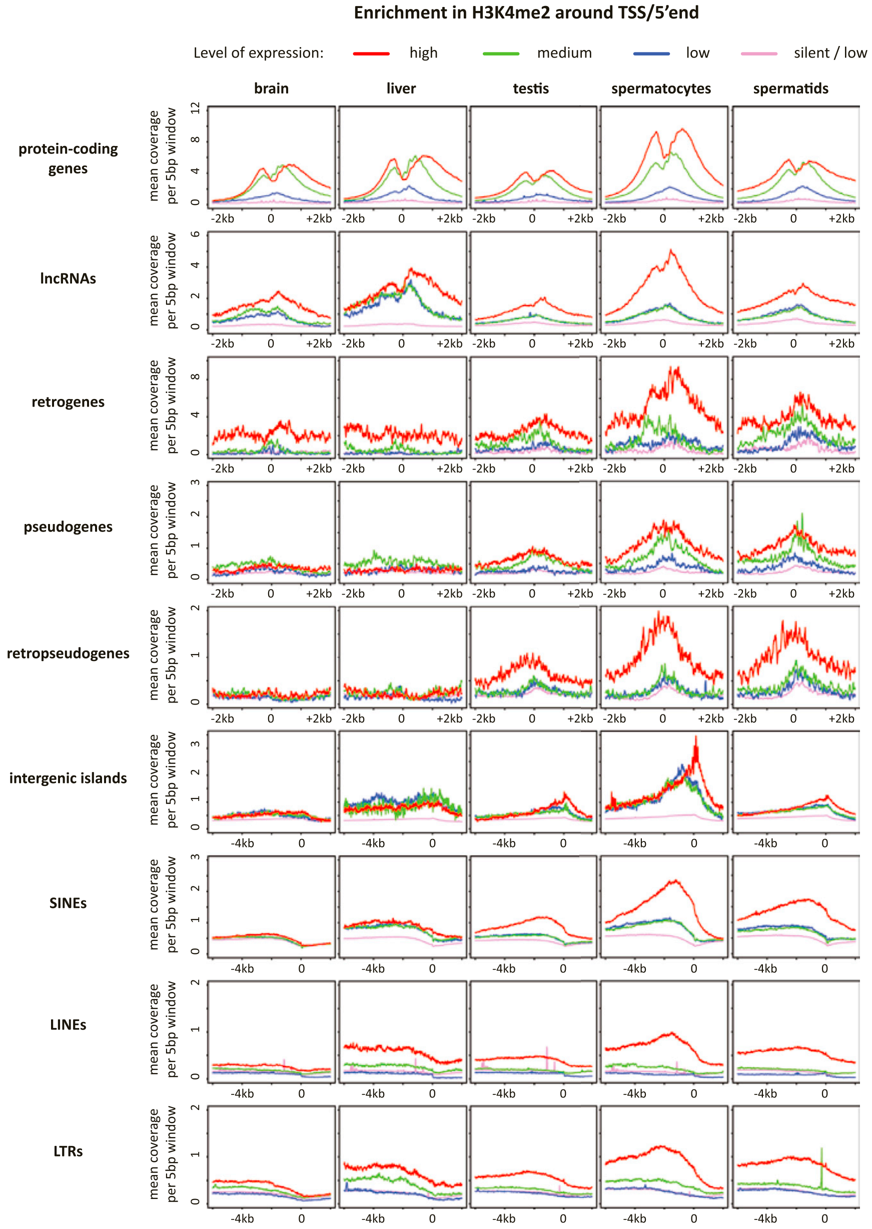

Figure 7. H3K4me2 Levels Upstream of Genomic Elements

Profiles of the H3K4me2 levels (based on ChIP-seq mean read coverage per 5 bp window) near the TSS (protein-coding and lncRNA candidate genes) or annotated 5′ end of autosomal transcribed elements (i.e., position 0) for four distinct subsets of elements based on their expression (high, medium, lower, and low/no expression; see also Experimental Procedures).

nongenic elements such as pseudogenes, transposable elements, and other putatively nonfunctional sequences. Notably, the measured transcriptome diversity in spermatids exceeds that of entire organs (i.e., brain and liver) in our study, although in the case of the brain, even deeper RNA-seq read coverage may be required to capture all transcripts, given that this organ consists of numerous different cell types (Clark et al., 2011). We also note, however, that heterogeneity of sampled spermatocyte and spermatid populations, due to the relatively long differentiation processes that give rise to these cell populations (Figure 2A), may contribute to their high transcriptome complexity. Interestingly, we observed that in contrast to the somatic Sertoli cells, spermatogonia and spermatozoa also have rather complex transcriptomes. The large transcriptome complexity in mature sperm cells is probably explained by residual transcripts from their precursors, elongated spermatids (Johnson et al., 2011), which are difficult to sample using current protocols and thus are not included in this study. Consistent with this notion, our analyses raise the possibility that many spermatozoa transcripts are degraded (Figure S2A), although extensive alternative polyadenylation may explain part of the decreased 3′ representation observed for spermatozoa transcripts (Sendler et al., 2013). It will be interesting to assess the extent to which (intact) spermatozoa transcripts, which could remain stable in the form of ribonucleoprotein particles (Johnson et al., 2011), ultimately contribute to the zygote transcriptome and proteome (in the case of translation of introduced mRNAs), and thus have functional roles in the zygote or early embryogenesis (Johnson et al., 2011).

Our analyses suggest that the high transcriptome complexity of the amniote testis stems largely from extensive transcriptional activity in round spermatids and pachytene spermatocytes. Although pachytene spermatocytes are not as abundant as round spermatids, their contribution is elevated by the large amount of RNA present in these cells. Presumably, nonfunctional transcripts contribute substantially to the high complexity of spermatid and spermatocyte transcriptomes. Moreover, not all transcription of protein-coding genes is necessarily functional in these cell types, given that previous studies of individual genes found that spermatocyte/spermatid mRNAs are often translationally repressed (Kleene, 2001). Nevertheless, we identified functionally relevant expression changes of subsets of protein-coding genes that are part of the complex process of germ cell differentiation.

