Widespread transcriptional scanning in testes modulates gene evolution rates

Bo Xia[1], Maayan Baron[1], Yun Yan[1], Florian Wagner[1], Sang Y. Kim[5], David L. Keefe[3], Joseph P. Alukal[3], Jef D. Boeke[2,4] and Itai Yanai[1,2]*

[1] Institute for Computational Medicine, New York University Langone Health, New York, NY 10016, USA

[2] Department of Biochemistry and Molecular Pharmacology, New York University Langone Health, New York, NY 10016, USA

[3] Department of Obstetrics and Gynecology, New York University Langone Health, New York, NY 10016, USA

[4] Institute for Systems Genetics, New York University Langone Health, New York, NY 10016, USA

[5] Department of Pathology, New York University Langone Health, New York, NY 10016, USA

*Correspondence to: Itai.Yanai@nyumc.org

Abstract:

A long-standing question in molecular biology relates to why the testes express the largest number of genes relative to all other organs. Here, we report a detailed gene expression map of human spermatogenesis using single-cell RNA-Seq. Surprisingly, we found that spermatogenesis-expressed genes contain significantly fewer germline mutations than unexpressed genes, with the lowest mutation rates on the transcribed DNA strands. These results suggest a model of 'transcriptional scanning' to reduce germline mutations by correcting DNA damage. This model also explains the rapid evolution in sensory- and immune-defense related genes, as well as in male reproduction genes. Collectively, our results indicate that widespread expression in the testes achieves a dual mechanism for maintaining the DNA integrity of most genes, while selectively promoting variation of other genes.

Main Text:

Human tissues and organs are distinguished by the genes that they express and those that they do not [1,2]. Tissues have transcriptomes of different complexities in terms of uniquely-expressed genes, as well as those genes expressed at differential levels [3–6]. One overarching goal in the life sciences is to characterize the specific transcriptomic signatures of all human tissues, and ultimately each different cell type at the single-cell level [7].

In males, the testis is unique in comparison with somatic tissues in that it contains germ cells which pass the genetic information on to the next generation [8]. Interestingly, it has been known for many years that the testis stands out as having the most complex transcriptome with the highest number of expressed genes [9–12]. Widespread transcription in the testes has been reported to account for an amazing expression of over 80% of all our protein-coding genes [10,11,13], as well as across many other mammals [3,10].

Several hypotheses have been proposed to explain this observation. Widespread expression may represent a functional requirement for the gene-products in question [12]. However, other more complex organs such as the brain do not exhibit a corresponding number of expressed genes despite the fact that they consist of a substantially greater number of distinct cell types [3,10,14–16]. Moreover, recent animal studies have shown that many testis-enriched and evolutionarily-conserved genes are not required for male fertility in mice [17]. A second hypothesis implicates leaky transcription during the massive chromatin remodeling that occurs throughout spermatogenesis [12,18,19]. However, this model predicts more expression during later stages of spermatogenesis – when the genome is undergoing the most chromatin changes – contradicting the observation [13,18]. Additionally, the energetic requirements for the observed widespread expression are sufficiently large that such leaky expression would be expected to be under tighter

control [20]. Given this lack of a compelling explanation for widespread testes transcription, the topic remains an interesting and yet unanswered question.

Here we propose a model that widespread testis transcription modulates gene evolution rates. Beyond functional requirements for reproduction, widespread transcription acts as a scanning mechanism through the majority of human genes, detecting and repairing bulky DNA damage events through transcription-coupled repair (TCR) [21,22], which ultimately reduces germline mutations rates and gene evolution rates. Genes that are not expressed in the male germline do not benefit from the reduced mutation rates. These genes do not constitute a random set but rather are enriched in sensory and defense-immune system genes, accounting for previous observations that these genes evolve faster [23,24]. We also found that transcription-coupled damage (TCD) overwhelms this pattern in the very highly expressed genes, which are enriched in spermatogenesis-related functions, implicating TCD-modulated gene evolution. By understanding the uneven germline mutation patterns and the intrinsic mechanism of germline DNA damage removal, we will be in a better position to understand human genome evolution and genetic diseases [25].

Single-cell RNA-Seq reveals the developmental trajectory of spermatogenesis.

The developmental process of spermatogenesis includes mitotic amplification, meiotic specification to generate haploid germ cells, and finally differentiation and morphological transition to mature sperm cells (Fig. 1A) [26]. Technical limitations confined previous gene expression analyses of spermatogenesis to its broad stages: spermatogonia, spermatocytes, round spermatids and spermatozoa [10,13]. To systematically characterize the detailed transcriptomic

signatures throughout the entirety of spermatogenesis, we applied high-throughput single-cell RNA-Seq to the human testes (Fig. S1A) [27].

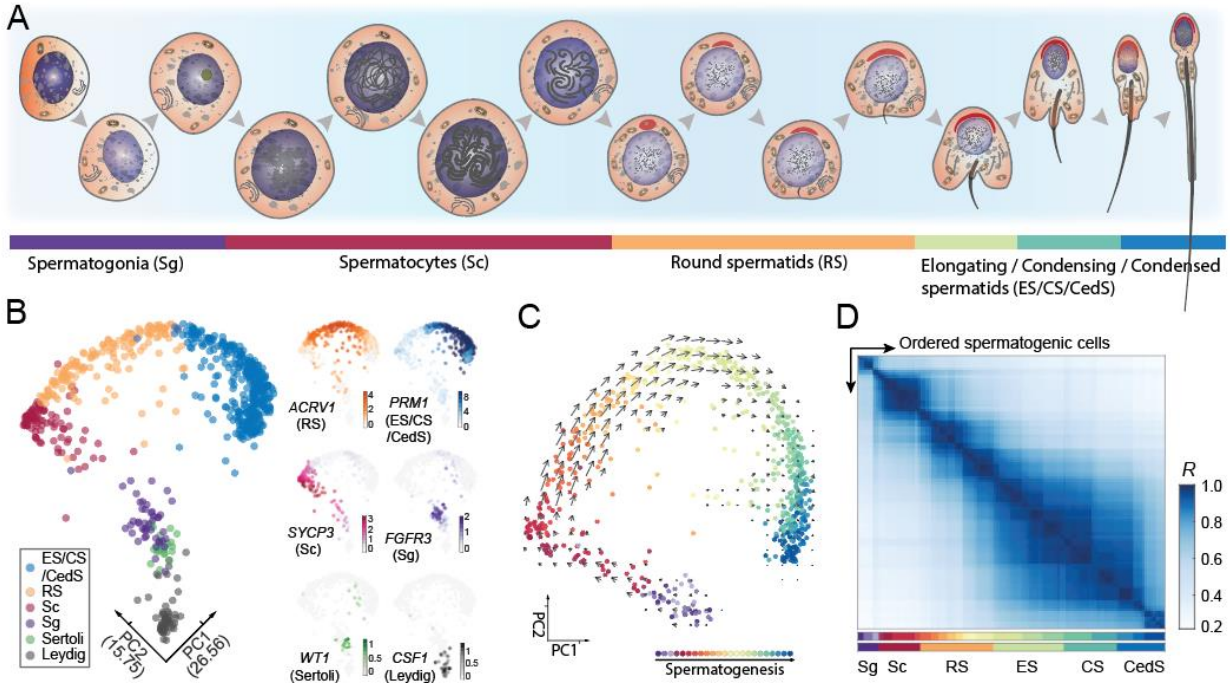

Fig. 1. Single-cell RNA-Seq (scRNA-Seq) reveals a detailed molecular map of human spermatogenesis. (A) Developmental stages of human spermatogenesis. **(B)** Principal components analysis of testis scRNA-Seq data. Colors indicate the main spermatogenic and somatic cell types, as defined by marker genes (insets). **(C)** Principal components analysis on the spermatogenic-complement of the single-cell data. Arrows indicate the developmental trajectory as inferred from the relationship between the spliced and unspliced transcriptomes [28] (SI methods). **(D)** Heatmap of the correlation coefficients between single-cell spermatogenesis transcriptomes.

A principal component analysis (PCA) revealed clusters of cells including a large continuous cluster (Fig. 1B). Using previously determined stage markers to infer the identity of the cells, we annotated the main spermatogenic stages, as well as the somatic Leydig and Sertoli cells (Fig. 1B, right). Excluding the somatic cells, PCA on the germ cells revealed a horseshoe-shaped cluster suggesting that the order of the cells corresponds to developmental time (Fig. 1C, S1C, SI methods). Three independent lines of evidence support this projection. First, the order of expression of known marker genes across the horseshoe-shaped cluster matches their developmental order (Fig. S1C). Second, the Monocle2 algorithm which identifies developmental trajectories also revealed the same order of cells (Fig, S1D-E) [29]. Finally, using the pattern of unspliced versus spliced transcripts across the cluster as a means to predict the developmental trajectory [28] also reinforced this interpretation (Fig, 1C and SI methods). The arrows in Figure 1C relate the unspliced transcriptome of cells with the spliced transcriptome of other cells, allowing the inference of developmental time. From these lines of evidence, we concluded that the germ cell transcriptomes could be ordered as successive stages throughout spermatogenesis. This detailed delineation of spermatogenic stages provides stage-specific marker-gene expression with unprecedented resolution of molecular signatures of spermatogenesis (Figs. 1D and S2).

TCR-induced reduction of germline mutation rates

We hypothesized that the widespread transcription in spermatogenesis may lead to two scenarios (Fig. 2A): 1) open chromatin in transcribed regions leads to a higher mutagenic likelihood by transcription-coupled damage (TCD) [30], and consequently to higher germline mutation rates and divergence across species; and/or 2) the transcribed regions are subject to transcription-coupled

repair (TCR) of the DNA [21], thus reducing germline mutation rates and safeguarding the germline genome, leading to lower divergence across species. To study these hypotheses, we first utilized our single-cell RNA-Seq data and assigned a spermatogenic stage to each gene according to its period of maximal expression (Fig. 2B, SI methods). Overall, we detected the expression of 87% of all protein-coding genes in one or more stages throughout spermatogenesis (Fig. 2B), consistent with previous observations [10,13].