Several mechanisms likely underlie the widespread transcription in spermatocytes and spermatids. Our analyses suggest that both erasure of DNA methylation at CpG promoters and other more genome-wide mechanisms (as assessed by our H3K4me2 analyses) result in an overall active chromatin state in meiotic and postmeiotic cells. Our results are in agreement with previous observations that the substantial remodeling of chromatin during male germ cell development involves the replacement of standard histones with histone variants (H1t, H1a, H3.3A, and H3.3B) and testis-specific histones (TH2A, TH2B, and TH3) that should favor an open chromatin conformation in these cells (Kimmins and Sassone-Corsi, 2005). Notably, the H3K4me2 patterns presented in this study likely reflect the open chromatin state conferred by H3.3A and H3.3B, because H3.3 variants were shown to be generally enriched with this "active" chromatin modification (Kimmins and Sassone-Corsi, 2005). The recruitment of such histones probably facilitates the substantial and continuous repackaging of DNA during meiosis and spermiogenesis. The particular enrichment of the H3K4me2 in spermatocyte genomes is consistent with the fact that the main replacement of histone H3 by the H3.3A and H3.3B variants takes place during meiosis (Kimmins and Sassone-Corsi, 2005). This may facilitate transcription in spermatocytes through opening of the chromatin during histone replacement, as well as through the generally permissive chromatin after incorporation, which then also affects round spermatids. Together, these observations suggest that the "leaky" transcription of the genome in spermatocytes and spermatids reported here is a secondary, functionally irrelevant consequence of chromatin remodeling.

The widespread genomic transcription in male germ cells has important evolutionary implications. First, our analyses suggest that it facilitated the initial transcription of duplicate protein-coding gene copies in spermatocytes and spermatids. Furthermore, we discovered a striking overabundance of transcribed lncRNA genes in spermatocytes and spermatids, which suggests that the unique chromatin environment in these cells facilitates the origination of new lncRNA genes. Thus, the testis, or rather these specific germ cell stages, indeed seem to represent a "crucible" for the emergence of new genes, as previously hypothesized (Kaessmann, 2010). Second, the extensive nonfunctional transcription in these germ cells suggests that expression levels may be under less evolutionary constraint. Thus, in addition to positive selection (Brawand et al., 2011), the promiscuous transcription of the genome during and after meiosis may have facilitated the rapid divergence of testis transcriptomes during mammalian evolution.

EXPERIMENTAL PROCEDURES

Isolation of Spermatogenic Cells

We used C57BL/6J mice for the preparations of different testis cell populations. Spermatogonia were isolated from the testes of 6-day-old mice by enzymatic dispersion (Kokkinaki et al., 2010). Purity was estimated at ~85% on the basis of immunostaining with an anti-GFRA1 antibody. Pachytene spermatocytes and round spermatids were purified by centrifugal elutriation (Buard et al., 2009). The purity of the round spermatid cell fraction was estimated to be ~90% based on cellular morphology. The purity of the pachytene spermatocyte sample was estimated at ~70% based on fluorescence analysis using anti-SYCP3 (a marker of the synaptonemal complex) and anti-phospho-H2AX (a marker of double-strand breaks and the sex body). Most contamination stemmed from other types of spermatocytes, with only a little contamination from spermatids. To specifically assess the contamination of the spermatocyte sample by spermatids, we compared the RNA-seq transcript levels of five genes specifically expressed in spermatids (three protamine genes [*Prm1*, *Prm2*, and *Prm3*] and two transition protein genes [*Tnp1 and Tnp2*]) between the spermatid and spermatocyte samples, which indicated a contamination of spermatocytes by spermatids of <7% (i.e., the gene-expression levels in the spermatocyte sample were <7% of those in the spermatid sample). Spermatozoa were isolated by dissection of the vas deferens (Stouder et al., 2009) and purity was estimated at ~95% based on cell morphology. Sertoli cells were isolated from 3-week-old animals using Datura stramonium agglutinin (DSA)-coated dishes (Scarpino et al., 1998), with a purity of ~95%.

RNA-Seq and Data Processing

We retrieved amniote RNA-seq data from a previous study (Brawand et al., 2011) and generated additional strand-specific RNA-seq libraries for male mouse brain, liver, testis, and five testis cell types according to the Directional mRNA-seq Library Prep Pre-Release Protocol from Illumina. Each library was sequenced (76 cycles, single end) in three lanes using the Illumina Genome Analyzer IIx platform, yielding a total of ~900 million reads (mean: ~114 million reads per sample) that were processed and mapped based on Ensembl 57 protein-coding gene annotations (Brawand et al., 2011). For specific mouse analyses, mouse genome annotations were refined using deep-coverage RNA-seq data as previously described (Brawand et al., 2011; see Extended Experimental Procedures for the de novo detection of candidate lncRNA genes). Mouse retrocopy coordinates were derived from a previous study (Potrzebowski et al., 2008). Mouse pseudogene coordinates were extracted from Ensembl release 57 (http://www.ensembl.org/). Coordinates of transposable elements (SINEs, LINEs, and LTRs) were retrieved from the UCSC database (http://genome.ucsc.edu/). We defined a set of "nonannotated" intergenic islands by selecting intergenic islands that did not overlap with any of the annotated intergenic elements described above. For all types of genomic elements, expression levels were calculated based on the $\log_2$ mean coverage of RNA-seq reads per base. Expression values were normalized across samples using a median scaling procedure (Brawand et al., 2011).

Detection of Transcribed Elements

A genomic element is defined as being transcribed when its RNA-seq mean base coverage is >0. We built detection saturation plots by quantifying the number of transcribed elements using 10–60 million randomly selected mapped reads. For protein-coding genes and lncRNAs, we modeled detection saturation through the logistic-like function

$$y(x) = \frac{a}{1 + (1/(bx + c))},$$

where *y* is the number of detected genes, and *x* is the number of available reads. We estimated the parameters *a*, *b*, and *c* with a nonlinear least-squares method, implemented in the *stats* package in R. The value *a* represents the theoretical number of genes we could detect given unlimited available reads.

Read Coverage Variation Analysis

We assessed the read coverage variation along the gene length by computing the mean read coverage in 20 nonoverlapping, equal-size transcript (exonic) windows. Genes with <1 kb exonic length were discarded.

Alternative Splicing

We used the mouse high-coverage data to analyze different classes of alternative splicing events (see Extended Experimental Procedures for details).

Principal-Component Analysis

We performed principal-component analyses (PCA) using the dudi.pca function from the ade4 package in R, using as input the matrices of log-transformed normalized expression levels. To exploit all the information present in the data sets, we did not apply any further scaling procedures to the variables.

Segmental Duplicate Regions

C57BL/6J mice segmental duplicon region coordinates were retrieved from a previous study (She et al., 2004; http://mouseparalogy.gs.washington.edu/), and the coordinates were merged to obtain a single set of coordinates encompassing all duplicated regions.

Clustering Analysis

We performed soft-clustering analysis using the Mfuzz software package (Kumar and Futschik, 2007), and GO term enrichment analysis using the FatiGO module of the Babelomics web-based tools (http://bioinfo.cipf.es/babelomicswiki/tool:fatigo). PhastCons scores were extracted from the UCSC table browser. Mouse SNP data were downloaded from ftp://ftp.sanger.ac.uk/pub/mouse_genomes/current_snps/ (Keane et al., 2011).

RNA Quantitation

We extracted RNA and DNA for nine mouse samples simultaneously using the QIAGEN AllPrep DNA/RNA Mini Kit for at least three biological replicates and at least five technical extraction replicates per tissue. We used the amount of extracted DNA to define the number of cells from which DNA/RNA was extracted based on the amount of DNA per diploid/haploid genome-equivalent, and then estimated the amount of RNA produced per cell for each of these tissues based on the amount of RNA extracted.

Reduced Representation Bisulfite Sequencing and Analysis

Two sets of DNA methylation bisulfite sequencing libraries (with different insert sizes) were generated for each of the studied tissues (same samples as used for RNA-seq) according to a previously described protocol (Smith et al., 2009). These libraries were sequenced (38 and 76 cycles, respectively) on the Illumina Genome Analyzer IIx platform, yielding a total of ~322 million reads (mean of ~64 million reads per sample). Reduced representation bisulfite sequencing (RRBS) reads were mapped on the mouse genome with the use of the BSMAP mapping tool (Xi and Li, 2009). The percentages of DNA methylation levels based on bisulfite conversation yield were computed at the single-nucleotide scale for positions represented by at least 10× coverage of uniquely mapped reads. We were able to establish the DNA methylation state for 7,492,706 C positions, including 1,519,053 C positions from CG dinucleotides. DNA methylation levels for promoter/upstream regions of the different genomic elements were computed as a mean of percentages of DNA methylation levels for all CG dinucleotides for which data were available in these regions.

ChIP-Seq

ChIP with antibodies against H3K4me2 was carried out (for the same samples as used for RNA-seq and RRBS) as previously described (Bernstein et al., 2005). Libraries were prepared according to the Illumina protocol for ChIP-seq and sequenced (38 cycles) using the Illumina Genome Analyzer IIx platform (total number of reads: ~131 million reads; mean: ~26 million reads per sample). Reads were mapped with the use of Bowtie (Langmead et al., 2009). We determined the base coverage for each sample by restricting our data to 16 million randomly resampled uniquely mapped reads. For the assessment of H3K4me2 levels, elements of a given type were separated into four sets of the same size on the basis of their expression levels. The mean coverage of H3K4me2 enrichment was then calculated for nonoverlapping 5 bp windows for specific regions upstream of genomic elements. In order to assess enrichment patterns for the different genomic elements in a given tissue, we first defined putative enrichment regions (promoter regions/regions around the TSS: −1 kb to +1 kb relative to the TSS for all genomic elements except transposable elements; −3 kb to −1 kb relative to TSS for transposable elements) and genomic background regions (−4 kb to −2 kb relative to TSS for all genomic elements except transposable elements; −6 kb to −4 kb relative to TSS for transposable elements). We then computed the mean of the maximum H3K4me2 level across all transcribed elements (or elements belonging to the highest-expression class in the case of intergenic islands and transposable elements) of a given type in a given tissue for these regions. The fold enrichment for each element and tissue was then calculated as the ratio of the means between the putative promoter and genomic background region.

ACCESSION NUMBERS

Sequencing data have been deposited in the Gene Expression Omnibus with accession numbers GSE43717, GSE43719, and GSE43721.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Discussion, Extended Experimental Procedures, four figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.05.031.

LICENSING INFORMATION

ACKNOWLEDGMENTS

We thank K. Harshman and the staff of the Lausanne Genomics Technologies Facility for high-throughput sequencing support; I. Xenarios and the staff of the Vital-IT computational facility for computational support; E. Ait Yahya Graison for helpful discussion and advice on the soft-clustering analysis; D. Cortez, K. Guschanski, J. Meunier, and M. Warnefors for comments on the manuscript; and the Kaessmann group in general for helpful discussions. This research was supported by grants from the European Research Council (Starting Independent Grant 242597, SexGenTransEvolution) and the Swiss National Science Foundation (grant 130287) to H.K.; grants from the Centre National de la Recherche Scientifique (CNRS) and the Association pour la Recherche sur le Cancer and the Agence Nationale de la Recherche (ANR-09-BLAN-0269-01) to B.d.M.; and grants from the Swiss National Science Foundation (31003A_135227) and the NCCR Frontiers in Genetics program to S.N. A.N. was supported by a long-term FEBS postdoctoral fellowship. P.B. was supported by a PhD fellowship from the French Ministry MENRT.