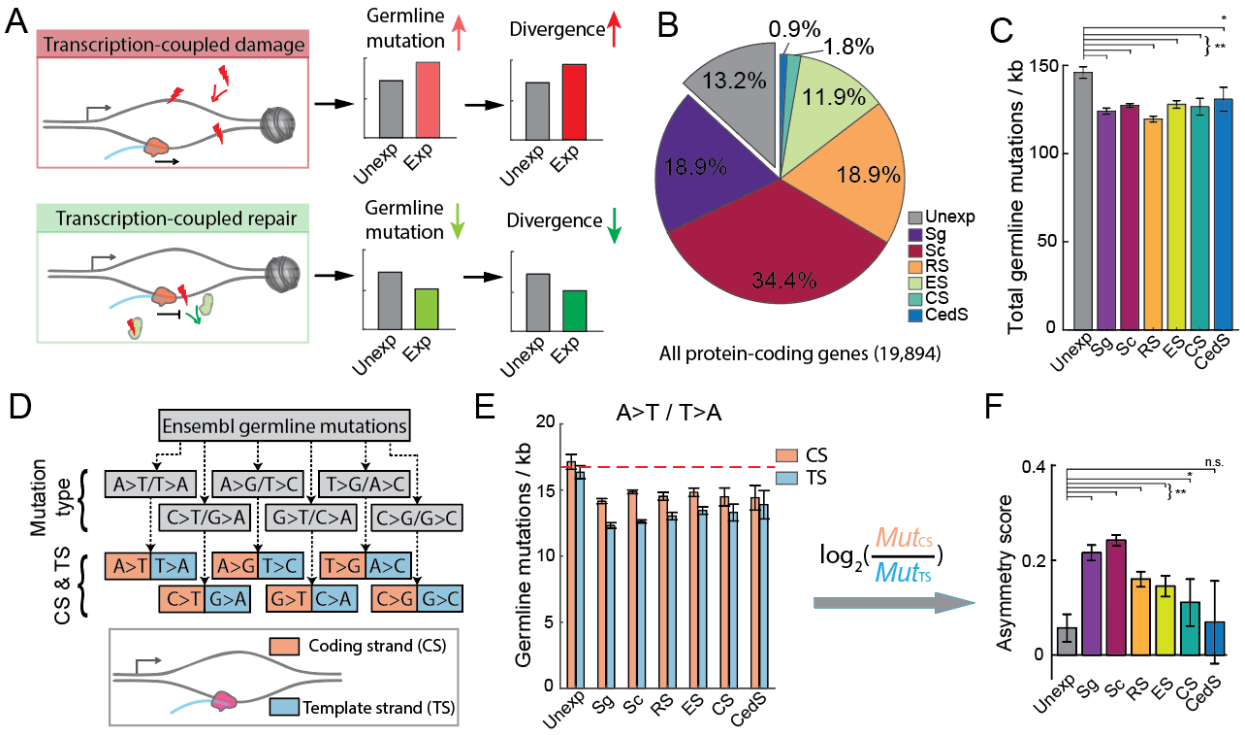

Fig. 2. Widespread transcription in spermatogenic cells is associated with reduced germline mutation rates. (A) Two possible consequences of widespread transcription in spermatogenic cells. **(B)** Pie chart indicating the number of genes expressed at each spermatogenic stage. Genes are associated with the stage in which they are maximally expressed (SI methods). **(C)** Total germline mutation rates across the gene categories of spermatogenesis stages. **(D)** Germline

mutations associated with genes were retrieved from Ensembl [31] and classified into the six mutation classes, which were further distinguished in terms of coding and template strands, as previously introduced [32]. **(E)** A>T transversion mutation rates for the coding and the template strands for the spermatogenic gene categories. Dashed lines indicate the average level of mutations in the unexpressed genes. **(F)** Asymmetry scores throughout spermatogenic gene categories, computed as the $\log_2$ ratio of the coding to the template mutation rates (shown in **E**). Significance is computed by the Mann-Whitney test. *, $P<0.005$; **, $P<0.00001$; n.s., not significant. Error bars indicate 99% confidence intervals.

The public databases have amassed over 200 million germline variants detected in the human population, providing a rich resource for studying germline mutation rates [31]. Since ~80% of these germline variants are thought to have originated in males [33,34], we used this dataset to query for widespread transcription-induced effects on the pattern of germline mutations. We thus sought to compare the number of DNA variants between genes expressed and unexpressed in spermatogenesis as a proxy for a difference in the level of DNA damage [35,36]. Interestingly, we found that spermatogenesis expressed-genes, regardless of spermatogenic stage of expression, generally have a lower level of germline mutations, relative to the unexpressed genes (Fig. 2C), consistent with previous notion of transcription-coupled repair in spermatogenic cells [37,38]. This difference is not observed in the gene flanking sequences (5kb of upstream and downstream), indicating a stronger effect in the genic region (Fig. S3) and supporting the notion that the widespread spermatogenesis transcription reduces the level of germline mutations.

If the reduction of mutations follows from a TCR-induced process, we would expect an asymmetry between the mutation levels of the coding and the template strands in the

spermatogenesis expressed genes, but not in the unexpressed genes [32,38–41]. The asymmetry would be such that the template strand accumulates fewer mutations since, in TCR, the RNA polymerase on the template strand detects DNA damage [21]. To distinguish between mutations occurring on the coding and template strands, we adapted previous approaches to identify strand-asymmetries in the mutation rate (Fig. 2D) [32,38]. By studying mutation categories with reference to the coding and template strand, Haradhvala *et al.* inferred a bias in mutation rates (Fig. 2D, schematic) [32] and such strategy was also utilized by Chen *et al.* [38]. We applied this approach to germline mutations and found that a lower mutation rate was inferred on the template strands of expressed genes during spermatogenesis, while such effect is unapparent in the unexpressed genes, as represented by A>T transversion mutations in Figure 2E and in the other mutation types (Fig. S4A). In addition, for the coding strand, we observed an inferred rate of mutations that is lower in the expressed relative to that in the unexpressed genes, suggesting that antisense transcription in spermatogenesis may be used to further reduce mutation levels [42].

We next computed an 'asymmetry score' to study the ratio between mutation levels inferred to occur in the coding and template strands (Fig. 2E-F) [32]. As expected, the unexpressed group of genes has minimal level of asymmetry scores (Fig. 2F and Fig. S4E), indicating no transcription-induced removal of DNA damage. Examining this measure across the spermatogenic stages, we observed that the asymmetry scores are highest in the early stages of spermatogenesis (spermatogonia and spermatocytes) and gradually decrease along the spermatogenesis lineage (Figs. 2F, S4D), consistent with a stronger transcription-induced removal of DNA damage earlier in spermatogenesis. Such a pattern is also reflected in the expression levels of TCR genes which show higher expression levels in early spermatogenesis (Fig. S7). As negative controls, we found that mutational asymmetry was not observed when

comparing Watson and Crick strands (instead of gene-specific coding and template strands, Fig. S5), nor did we detect difference between the gene groups when shuffling the spermatogenic gene group assignments (while maintaining the group sizes, Fig. S6).

Bidirectional transcription signatures of mutation asymmetries

While the Figure 2 analysis examined transcription in the gene body (start to end of mRNA transcription), transcription in the human genome contains additional levels of complexity. For example, while expression is usually considered as transcribing the gene body, transcription in the opposite direction is common [43,44], leading to bidirectional transcription initiation on opposite strands (Fig. 3A). If lower mutation rates are indeed transcription-induced, we would predict that mutation asymmetry scores would display an inverse pattern between the opposite strands of the initiation of bidirectional transcription. Consistently, we detected an inverse pattern of asymmetry scores between the gene body and the upstream sequences (Figs. 3B,C, S4). Since transcription may extend beyond the annotated end or alternative polyadenylation sites (Fig. 3A) [45], we would also predict that the asymmetry scores between the gene body and the downstream sequences would display a coherent pattern. Again, we find the expected pattern whereby the gene body and the downstream sequences have the same pattern of asymmetry scores (Figs. 3B,D, S4). Together, these analyses provide striking support for transcription-induced germline mutation reduction.

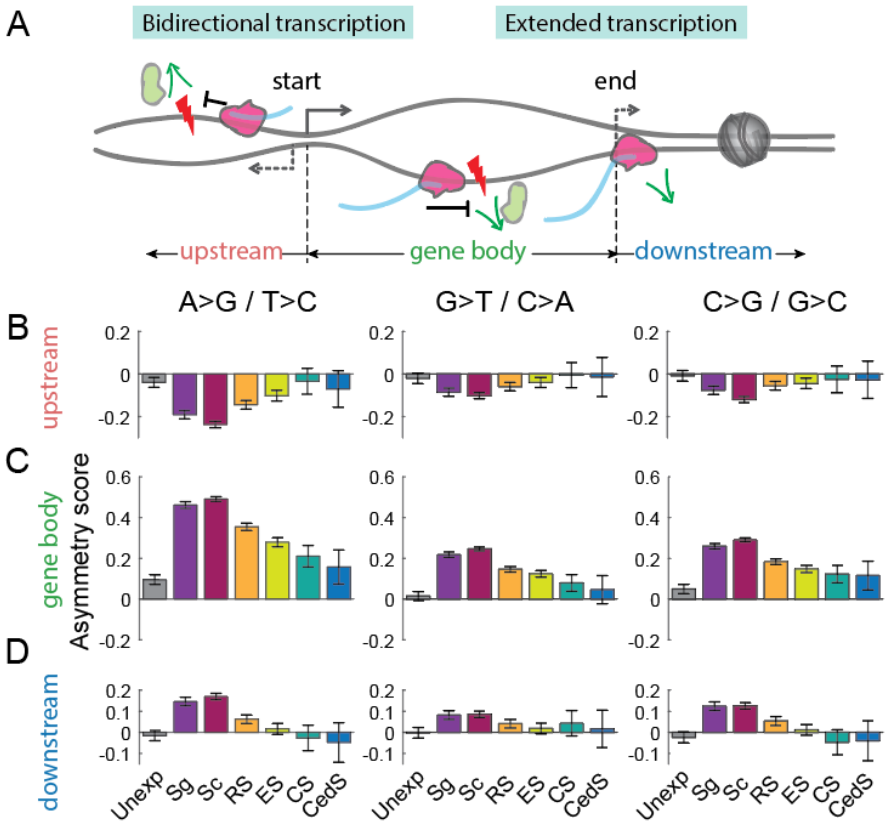

Fig. 3. TCR-associated mutation asymmetry scores show bidirectional transcription and extended transcription signatures. (A) Gene model indicating bidirectional and extended transcription. The model shows that relative to the promoter, upstream and gene body transcription occur on opposite strands, while downstream transcription occurs on the same strand as the gene body. **(B-D)** Asymmetry scores in the upstream 5kb region **(B)**, gene body **(C)** and downstream 5kb region **(D)**. Three mutation types are shown here (A>G, G>T and C>G); the rest are shown in Fig. S4.

'Transcriptional scanning' is tuned by gene-expression level.

Our results led us to propose a model whereby widespread spermatogenesis transcription functions for 'transcriptional scanning' to reduce DNA damage-induced mutagenesis and thus

safeguard the germline genome (Fig. 4A). Such a model suggests that mutation rates of scanned genes might be tuned by their expression levels in the testis. First, we expect that even minimally expressed genes should show fewer mutations than unexpressed genes, since a single round of transcription would pick up any damage. To test this, we binned all genes into seven groups according to their peak level of expression (Fig. 4B, SI methods). Consistently, we found that even the most lowly-expressed genes have lower levels of germline mutations than the unexpressed genes (Figs. 4C, S8A-B).