Received: December 19, 2012
Revised: April 17, 2013
Accepted: May 21, 2013
Published: June 20, 2013

REFERENCES

Arnheim, N., and Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. Nat. Rev. Genet. *10*, 478–488.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169–181.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343–348.

Buard, J., Barthès, P., Grey, C., and de Massy, B. (2009). Distinct histone modifications define initiation and repair of meiotic recombination in the mouse. EMBO J. *28*, 2616–2624.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927.

Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A., et al. (2011). The reality of pervasive transcription. PLoS Biol. *9*, e1000625, discussion e1001102.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Grabske, R.J., Lake, S., Gledhill, B.L., and Meistrich, M.L. (1975). Centrifugal elutriation: separation of spermatogenic cells on the basis of sedimentation velocity. J. Cell. Physiol. *86*, 177–189.

Johnson, G.D., Lalancette, C., Linnemann, A.K., Leduc, F., Boissonneault, G., and Krawetz, S.A. (2011). The sperm nucleus: chromatin, RNA, and the nuclear matrix. Reproduction *141*, 21–36.

Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., Daish, T., Grützner, F., and Kaessmann, H. (2012). Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol. *10*, e1001328.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res. *20*, 1313–1326.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. *10*, 19–31.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

Kimmins, S., and Sassone-Corsi, P. (2005). Chromatin remodelling and epigenetic features of germ cells. Nature *434*, 583–589.

Kleene, K.C. (2001). A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech. Dev. *106*, 3–23.

Kokkinaki, M., Lee, T.L., He, Z., Jiang, J., Golestaneh, N., Hofmann, M.C., Chan, W.Y., and Dym, M. (2010). Age affects gene expression in mouse spermatogonial stem/progenitor cells. Reproduction *139*, 1011–1020.

Kumar, L., and Futschik, M.E. (2007). Mfuzz: a software package for soft clustering of microarray data. Bioinformation *2*, 5–7.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Mueller, J.L., Mahadevaiah, S.K., Park, P.J., Warburton, P.E., Page, D.C., and Turner, J.M. (2008). The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression. Nat. Genet. *40*, 794–799.

Ponting, C.P., and Belgard, T.G. (2010). Transcribed dark matter: meaning or myth? Hum. Mol. Genet. *19*(R2), R162–R168.

Potrzebowski, L., Vinckenbosch, N., Marques, A.C., Chalmel, F., Jégou, B., and Kaessmann, H. (2008). Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol. *6*, e80.

Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput. Biol. *5*, e1000598.

Scarpino, S., Morena, A.R., Petersen, C., Fröysa, B., Söder, O., and Boitani, C. (1998). A rapid method of Sertoli cell isolation by DSA lectin, allowing mitotic analyses. Mol. Cell. Endocrinol. *146*, 121–127.

Schmidt, E.E., and Schibler, U. (1995). Cell size regulation, a mechanism that controls cellular RNA accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-Y and the liver-enriched transcription factor DBP. J. Cell Biol. *128*, 467–483.

Schultz, N., Hamra, F.K., and Garbers, D.L. (2003). A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. Proc. Natl. Acad. Sci. USA *100*, 12201–12206.

Sendler, E., Johnson, G.D., Mao, S., Goodrich, R.J., Diamond, M.P., Hauser, R., and Krawetz, S.A. (2013). Stability, delivery and functions of human sperm RNAs at fertilization. Nucleic Acids Res. *41*, 4104–4117.

She, X., Horvath, J.E., Jiang, Z., Liu, G., Furey, T.S., Christ, L., Clark, R., Graves, T., Gulden, C.L., Alkan, C., et al. (2004). The structure and evolution of centromeric transition regions within the human genome. Nature *430*, 857–864.

She, X., Cheng, Z., Zöllner, S., Church, D.M., and Eichler, E.E. (2008). Mouse segmental duplication and copy number variation. Nat. Genet. *40*, 909–914.

Shima, J.E., McLean, D.J., McCarrey, J.R., and Griswold, M.D. (2004). The murine testicular transcriptome: characterizing gene expression in the testis during the progression of spermatogenesis. Biol. Reprod. *71*, 319–330.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285.

Smith, Z.D., Gu, H., Bock, C., Gnirke, A., and Meissner, A. (2009). High-throughput bisulfite sequencing in mammalian genomes. Methods *48*, 226–232.

Stouder, C., Deutsch, S., and Paoloni-Giacobino, A. (2009). Superovulation in mice alters the methylation pattern of imprinted genes in the sperm of the offspring. Reprod. Toxicol. *28*, 536–541.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. USA *101*, 6062–6067.

Turner, J.M. (2007). Meiotic sex chromosome inactivation. Development *134*, 1823–1831.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial analysis of gene expression. Science *270*, 484–487.

Wang, P.J., McCarrey, J.R., Yang, F., and Page, D.C. (2001). An abundance of X-linked genes expressed in spermatogonia. Nat. Genet. *27*, 422–426.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232.

Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A., Mitsakakis, N., Mohammad, N., Robinson, M.D., Zirngibl, R., Somogyi, E., et al. (2004). The functional landscape of mouse gene expression. J. Biol. *3*, 21.

Zhang, Y.E., Vibranovski, M.D., Landback, P., Marais, G.A., and Long, M. (2010). Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS Biol. *8*, 8.

Supplemental Information

EXTENDED DISCUSSION

Our soft-clustering analysis revealed four distinct clusters of groups of autosomal genes that show specific expression level profiles and are significantly enriched with respect to specific biological processes, molecular functions and cellular localizations of encoded proteins (Figure 5A). The first cluster (2,764 genes) is strongly upregulated in spermatocytes and spermatids relative to spermatogonia and spermatozoa, and is significantly enriched with genes involved in spermatogenesis, gamete generation, sperm motility, fertilization and similar processes (Benjamini-Hochberg corrected $p < 0.05$, Fisher's exact test). Consistently, an excess of genes in this cluster produce proteins localizing to the flagellum and mitochondria, both essential for sperm motility (Table S2).

The 2,347 genes in the second cluster, which are downregulated in spermatocytes/spermatids and subsequently strongly upregulated in spermatozoa, may underlie the dramatic cellular changes during spermatogenesis, in particular those occurring during spermiogenesis (i.e., the formation of spermatozoa from round spermatids via elongated spermatids). The most significant overrepresented biological processes in this cluster are related to apoptosis, consistent with the important role of cell death processes during spermatogenesis (Shaha et al., 2010). The overrepresentation of biological processes and molecular functions related to signal transduction and the overrepresentation of gene products localizing to the Golgi and actin cytoskeleton (Table S2) are likely reflecting the profound changes in cellular morphology occurring during spermiogenesis. For example, the major restructuring of the Golgi apparatus during spermiogenesis (the Golgi is ultimately transformed into the sperm acrosome) requires specific activities of protein kinases localizing to the Golgi and reorganization of the actin cytoskeleton (Sun et al., 2011).

Overrepresented GO categories in the third cluster (2,347 genes), which shows strong upregulation in spermatids, are mainly related to sensory perception (Table S2). This observation may seem surprising but is consistent with previous observations that a subset of olfactory receptor gene expression family members are expressed in spermatids (e.g., Parmentier et al., 1992) and the hypothesis that similar signal transduction systems might be used in olfaction and spermiogenesis (Gautier-Courteille et al., 1998).

Finally, genes in the fourth cluster (697 genes), which are predominantly functioning in protein translation (Table S2), show strongly reduced expression levels during spermiogenesis. This observation is consistent with the generally accepted view that the process of protein translation is shut down during late spermiogenesis (spermatozoa are thought to be largely transcriptionally silent; (Johnson et al., 2011)). Overall, our results seem to capture major aspects of the sophisticated transcriptional program that underlies the complex process of spermatogenic cell differentiation.

EXTENDED EXPERIMENTAL PROCEDURES

Detection of Candidate LncRNA Genes

We used our RNA-Seq data to detect candidate lncRNA genes de novo. To do this, we aligned the reads on the genome and detected splice junctions using TopHat (Trapnell et al., 2009). We then defined transcribed islands (i.e., contiguous genomic regions with nonnull read coverage) and determined the set of clusters of transcribed islands that are connected through splice junctions. These clusters of transcribed islands represent our set of candidate multi-exonic transcribed loci. For the mouse high-coverage RNA-Seq data, which is strand-specific, this detection was done independently on each strand; for the nonstrand specific RNA-Seq data, the strand of the locus was defined based on the orientation of the splice junctions. We considered only junctions with GT-AG and GT-AC splice sites, for which the strand can be reliably inferred. We next selected the multi-exonic transcribed loci that are found in regions annotated as intergenic in the Ensembl database (release 57); these intergenic multi-exonic transcribed loci were used as a basis for further analyses.

We first applied the procedure described above for the multiple species analyses based on a previously established RNA-seq data set (Brawand et al., 2011). In the multiple species analyses, to increase our sensitivity and to minimize prediction inequalities among species (which can be due to differences in RNA-seq read coverage and in existing genome annotations), we projected the coordinates of multi-exonic transcribed loci of each species on the genomes of the other species (A.N., M.S., A. Liechti, T. Daish, J.C. Baker, F. Grützner, and H.K., unpublished data). For the mouse, we separately predicted the multi-exonic transcribed loci based on the strand-specific deep coverage data for this species. Thus, the numbers of putative candidate lncRNA loci are different in the multiple species and dedicated mouse analyses.

To determine whether these multi-exonic transcribed loci are likely protein-coding or noncoding, we applied the codon substitution frequency (CSF) metrics (Lin et al., 2008), which was previously used to discriminate between noncoding and protein-coding genes in the mouse (Guttman et al., 2009). This method compares the codon substitution pattern in a candidate region with that of known coding and noncoding sequences. The substitution pattern was computed based on a multiple genome alignment of 46 vertebrate species, downloaded from the UCSC Genome Browser (Karolchik et al., 2011). The CSF method was trained on as set of Ensembl-annotated protein-coding sequences, and on a set of 10,000 random intergenic regions, used as a noncoding standard, for each species. The CSF score cut-off was chosen in order to have a false discovery rate of 10%, i.e., at most 10% of protein-coding exons can be classified as noncoding with this cut-off. In addition to the CSF metrics, we also searched for sequence similarities between our candidate regions and known protein-coding genes. To do this, we used blastx to align the nucleotide sequences of our candidate genes with the Ensembl-annotated protein sequences of 16 vertebrate species (human, chimpanzee, gorilla, orangutan, macaque, mouse, rat, guinea pig, cat, dog, pig, opossum, platypus, chicken, *Xenopus*, zebra fish). We considered that a sequence has significant similarity with a known protein sequence if the blastx e-value was below 0.01. For a multi-exonic transcribed locus to be considered as a candidate lncRNA gene, it had to be classified as noncoding with both the CSF and the blastx methods.