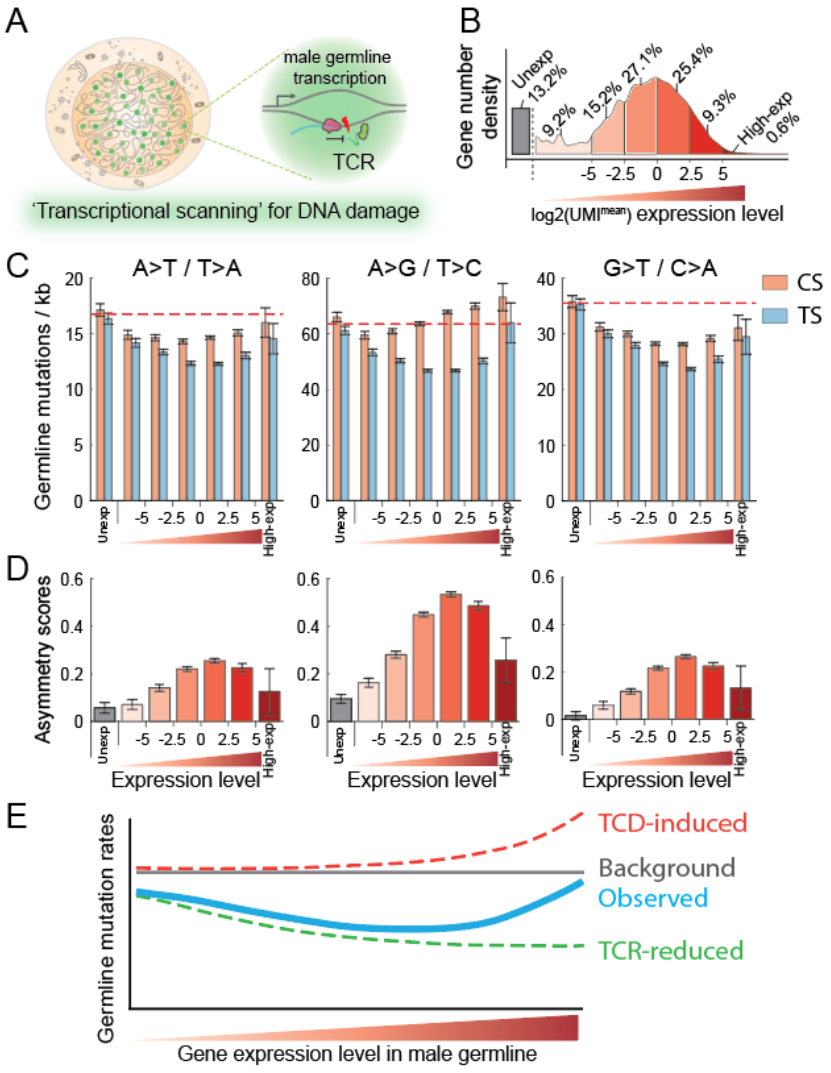

Fig. 4. 'Transcriptional scanning'-induced mutation reduction is tuned by gene-expression level. (A) Model for transcriptional scanning of DNA damage in male germ cells. **(B)** Genes were binned to seven gene expression level groups, from unexpressed (Unexp) to highly-expressed (High-exp) (SI methods). **(C)** Distributions of the indicated germline mutation types across gene expression level categories, and distinguished by coding and template strands. Dashed lines indicate the average level of mutations in the unexpressed genes. **(D)** Distribution of asymmetry scores between coding and template strand for the mutation types indicated in **(C)**. **(E)** A model for gene expression level tuning of germline mutation rates following additive contributions by transcription-coupled repair (TCR-reduced) and transcription-coupled damage-induced (TCD-induced) effects.

The 'transcriptional scanning' model predicts that higher expression levels would lead to additional scanning, and consequently further reduced mutation rates on the template strand. Indeed, examining our asymmetry score according to different expression levels, we observed that as expression level increases, the overall mutation level drops (Fig. 4C). Surprisingly, however, the very highly expressed genes showed the opposite effect: asymmetry between the strands is reduced and a paradoxically higher level of germline mutations relative to the unexpressed genes is observed (Figs. 4C,D, S8A,B). This pattern is consistent with observations that very high expression levels can lead to transcription-coupled DNA damage (Fig. 2A), as previously reported for transcription-associated mutagenesis in highly expressed genes in other systems [46]. The mutation type in which TCD is most evident is A>G (Fig. 4C), and similarly, such TCD was readily observed in somatic A>G mutation in liver cancer samples [32]. Our

findings therefore extend support for TCD occurring for all mutation types in highly expressed genes (Figs. 4C-D, S8).

Our analyses suggest that spermatogenesis gene-expression levels tune germline mutation levels and we interpret our results as follows (Fig. 4E). 'Transcriptional scanning' reduces mutation rates even in genes with low-expression. Increasing expression levels are correlated with further reductions in mutation rates, but only to a point. In the very highly expressed genes, TCD overwhelms the TCR-induced reductions, and produces an overall higher mutation rate than genes expressed at low and moderate levels (Fig. S8A).

Transcriptional scanning and differential rates of genome evolution

We hypothesized that the reduction in mutation rates by transcriptional scanning would have cumulative effects over evolutionary time-scales. Specifically, since we observed lower mutation rates for spermatogenesis expressed genes at the level of the human population, we expected that these genes would be more conserved at the sequence level across orthologues in other apes (Fig. S9A), than the unexpressed genes. Consistently, examining across our stage-specific gene groups, we found that unexpressed genes show the highest level of divergence when comparing across the apes (Fig. 5A). Examining divergence across expression levels, we found a negative correlation between increased expression and divergence (Fig. 5D). However, the most highly expressed genes showed higher divergence. These observations are fully consistent with our analyses implicating higher mutation rates by TCD (Fig. 4). Collectively, as expected, the same mutation-level pattern is detected both in the population (Figs. 2-4) and across species (Fig. 5).

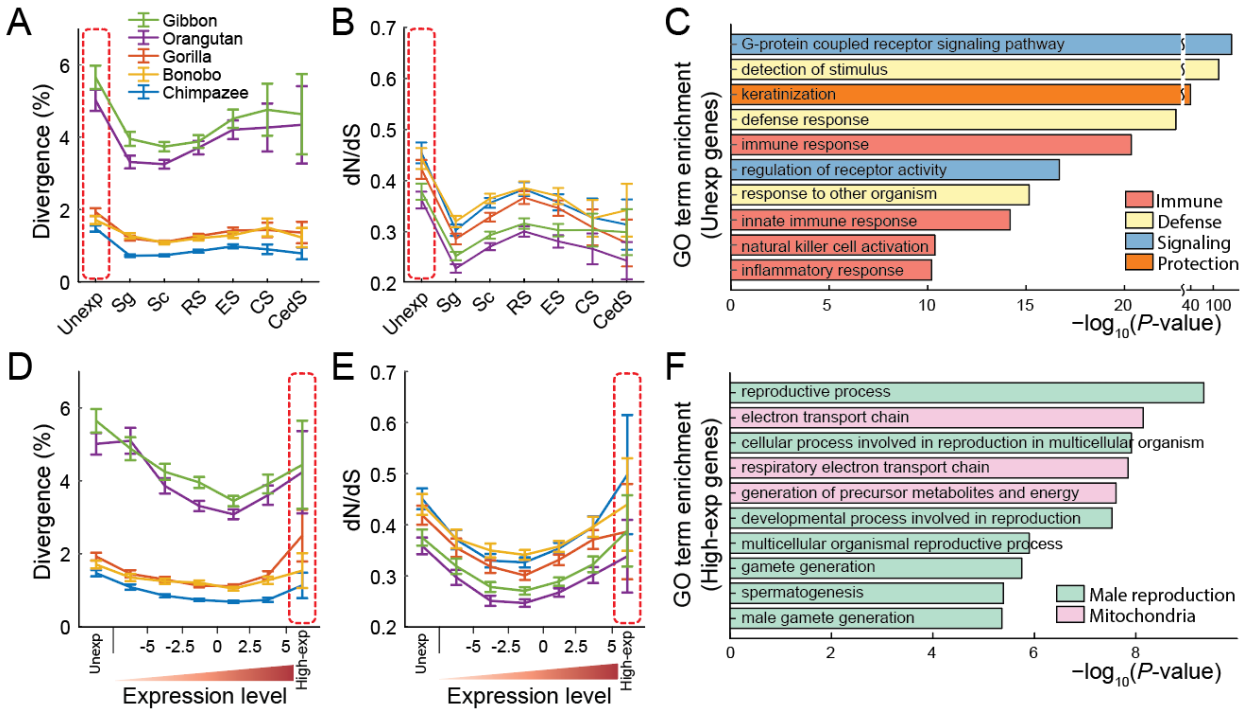

Fig. 5. Evolutionary consequences of 'transcriptional scanning' in male germ cells. (A) DNA divergence levels of human genes with their ortholog in the indicated apes, according to spermatogenic stages. **(B)** Same as **(A)** for dN/dS values. **(C)** Gene ontology categories enriched in the set of genes unexpressed during spermatogenesis (*P*-value is indicated). **(D)** Same as **(A)**, according to gene expression level categories. **(E)** Same as **(D)** for dN/dS values. **(F)** Gene ontology categories enriched in the set of genes that are very highly expressed during spermatogenesis.

The observation of different evolutionary rates between spermatogenesis expressed and unexpressed genes suggests a distinct selective regime acting upon the unexpressed genes. To test this, we studied the ratio of nonsynonymous to synonymous substitution rates (dN/dS) of evolution for stage-specific and expression-level specific gene groupings. We found that the unexpressed genes have a higher dN/dS ratio than the expressed genes, indicating that they are

subject to weaker levels of purifying selection (Figs. 5B, S9B,C). Thus, the higher divergence levels of the unexpressed genes follows from both their higher mutation rates (Fig. 2C) and their weaker levels of purifying selection. Studying the set of 2,623 unexpressed genes at the functional level, we found that this set is enriched for environmental sensing, immune and defense systems, and signaling genes (Fig. 5C and Table S1). These functions strikingly coincide with those known to be fast-evolving in the human genome [23,24]. Our results suggest that, beyond differential levels of purifying selection, the underlying levels of mutations are increased in this important set of genes by virtue of their being unexpressed during spermatogenesis. Our analysis into expression levels further revealed that the very highly expressed genes will also have high mutation levels (Fig. 4). We found that the very highly expressed genes also exhibit low levels of purifying selection (high dN/dS, Fig. 5E). Functionally, this set of genes is enriched for roles in male reproduction and mitochondrial function (Fig. 5F and Table S2).

Discussion

Our findings led us to propose a model whereby widespread transcription at fine-tuned levels of expression leads to a rugged landscape of germline mutations by transcriptional scanning (Fig. 6). Given that this process is carried out in the germline, the variable mutation rates have important implications for genome evolution. In this model, the widely transcribed genes in male germ cells benefit from transcription-coupled repair (TCR), which scans through the expressed genes, thereby reducing germline mutations and safeguarding the germ cell genome. Over long time-scales these genes evolve slower (Fig. 6 middle). The small group of genes that are unexpressed throughout spermatogenesis are enriched for sensory and defense-immune system genes (Fig. 5C) and exhibit higher mutation rates, which in our model is explained by the lack of

a TCR-induced germline mutation reduction (Fig. 6 left). Defense and immune system genes are known to evolve faster [23,24] and our selective transcriptional scanning model provides insight into how variation is preferentially provided to this class of genes. Such rapid evolution may be under strong selective biases for adaptation at the population-level in rapidly changing environments. A third class of genes are characterized by very high germline expression. These genes have higher germline mutation rates since their transcription-coupled DNA damage obscures the effect of transcription-coupled repair (Fig. 6 right). This model provides more comprehensive view of TCR-TCD crosstalk in spermatogenic cells with expression level-tuned mutation rates fluctuation (Fig. 4E), and corrects the previous observation that the germline mutation rates increase with expression levels [38]. In this Discussion, we address the issues of the full spectrum of mutagenesis pattern in the male germline, a proxy for detecting important genomic regions, and testable predictions of our model.

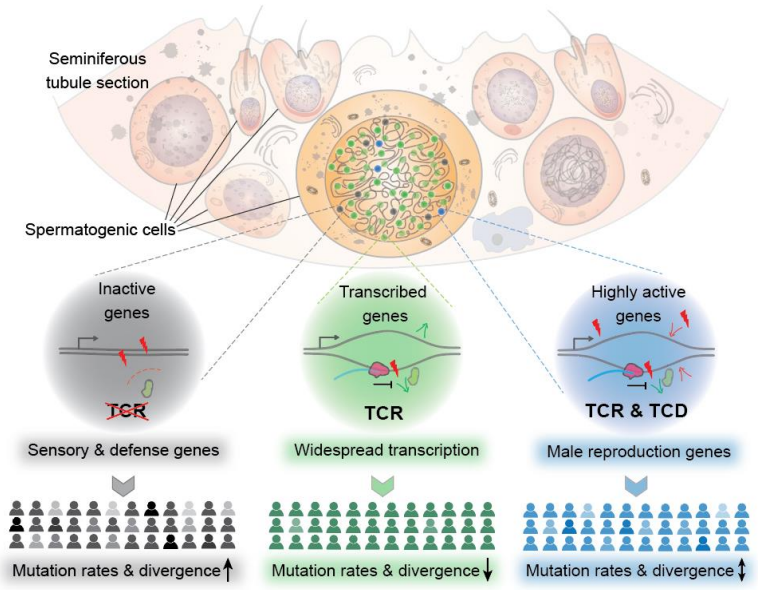

Fig. 6. A model for widespread transcriptional scanning in male germ cells. The transcriptional scanning model predicts reduced germline mutation rates across most expressed genes. Genes unexpressed in spermatogenesis have higher relative mutation rates and consequently experience more evolutionary divergence. In the very highly-expressed genes, transcription-coupled DNA damage overwhelms the effects of TCR, resulting in higher mutation rates in these genes, highly enriched for male reproductive function genes.

The transcriptional scanning model can account for a reduction of ~15-20% of mutagenic DNA damage by detecting and removing bulky germline DNA damage (as estimated from the Fig. 2C analysis). Such a mechanism is critical for germ cell viability as retained bulky DNA damage may lead to cell death[47]. On the other side, the expressed genes of male germ cells still retain mutations that cannot be repaired by the TCR machinery[22,48]. These male germline mutations likely originate from DNA replication errors, accumulating with paternal age [49]. Thus, it would be of great interest to further analyze the observed germline mutation pattern, in particular relative to replication fork directionality [50].

Beyond the protein-coding genes expressed here, it would be interesting to study non-coding genomic regions that are also expressed in the testes. Previous studies have reported that testis also expressed large numbers of non-coding genes[10]. These genomic regions may be inferred to be biologically important given that they are subjected to TCR-induced mutation reduction. According to this logic, it might follow that sensory and defense-immune system genes are unimportant since they are not generally expressed in the testes. Instead, we argue that this gene set is the exception that highlights the rule. In other words, most genes benefit from TCR mutation reduction excepting those under selection for faster evolution. Similarly to

phylogenetic profiling for identifying functionally important regions of the genome [51], identification of testis-expressed regions – for example non-coding genes and retrotransposons – may be an efficient method for identifying these important regions.

Our model leads to important testable predictions and may provide deeper insights into human genetics and diseases originated from *de novo* germline mutations. First, we predict that *de novo* male-derived mutations would be enriched for genes unexpressed in spermatogenesis. Second, the same process should also hold in other mammals. Finally, we would expect that TCR-deficient animals should produce offspring with an increase in the number of *de novo* mutations. For patients with TCR gene-associated mutations, such as Cockayne syndrome and xeroderma pigmentosum [52], our model predicts higher germline mutation rates. It would also be of interest to study TCR/TCD processes in the female germline, though widespread gene expression has not been reported in the ovaries [11]. The brain is another organ with a highly complex transcriptome [3,10], and it would be interesting to explore whether transcriptional scanning might have a function in certain somatic tissues. For example, such a function might help prevent somatic mutation induced neurodegenerative diseases in the aging brain [53].

Acknowledgments: We thank Yael Kramer for coordinating the human sample collection. We thank Molly Przeworski, Hannah Klein, Huiyuan Zhang and the members of the Yanai lab for constructive comments and suggestions to the manuscript. We thank Megan Hogan and Matthrew Maurano for assistance with sequencing.

Funding: This work was supported by the NYU School of Medicine with funding to I.Y.

Author contributions: B.X. and I.Y. conceived the project, interpreted the results and drafted the manuscript. B.X. led the experimental and analysis components. M.B. contributed expertise in the inDrop analysis and sequencing. Y.Y. contributed to RNA velocity and Monocle2 analysis, and mutation data processing. F.W. contributed to raw data processing of scRNA-seq and cell ordering. J.A., S.Y.K., and D.K. contributed to the sample collection. All authors edited the manuscript.

Competing interests: Authors declare no competing interests.

Data and materials availability: Raw sequencing data will be deposited to GEO and will include gene expression matrices including both smoothed and unsmoothed UMI counts matrices.

Materials and Methods

Human testes sample

Human testis tissue was obtained from New York University Langone Health (NYULH) Fertility Center; this was approved by the NYULH Institutional Review Board (IRB). Fresh seminiferous tubules were collected from testicular sperm extraction (TESE) surgery of a healthy patient with an obstructive etiology for infertility; there were no drug or hormonal treatments prior to TESE surgery. The research donor was fully informed before signing consent to donating excess tissue for research use; this was again done in fashion consistent with the IRB (including tissue sample de-identification).

Single cell suspension preparation

After TESE surgery, samples were kept in cell culture PBS and transported to the research lab on ice within 1h of surgery for single-cell preparation. Testicular single-cell suspension was prepared by adapting existing protocols [54]. Specifically, samples from TESE surgery was washed once with PBS and resuspended in 5mL PBS. Seminiferous tubules were minced quickly in a cell culture dish and spun down at 100g for 0.5min to remove supernatants. The minced tissue was resuspended in 8mL of 37°C pre-warmed tissue dissociation enzyme mix (See below). Tissue dissociation was done by incubating at 37°C for 20min with mechanical dissociation with pipetter every 5min. After digestion, the reaction was quenched by adding 2mL of 100% FBS (Gibco, Cat. 16000044) to a final concentration of 10%. Dissociation mix was filtered through a 100um strainer to remove remaining seminiferous tubule chunks. Cells were washed once with DMEM medium (Gibco, Cat. 11965092) with 10% of FBS and twice with PBS. Cell viability was checked with Trypan-blue staining (with expectation of over 85% viable cells) before moving to the inDrop microfluidics platform. The tissue dissociation enzyme mix (8mL) was composed of 7.56mL of 0.25% Trypsin-EDTA (Gibco, Cat. 25200056), 400uL of 20mg/mL type IV Collagenase (Gibco, Cat. 17104019) and 40uL of 2U/uL TURBO DNase (Invitrogen, Cat. AM2238).

Single-cell RNA-Seq

Single-cell barcoding was carried out with the inDrop microfluidics platform [27] as instructed by the manufacturer (1CellBio). Briefly, the microfluidic chip and barcoded hydrogel beads were primed ahead of single cell preparation. The ready-to-use single-cell suspension in PBS (after two times wash with PBS buffer) was adjusted to 0.1 million/mL by counting with hemocytometer. Next, the prepared cells, reverse transcription reagents (SuperScript III Reverse Transcriptase, Invitrogen, Cat. 18080085), barcoded hydrogel beads and droplet-making oil were loaded onto the microfluidic chip sequentially. Encapsulation was done by adjusting microfluidic flow rates as instructed. Single-cell barcoding and reverse transcription in the droplets were done by incubating at 50°C for 2h followed by heat inactivation at 70°C for 15min. Barcoded single-cells in droplets were aliquoted as desired and then decapsulated by adding demulsifying agent.

Sequencing library preparation

Single-cell RNA-Seq library preparation after inDrop was carried out as instructed by the manufacturer (1CellBio) and similar to the CEL-Seq2 method [55]. Basically, barcoded single-cell cDNA was purified with Agencourt RNAClean XP magnetic beads (Beckman Coulter, Cat. A63987) followed by second-strand synthesis reaction with NEBNext mRNA Second Strand

Synthesis KIT (New England Biolabs, Cat. E6111S). Then linear amplification of cDNA was carried out through *in vitro* transcription (IVT) using HiScribe T7 High Yield RNA Synthesis kit (New England Biolabs, Cat. E2040S). IVT-amplified RNA was fragmented and purified again with Agencourt RNAClean XP magnetic beads. The second reverse transcription was done with PrimeScriptTM Reverse Transcriptase (Takara Clonetech, Cat. 2680A) followed with cDNA purification with Agencourt AMPure XP magnetic beads (Beckman Coulter, Cat.A63881). cDNA quantity was determined by qPCR on a fraction (5%) of purified cDNA. Final PCR amplification was done according to qPCR results and purified with Agencourt AMPure XP magnetic beads. Library concentration was determined by Qubit dsDNA HS Assay Kit (Invitrogen, Cat. Q32851). Library size was determined by Bioanalyzer High Sensitivity DNA Kit (Agilent, Cat. 5067-4626).