Alternative Splicing

We used the mouse high-coverage RNA-Seq data to assess four classes of alternative splicing events: skipped (or cassette) exons, alternative 5′ and 3′ splice sites and retained introns. We used TopHat (Trapnell et al., 2009) version 1.0.13 to align the RNA-Seq reads on the genome and to detect splice junctions de novo, without relying on the existing genomic annotations. For the analysis of alternative splicing patterns, we considered only unambiguously mapped reads, extracted from the 60 million mapped reads resampled for each tissue. We detected skipped exons and alternative splice sites based on the positions of the splice junctions within genes and exons, as previously described (Wang et al., 2008). We inferred retention events for introns whose exact positions were confirmed with our RNA-seq data, for which at least 50% of the intron length was covered by reads, and for which both exon-intron boundaries were also spanned by reads (this definition is similar to the one used by Wang et al., but is more stringent).

To determine whether there are global differences in alternative splicing frequencies among tissues and cell types, we estimated for each sample the proportion of genes that had at least one alternative splicing event of a given class. Given that the distribution of gene expression levels differs significantly among tissues and cell types, we took into account the dependence between the mean read coverage of a gene and the probability of detection of alternative splicing events. To do this, we ordered the genes according to their expression level, and divided them into 15 equal size classes, independently for each sample. We then computed the proportion of genes with detected alternative splicing events in each expression class, and analyzed its relationship with the mean expression level in each class. This relationship can be modeled mathematically with the logistic-like function

$$y(x) = \frac{1-a}{1+\exp(-bx+c)},$$

where y is the proportion of alternatively spliced genes and x is the log-transformed mean read coverage of the genes. We estimated the parameters a, b and c with a nonlinear least-squares method, implemented in the stats package in R. The value (1-a) represents the theoretical proportion of alternatively spliced genes that would be observed given unlimited read coverage.

We developed a randomization procedure that allows us to compare the frequency of occurrence of alternative splicing events between samples, while taking into account the differences in expression level distributions between samples. For the comparison between two given samples, we determined the 10% - 90% quantile interval of the mean read coverage distribution of each of the samples, and selected those genes which are found in the intersection of the two intervals, for both samples. We divided the range of expression levels of the selected genes into ten equal-width intervals, and selected randomly a fixed number of genes in each interval, for each sample. We then computed the proportion of genes for which we could detect alternative splicing events, for each randomized sample. This procedure was repeated 1000 times. We derived a one-tailed p value from the comparisons of the 1000 simulations, for each comparison between two samples.

SUPPLEMENTAL REFERENCES

Gautier-Courteille, C., Salanova, M., and Conti, M. (1998). The olfactory adenylyl cyclase III is expressed in rat germ cells during spermiogenesis. Endocrinology *139*, 2588–2599.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Karolchik, D., Hinrichs, A.S., and Kent, W.J. (2011). The UCSC Genome Browser. Curr. Protoc. Hum. Genet. *Chapter 18*, Unit 18.16.

Lin, M.F., Deoras, A.N., Rasmussen, M.D., and Kellis, M. (2008). Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. PLoS Comput. Biol. *4*, e1000067.

Parmentier, M., Libert, F., Schurmans, S., Schiffmann, S., Lefort, A., Eggerickx, D., Ledent, C., Mollereau, C., Gérard, C., Perret, J., et al. (1992). Expression of members of the putative olfactory receptor gene family in mammalian germ cells. Nature *355*, 453–455.

Shaha, C., Tripathi, R., and Mishra, D.P. (2010). Male germ cell apoptosis: regulation and biology. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 1501–1515.

Sun, X., Kovacs, T., Hu, Y.J., and Yang, W.X. (2011). The role of actin and myosin during spermatogenesis. Mol. Biol. Rep. *38*, 3993–4001.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470–476.

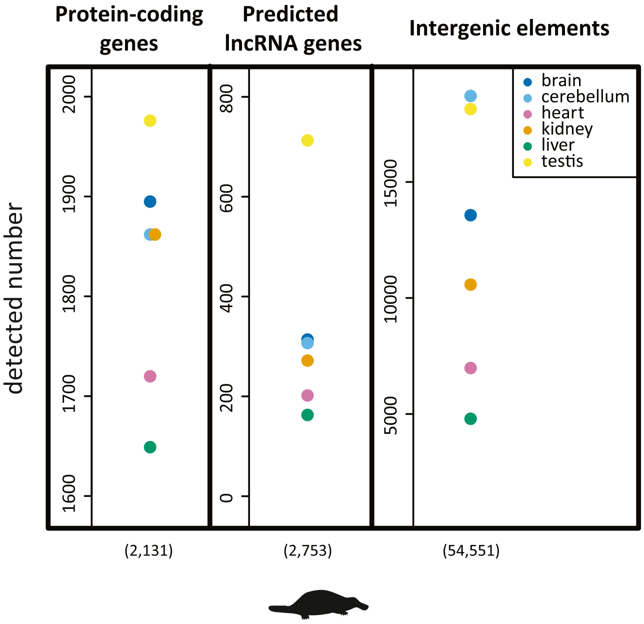

Figure S1. Transcriptome Complexity in Platypus, Related to Figure 1

Number of autosomal transcribed protein-coding genes, predicted long noncoding RNA (lncRNA) genes, and transcribed intergenic elements (pseudogenes, transposable elements, other intergenic sequences) in six platypus organs, based on eight million of randomly selected RNA-seq reads per sample (total number of elements of each type is indicated in brackets). Note: the substantially smaller total numbers of elements for platypus compared to the other species (Figure 1) is due to the small amount of genomic sequence specifically assigned to autosomes for this genome.