Sequencing

Single-cell RNA-Seq library sequencing was carried out with Illumina NextSeq 500/550 75 cycles High Output v2 kit (Cat. FC-404-2005). Custom sequencing primers were used as instructed by manufacturer [27]. In addition, 5% of PhiX Control v3 (Illumina, Cat. FC-110-3001) library was added to give more complexity to scRNA-Seq libraries. Pair-end sequencing was carried out with read1 (barcodes) for 34bp, index read for 6bp and read2 (transcripts) for 50bp.

Sequencing data processing

Raw sequencing data obtained from the inDrop method were processed using a custom-built pipeline, available at (https://github.com/flo-compbio/singlecell). Briefly, the "W1" adapter sequence of the inDrop RT primer was located in the barcode read (the second read of each fragment), by comparing the 22-mer sequences starting at positions 9-12 of the read with the known W1 sequence ("GAGTGATTGCTTGTGACGCCTT"), allowing at most two mismatches. Reads for which the W1 sequence could not be located in this way were discarded. The start position of the W1 sequence was then used to infer the length of the first part of the inDrop cell barcode in each read, which can range from 8-11 bp, as well as the start position of the second part of the inDrop cell barcode, which always consists of 8 bp. Cell barcode sequences were mapped to the known list of 384 barcode sequences for each read, allowing at most one mismatch. The resulting barcode combination was used to identify the cell from which the fragment originated. Finally, the UMI sequence was extracted, and reads with low-confidence base calls for the sex bases comprising the UMI sequence (minimum PHRED score less than 20) were discarded. The reads containing the mRNA sequence (the first read of each fragment) were mapped by STAR 2.5.1 with parameter "—outSAMmultNmax 1" and default settings otherwise[56]. Mapped reads were split according to their cell barcode and assigned to genes by testing for overlap with exons of protein-coding genes and long non-coding RNA genes, based on genome annotations from Ensembl release 90. For each gene, the number of unique UMIs across all reads assigned to that gene was determined (UMI filtering), corresponding to the number of transcripts expressed and captured. Cells with a total transcript count of less than 1,000 or more than 20% of transcripts originating from mitochondrial genes (i.e., genes that are part of the mitochondrial genome) were removed for downstream analysis. The resulting gene expression matrix contained UMI counts for 27,378 genes across 783 cells.

Inferring the transcriptomic trajectory of spermatogenesis

To obtain a temporal ordering of our cells that reflected the developmental process of spermatogenesis, we first filtered the expression matrix for protein-coding genes, retaining 19,788 genes. We then applied a variant of our recently proposed kNN-smoothing method [57], with *k*=3. This variant differed from the published version in that it relied on the Anscombe transform ($y = 2\sqrt{x + 3/8}$) instead of the Freeman-Tukey-transform as a variance-stabilizing transformation, and in that it identified all neighbors in a single step, rather than adopting a step-wise approach. Briefly, all single-cell expression profiles were normalized to median number of total transcripts per cell [58], the Anscombe transform was applied to all expression values, and the *k*=3 closest neighbors of each cell were identified using Euclidean distance. The expression profile of each cell was then combined with those of its neighbors, thus obtaining its smoothed expression profile.

We next transformed the smoothed data using principal component analysis, and applied multidimensional scaling (MDS) to the cell scores for the first four principal components. Based on the two-dimensional results, we constructed a nearest-neighbor graph in which we connected each cell to its closest 32 neighbors, with a maximum distance of 80. We calculated the minimum spanning tree of this nearest-neighbor graph, determined the longest path in the tree, and applied smoothing by averaging the x and y coordinates of four consecutive vertexes. This created a continuous "backbone" representing the transcriptomic trajectory of spermatogenesis. To obtain the temporal ordering of all cells, we then projected all cells onto this path in the manner described by Qiu et al [29] and excluded 42 cells (5.4 %) with a distance of 25 or greater, which likely presented rare cell types or damaged cells. We used the expression of the *PRM1* gene [59] to determine which "end" of the ordering corresponded to the last stage of spermatogenesis. Minimal manual adjustments to the cell ordering inferred through the aforedescribed process were made by comparison with unsupervised hierarchical clustering results. Finally, we obtained a temporal ordering (from early to late) for 741 cells that formed the basis for our downstream analyses.

Cell stage and cell type identification

Following MDS ordering of cells, several marker genes were used to determine cell types or spermatogenic stages. *CSF1*, *CYP11A1* and *IGF1* [60–62] genes were used to distinguish Leydig cells. *WT1* and *SOX9* [61,63] were used to distinguish Sertoli cells. Both Leydig cells and Sertoli cells were then excluded from the dataset to determine developmental stages of spermatogenesis. *FGFR3* and *DMRT1* [26,64] were used to determine spermatogonia. *SYCP3* and *TEX101* [61,65] were used to determine spermatocytes. *ACRV1* and *ACTL7B* [61,65] were used to determine round spermatids. *TNP1*, *PRM1*, *PRM2*, *YBX1* and *YBX2* [18,59,65,66] were used collectively to determine elongating spermatids, condensing spermatids and condensed spermatids. Based on the main spermatogenic stages, a more detailed spermatogenesis staging were defined by hierarchical clustering to increase resolution.

Principal component analysis (PCA)

The PCA plots in Figure 1 and S1 were perform on the UMI expression matrix of all testicular cells (741 cells, Fig. 1B) or spermatogenic cells (664 cells, Fig. 1C). In both cases, expression matrices were first normalized to 100,000 transcripts per cell. Fano factor or variance-to-mean ratio (VMR) was computed for each gene to determine dynamically expressed genes. PCA was then performed on the normalized and $\log_2$ transformed expression matrix using

the dynamically expressed genes. For all testicular cells (Fig. 1B), 860 dynamic expressed genes were included. For spermatogenic cells (Fig. 1C), 1648 dynamic expressed genes were used.

Spermatogenic cell ordering by Monocle2

With the same smoothed spermatogenic cell expression matrix for building developmental trajectory as input, we used Monocle2 (version 2.6.0) [29] to infer the pseudotime track. We performed the required processes with default parameters according to the user manual (http://cole-trapnell-lab.github.io/monocle-release/docs/): 1) Set "negbinomial.size()" for expression distribution, and estimated size factors and dispersions. 2) Selected genes detected among at least 5% of 664 cells to project cells to 2D space using "DDRTree" method. 3) Ordered cells and visualized pseudotime track as shown in Fig. S1D. The increasing order of pseudotime values was consistent to the pattern of marker genes during spermatogenesis (data not shown). Pseudotime values were unique so the index of cell order was determined. The Monocle2-determined and MDS-determined cell index were plotted and Pearson correlation coefficient was calculated as shown in Fig. S1E.

Cell fate prediction with "RNA velocity"

We used the R package velocyto.R (version 0.5) to estimate RNA velocity [28]. This required three separate counts matrices (emat, nmat, and spmat) which were composed of the intronic UMIs, exonic UMIs and intron/exon spanning UMIs, respectively. They were generated by the dropEst pipeline (https://github.com/hms-dbmi/dropEst). 1) The raw sequencing reads was tagged by droptag with the default "inDrop v1&v2" config file except "r1_rc_length" was set as 3. 2) The tagged reads were mapped to the human reference genome GRCh38 using STAR (version 2.5.3a) [56] with default settings. 3) The alignments were processed by dropest with gene annotation GTF file (Ensembl release 90) and the default settings except the "--merge-barcodes" option was additionally called as suggested. The result contained 655 of the 664 spermatogenic cells. Pearson correlation coefficient between the UMI count profile of each cell estimated by custom-built single-cell RNA-Seq pipeline (https://gitlab.com/yanailab/singlecell) and dropEst pipeline was calculated and the median of all 655 cells was 0.968.

We followed the velocyto.R manual (https://github.com/velocyto-team/velocyto.R) and used emat and nmat to estimate and visualize RNA velocity. With predefined cell stage, we performed gene filtering with the parameter "min.max.cluster.average" set to 0.1 and 0.03 for emat and nmat, respectively. RNA velocity using the selected 4266 genes was estimated with the default settings except parameter "kCells" and "fit.quantile" which were set to was 3 and 0.05, respectively. RNA velocity field was visualized on a separate PCA embedding as shown in Fig. 1C.

Stage-marker identification

To identify gene markers for stages throughout spermatogenesis, we searched for genes exclusively expressed in the corresponding stage. We constructed an idealized gene expression pattern exclusive to each stage (main or detailed), which was used as a reference to find gene expression pattern. A correlation coefficient higher than 0.5 and *P*-value lower than 0.0001 was used as thresholds to detect stage-specific marker genes. The top 50 genes with the highest correlation coefficient values to each stage are shown in Fig. S2.

Delineating the stage and expression level groups

To assign genes to specific stages, we computed for each, its average gene expression levels across the six main stages (Sg, Sc, RS, ES, CS, CedS). Genes were then assigned to a main stage in which they have highest level of expression. Unexpressed genes formed a separate group.

To assign groups based on expression levels, we binned the peak expression level to 7 groups:

Group 1: $\log_2(\mathrm{UMI}^{\mathrm{mean\_peak}}) = 0$, unexpressed;
Group 2: $\log_2(\mathrm{UMI}^{\mathrm{mean\_peak}}) \leq -5$;
Group 3: $-5 < \log_2(\mathrm{UMI}^{\mathrm{mean\_peak}}) \leq -2.5$;
Group 4: $-2.5 < \log_2(\mathrm{UMI}^{\mathrm{mean\_peak}}) \leq 0$;
Group 5: $0 < \log_2(\mathrm{UMI}^{\mathrm{mean\_peak}}) \leq 2.5$;
Group 6: $2.5 < \log_2(\mathrm{UMI}^{\mathrm{mean\_peak}}) \leq 5$;
Group 7: $5 < \log_2(\mathrm{UMI}^{\mathrm{mean\_peak}})$, highly expressed.

Human germline variations

Human germline variations were downloaded from the Ensembl FTP site (ftp://ftp.ensembl.org/pub/release-91/variation/vcf/homo_sapiens). We selected from these, the variations from dbSNP_150 and used BEDOPS together with custom Bash scripts to associate them with gene body, upstream 5kb and downstream 5kb genomic regions. The gene body region was defined as the genomic interval between the gene start site and gene end site annotated in GTF file (Ensembl release 91). Upstream and downstream 5kb region was defined according to gene body region and with reference to gene strand information. We classified the variants into the six mutation classes: (A>T/T>A; A>G/T>C; T>G/A>C; C>T/G>A; G>T/C>A; C>G/G>C). Each variant was them further distinguished in terms of the coding and the template strands, as previously introduced [32]. The same procedures were also performed on upstream and downstream genomic regions, with the strand specificity (coding strand versus template strand) being assigned in consistent with the associated genes.