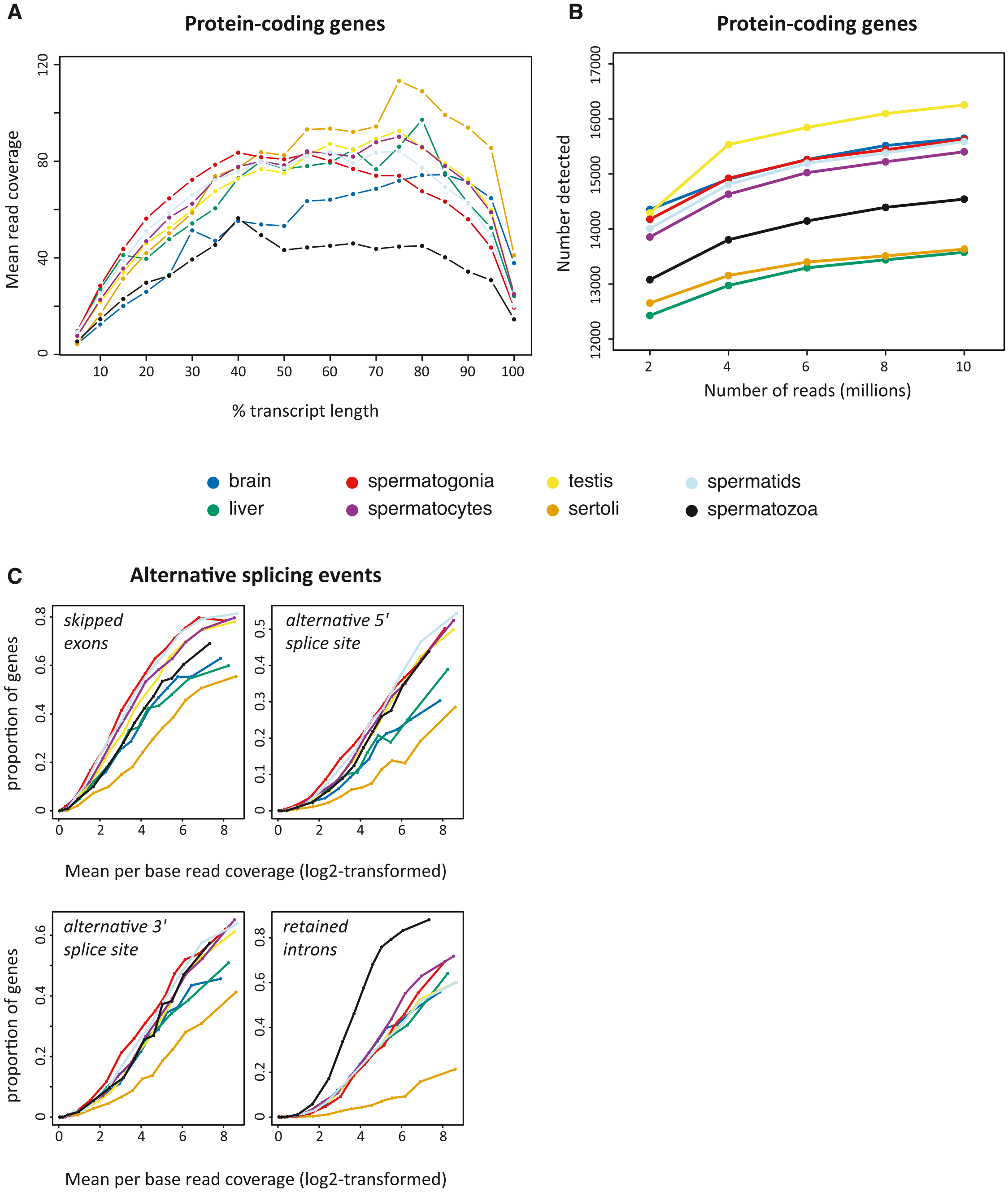

Figure S2. Read Coverage Variation, Read Resampling, and Alternative Splicing Patterns, Related to Figure 2

(A) Mean of read coverage along autosomal protein-coding gene transcripts (longer than 1kb), computed on 20 same-sized windows for each of the 8 samples.

(B) Number of autosomal protein-coding genes for which transcripts were detected in two somatic tissues, total testis, and five testicular cell types, on the basis of 2 to 10 million of randomly selected mapped reads excluding reads mapping on either the top 1,000 most highly expressed genes or on ribosomal genes.

(C) Proportion of protein-coding genes affected by alternative splicing events (skipped exons, 5′/3′ alternative splice sites and retained introns) for 15 classes of genes given their level of expression (Experimental Procedures). These data were used to obtain complete theoretical distributions of frequencies of genes with alternative splicing events for the different tissues.