The germline mutation rates of the coding and the template stands were calculated by normalizing to a length of 1kb. Specifically, for germline mutations in total, the mutation rates were calculated as the sum of all germline short variants normalized to a length of 1kb. For specific base substitution mutation type, the mutation rates were calculated as the number of specific mutation type normalized to 1kb of the reference base type.

Gene divergence datasets

The sequence divergence datasets of human to apes were downloaded from Ensembl release 91[31]. Percent divergences in Figure 5 were calculated as: Divergence = 100% − Identity (human to other apes). dN and dS values were also retrieved from Ensembl and we excluded genes zero dN or dS. The mean values shown in Figure 5 were computed on non-outlier values, where an outlier value is defined as more than three scaled median absolute deviations (MAD) away from the median. For a set of divergence or dN/dS values made up N genes, MAD is defined as: MAD = median ($|A_i - \mathrm{median}(A)|$), for $i = 1,2,...,N$.

Statistical Analysis

Statistical significance was computed by the Mann-Whitney test (Mann-Whitney-Wilcoxon test or rank-sum test) to test whether two groups of genes have distinct value distributions. Error bars of bar plots represents 99% percent confidence intervals, calculated as 2.58×standard error, as values are all normal distributed or close to normally distributed.

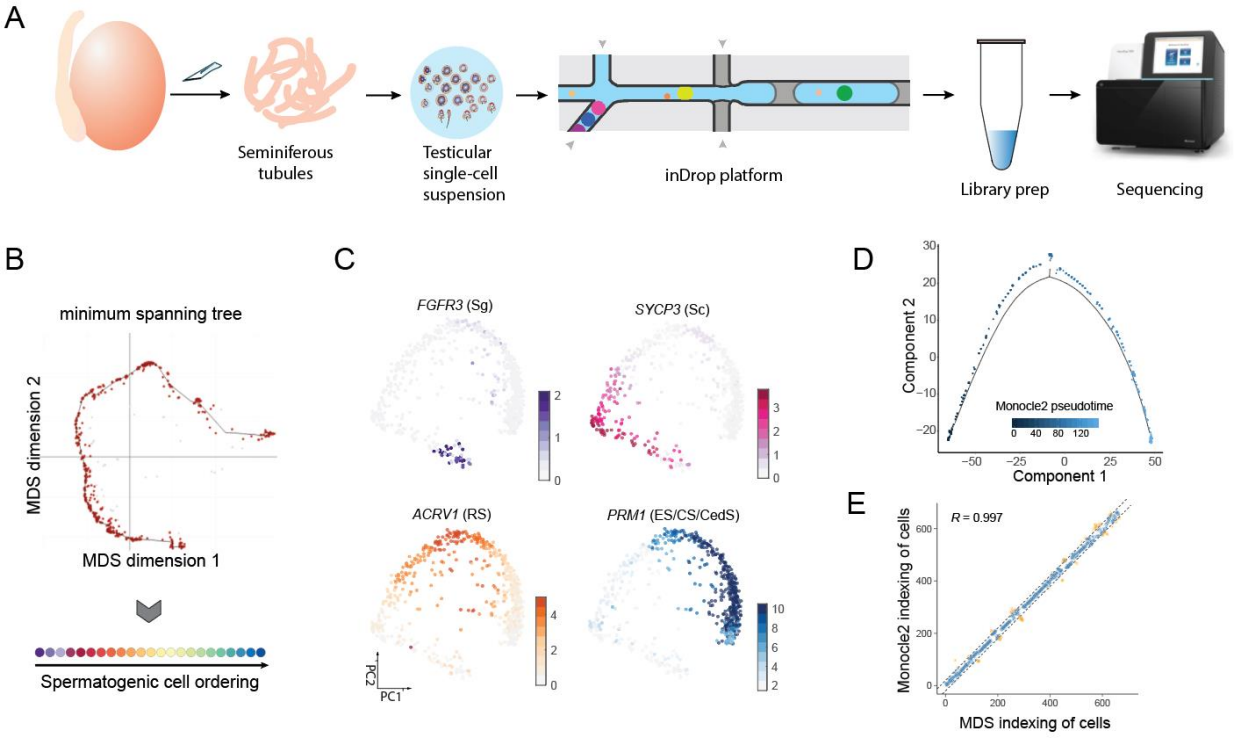

Fig. S1. Single-cell transcriptomic analysis of human spermatogenesis. (A) Schematic of single-cell RNA-seq of human testes sample with the inDrop microfluidics platform (see Methods). (B) Determining the developmental program of spermatogenic cells. A multidimensional scaling (MDS)-embedding on the single cell data was constructed using a no-branching minimum spanning tree, and the cell order is corrected with hierarchical clustering of the cells to determine the developmental time (see Methods). (C) Same PCA as in Figure 1C for the indicated markers of stages. Color indicates gene expression levels. (D) Monocle2-ordering of spermatogenic cells. (E) Comparison of MDS ordering with the Monocel2-determined cell ordering.

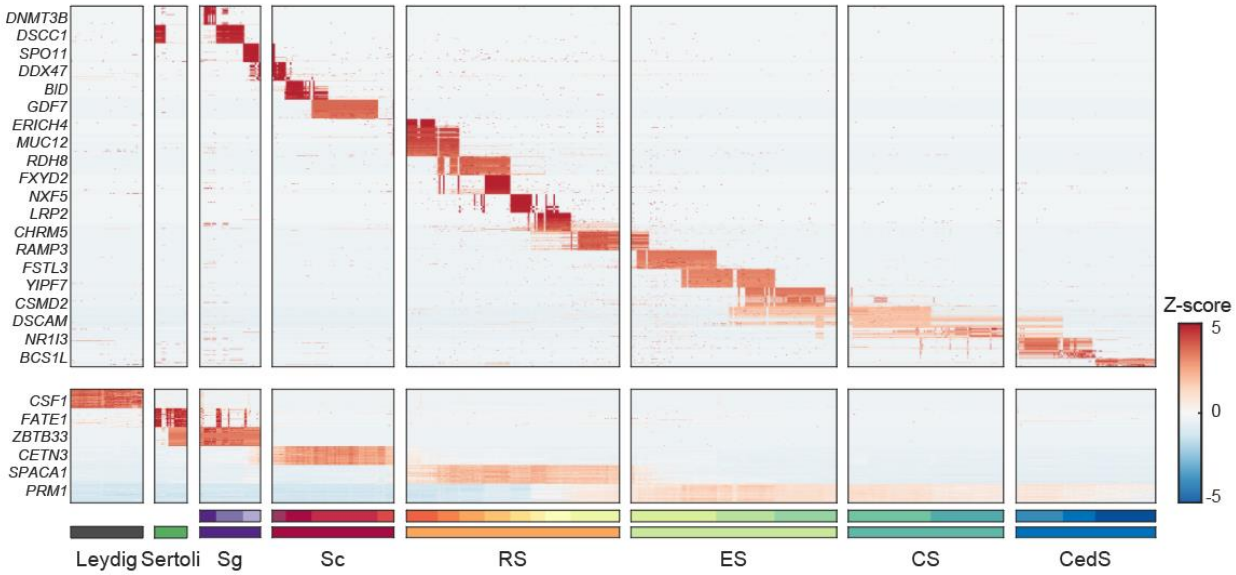

Fig. S2. Heatmap of stage-specific marker gene expression levels. Expression data for both main stages (bottom) and detailed spermatogenic stages (top) marker genes is shown. Gene names are indicated for one representative gene of each stage. Expression levels of at most 50 genes are displayed for each stage.

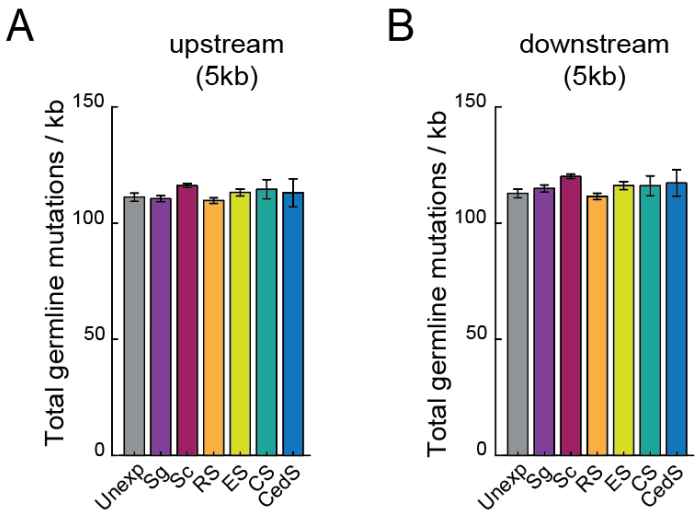

Fig. S3. Germline mutation rates in the flanking regions of human genes. Germline mutation rates in both upstream 5kb (A) and downstream 5kb (B) of genes are shown.