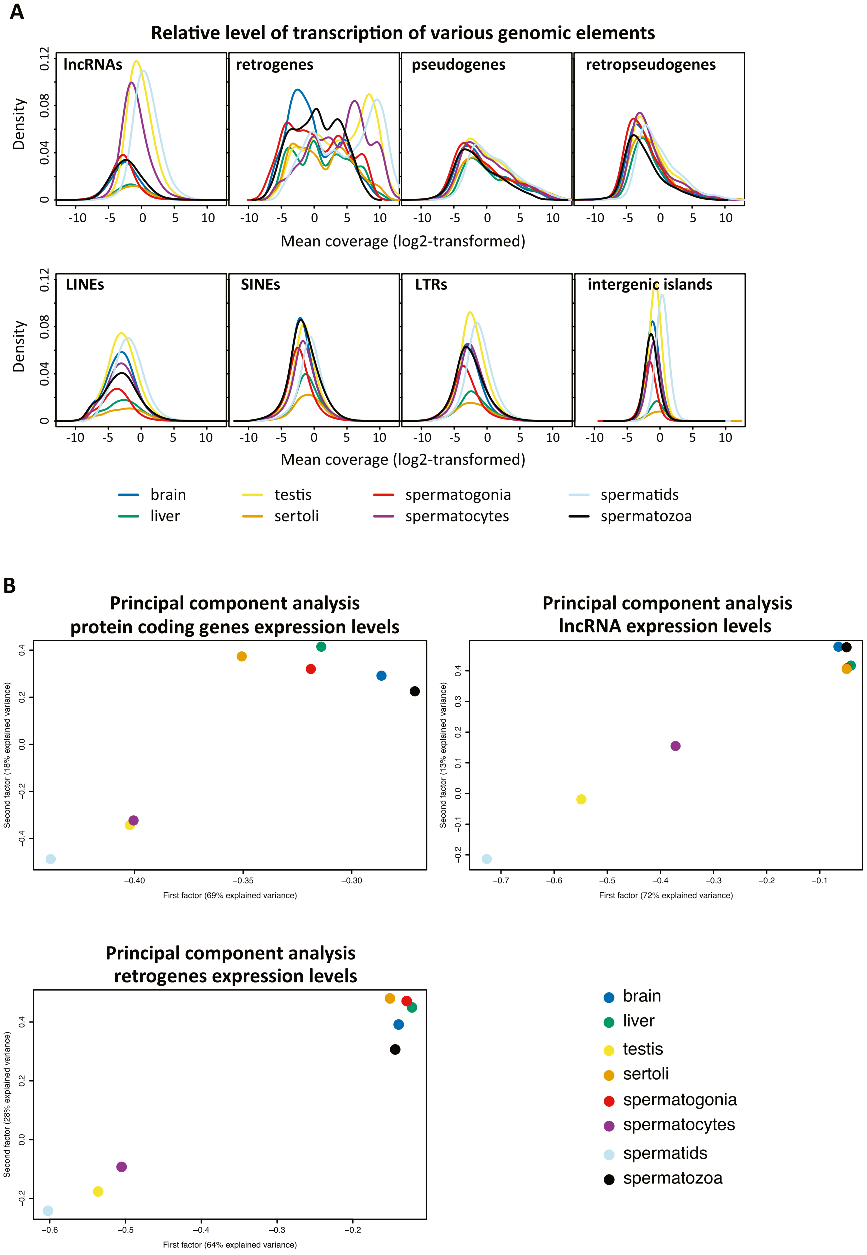

Figure S3. Relative Transcript Levels and PCA, Related to Figure 3

(A) Density plots of the total per-base read coverage for the different types of autosomal genomic elements in the different tissues (plots grouped by genomic elements; see Figure 3 for the same data grouped by tissues).

(B) Principal Component Analysis performed for protein-coding genes, lincRNAs, and retrogenes.

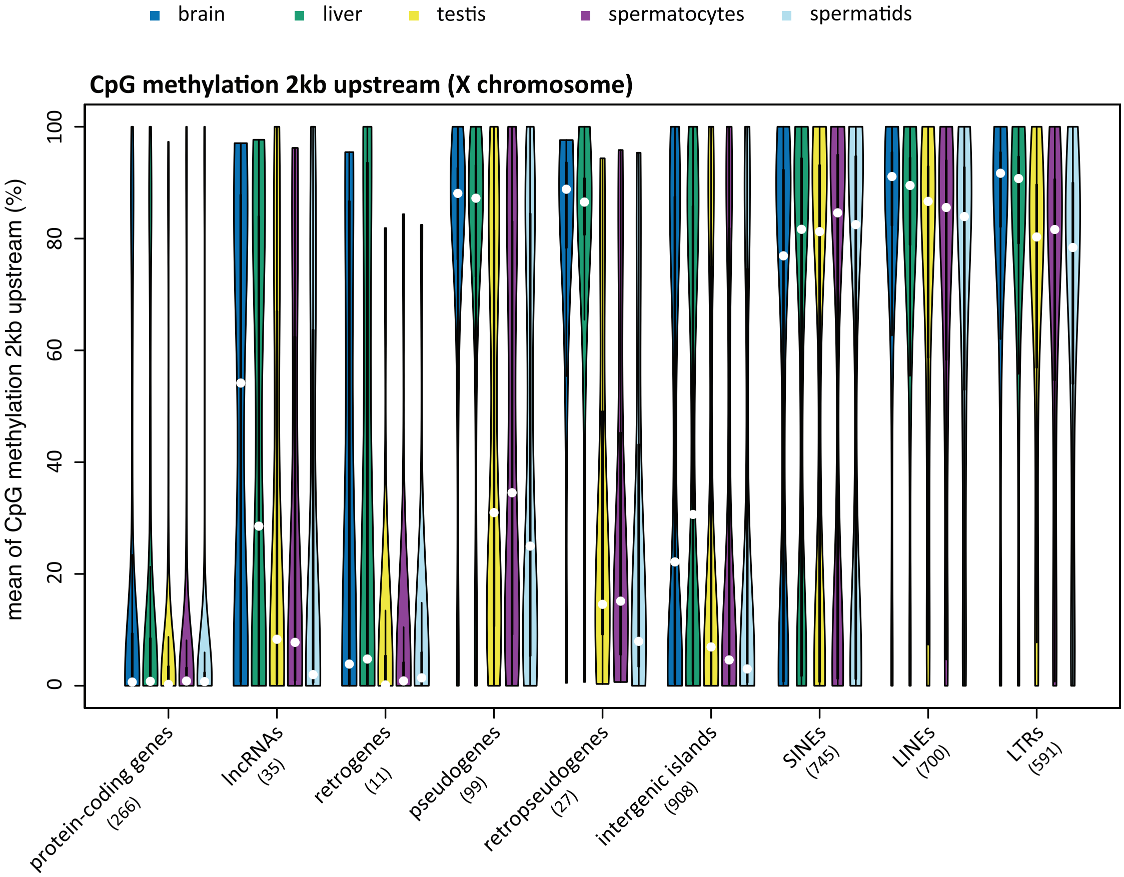

Figure S4. DNA Methylation Upstream of Genomic Elements, Related to Figure 6.

Violin plot showing the distribution of CpG methylation levels upstream of genomic elements on the X chromosome (mean percentages of CpG methylation levels 2kb upstream). The white point represents the median. See Figure 6 for autosomal patterns.