5

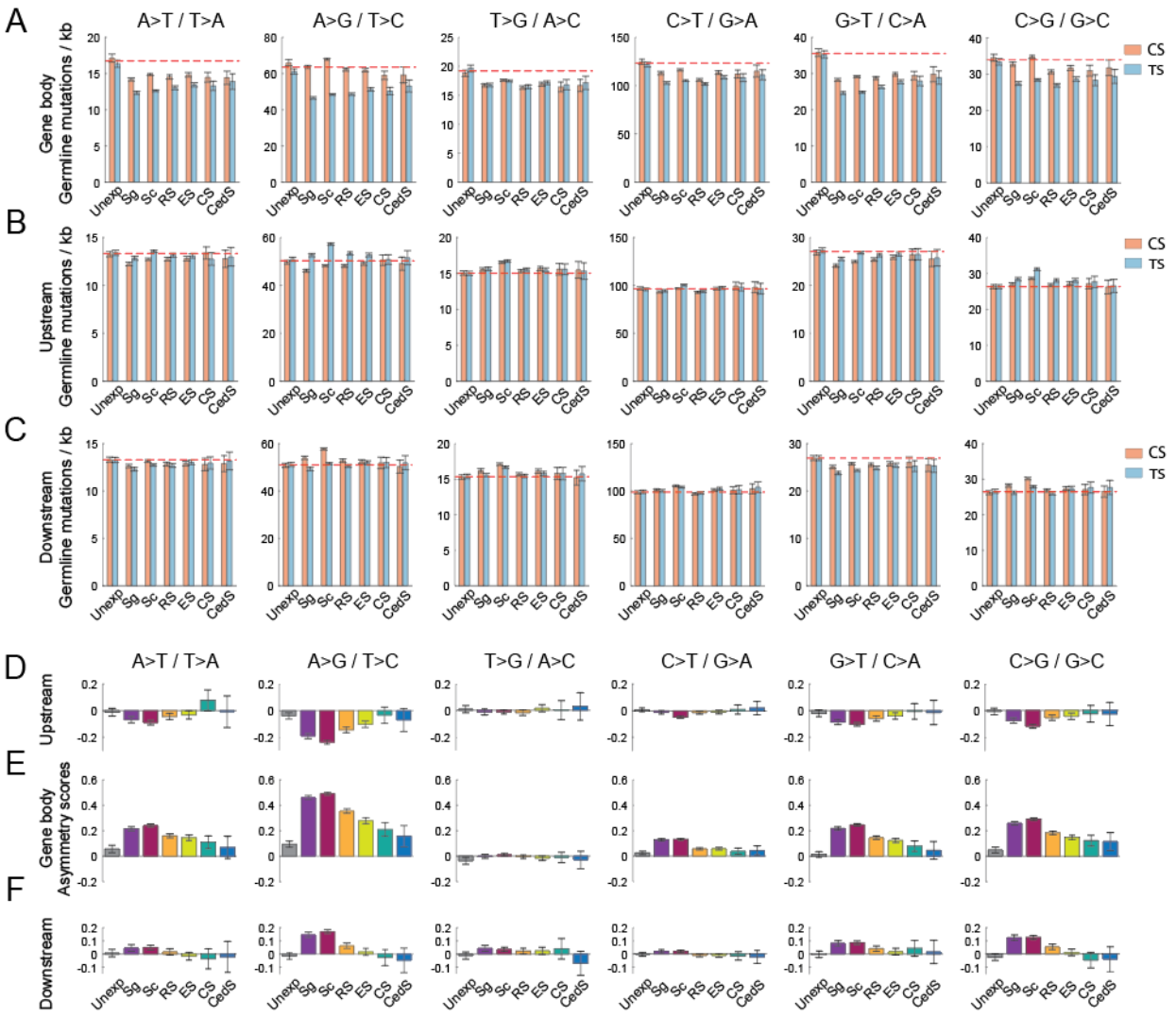

Fig. S4. Germline mutation rates and asymmetry scores of gene body and flanking regions of all base-substitution mutation types. (A-C) Germline mutation rates in the gene body region (A), upstream 5kb (B) and downstream 5kb (C). Dashed lines indicate the average level of mutations in unexpressed genes. (D-F) Germline mutation asymmetry scores between coding and template strands in the upstream 5kb (D), gene body region (E) and downstream 5kb (F).

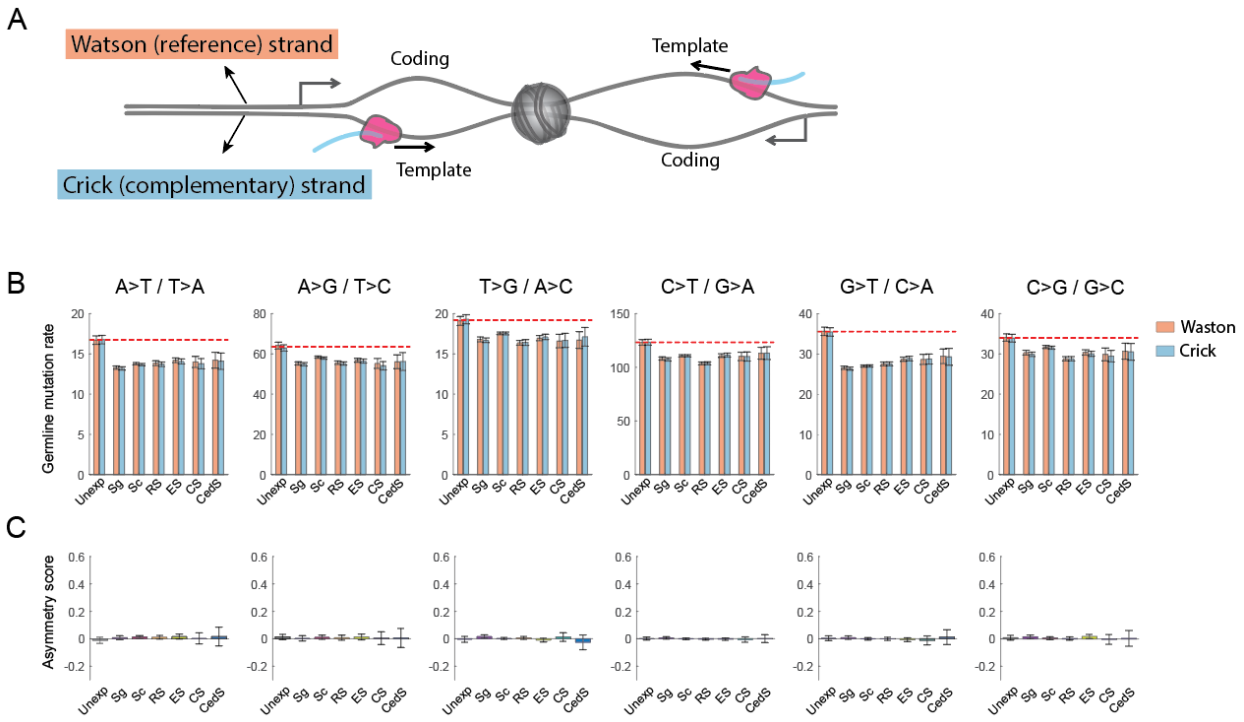

Fig. S5. Mutation rate asymmetry is not detected between the Watson and Crick strands in expressed genes. (A) Schematic of two neighboring genes, each on a different strand. Across the genome, genes are randomly disposed with respect to strand. (B-C) Germline mutation rates (B) and asymmetry scores (C) of all base substitution mutation types across spermatogenesis expressed and unexpressed genes. Mutation rates and asymmetry scores are computed by distinguishing between the Watson and Crick strands, instead of coding and template strands (as shown in Fig. 2D and S4). Dashed lines indicate the average level of mutations in unexpressed genes.

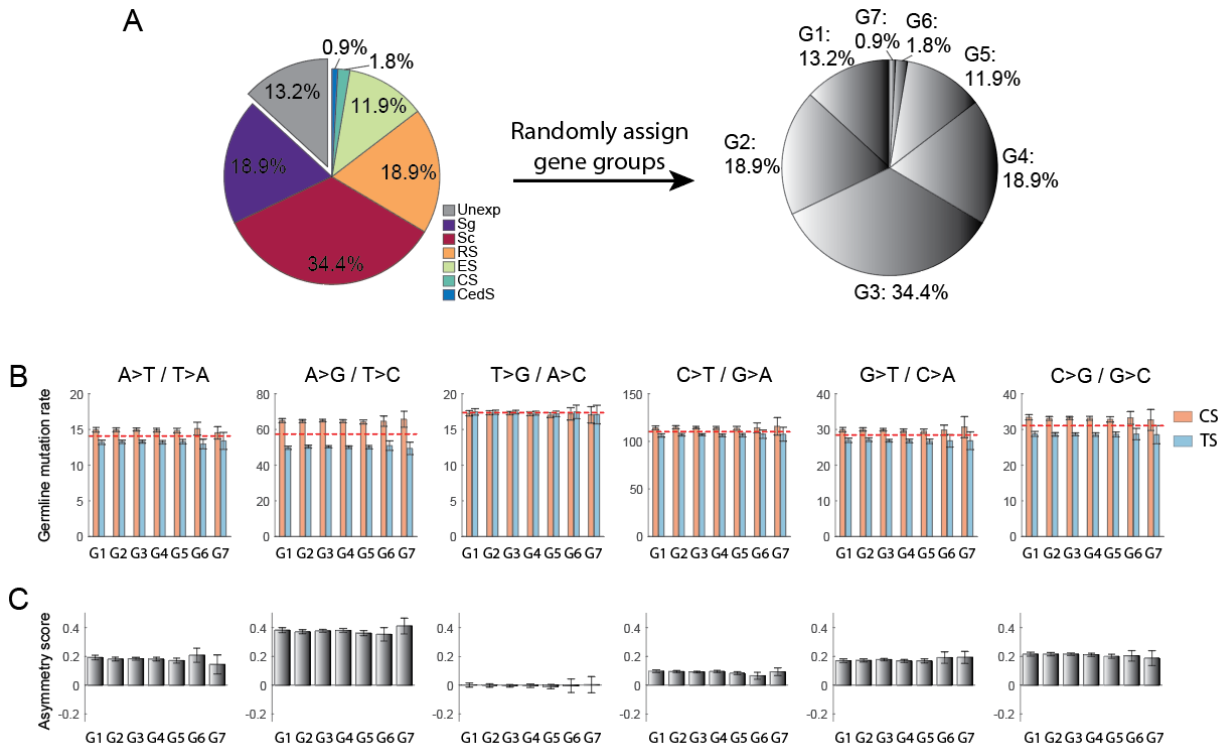

Fig. S6. Shuffling gene assignments loses the mutation-level difference between expressed- and unexpressed genes. (A) Shuffling gene group assignments. Genes assigned to all stages were shuffled, while maintaining the size of each group. (B-C) Germline mutation rates (B) and asymmetry scores (C) of all base substitution mutation types according to shuffled gene-grouping in (A). Mutation rates and asymmetry scores are computed by distinguishing between the coding and template strands (same as in Fig. 2D and S4). Dashed lines indicate the average level of mutations in unexpressed genes.

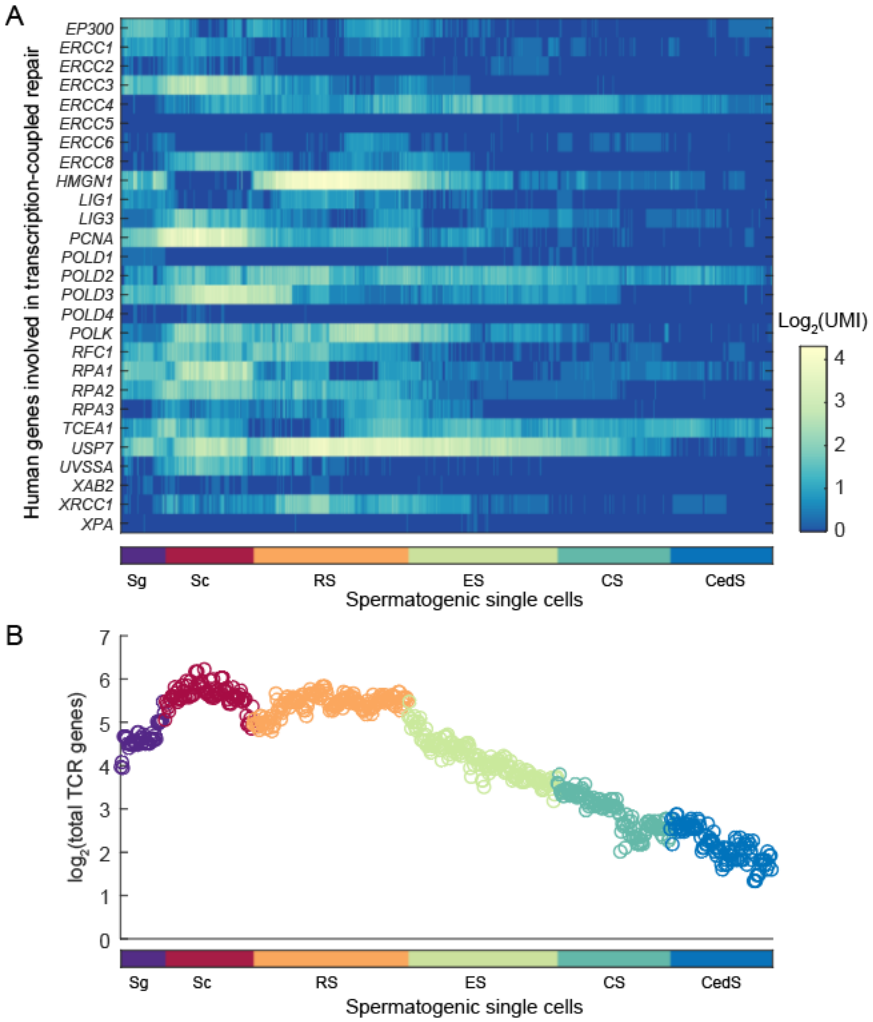

Fig. S7. Gene expression profiles of genes involved in transcription-coupled repair (TCR). Gene expression levels of each TCR gene (A) and their sum (B) across all spermatogenic single cells are displayed, respectively.

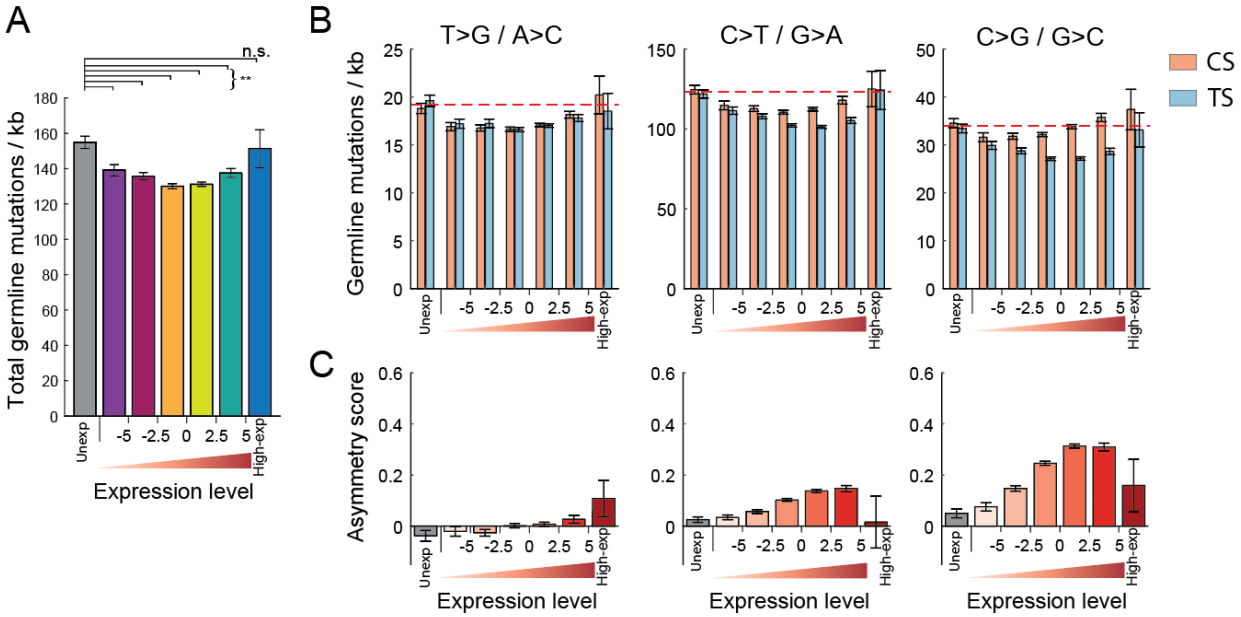

Fig. S8. 'Transcriptional scanning'-induced mutation reduction is tuned by gene-expression level. (A) Germline mutation rates across gene expression level categories. Spermatogenesis unexpressed- or highly expressed genes have higher level of germline mutations. (B-C) Same as Fig. 4C and D, showing more mutation types. Germline mutation rates (B) and associated asymmetry scores (C) of the indicated mutation types across gene expression level categories as determined in Fig. 4B. Dashed lines in (B) indicate the average level of mutations in unexpressed genes.

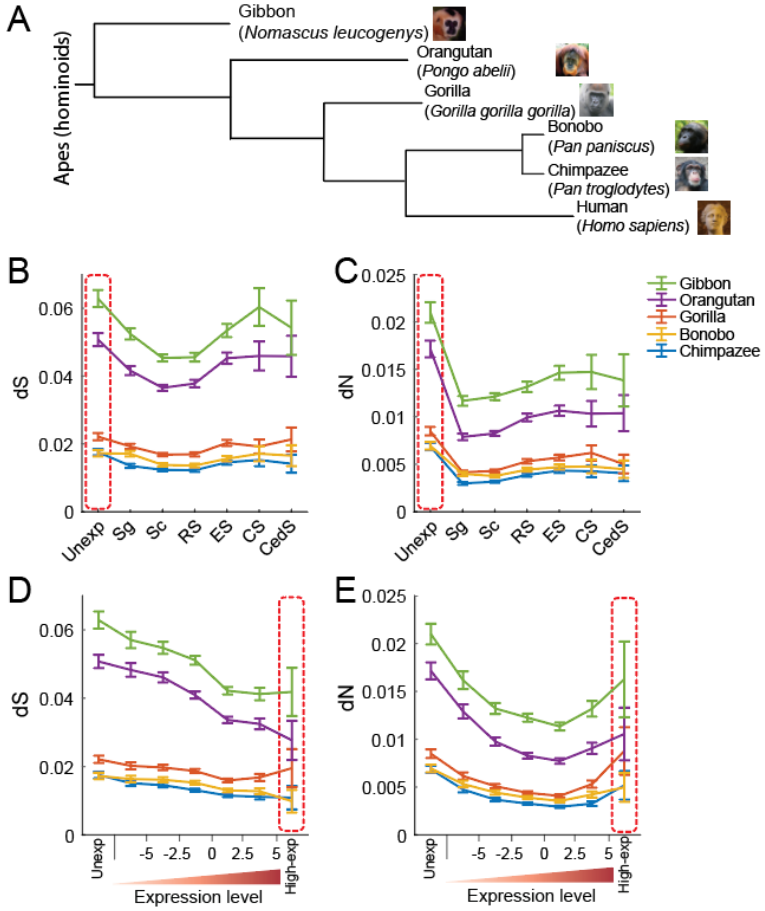

Fig. S9. Evolutionary consequences of 'transcriptional scanning' across apes. (A) Phylogenic tree of apes with sequenced genome data in Ensembl [31]. (B-C) dN (B) and dS (C) values of human genes with their orthologues across apes, according to stages of spermatogenesis expression. Red dashed box highlights the unexpressed genes. (D-E) Same as B-C, according to gene expression level categories.

Table S1. Gene Ontology (GO) terms showing enrichment in the set of genes unexpressed in spermatogenesis. The GO term analysis was done by GOrilla [67]. 'FDR q-value' is the correction of p-values for multiple testing using the Benjamini and Hochberg method [68]. Enrichment (N, B, n, b) is defined as 'Enrichment = (b/n) / (B/N)'. N, total number of genes; B, total number of genes associated with a specific GO term; n, number of genes in the input list; b, number of genes in the intersection. The highlighted GO terms are displayed in Fig. 5C.

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0050907	detection of chemical stimulus involved in sensory perception	3.88E-134	5.58E-130	5.54	15260	369	1793	240
GO:0007186	G-protein coupled receptor signaling pathway	1.48E-133	1.07E-129	3.43	15260	1029	1793	415
GO:0009593	detection of chemical stimulus	1.42E-132	6.83E-129	5.34	15260	394	1793	247
GO:0050906	detection of stimulus involved in sensory perception	1.98E-127	7.13E-124	5.15	15260	408	1793	247
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	1.18E-123	3.39E-120	5.62	15260	330	1793	218
GO:0051606	detection of stimulus	6.82E-107	1.64E-103	4.22	15260	529	1793	262
GO:0031424	keratinization	6.42E-45	1.32E-41	4.72	15260	175	1793	97
GO:0007165	signal transduction	6.02E-39	1.08E-35	1.52	15260	3825	1793	683
GO:0007606	sensory perception of chemical stimulus	3.28E-33	5.25E-30	4.28	15260	159	1793	80
GO:0007608	sensory perception of smell	8.62E-33	1.24E-29	4.73	15260	126	1793	70
GO:0050896	response to stimulus	3.58E-32	4.67E-29	1.44	15260	4167	1793	705
GO:0032501	multicellular organismal process	5.80E-28	6.95E-25	1.53	15260	2853	1793	513
GO:0006952	defense response	4.24E-25	4.69E-22	2.07	15260	824	1793	200
GO:0006955	immune response	4.51E-21	4.63E-18	2.06	15260	699	1793	169
GO:0098542	defense response to other organism	1.26E-20	1.21E-17	2.64	15260	319	1793	99
GO:0007600	sensory perception	8.91E-20	8.01E-17	2.27	15260	476	1793	127
GO:0003008	system process	9.19E-18	7.77E-15	1.73	15260	1112	1793	226
GO:0010469	regulation of receptor activity	2.05E-17	1.64E-14	2.21	15260	455	1793	118
GO:0051707	response to other organism	7.01E-16	5.30E-13	2.19	15260	424	1793	109
GO:0050877	nervous system process	1.01E-15	7.24E-13	1.87	15260	730	1793	160
GO:0045087	innate immune response	6.50E-15	4.45E-12	2.26	15260	358	1793	95
GO:0002323	natural killer cell activation involved in immune response	1.17E-13	7.62E-11	7.23	15260	20	1793	17
GO:0043207	response to external biotic stimulus	1.28E-13	8.02E-11	1.89	15260	600	1793	133
GO:0042742	defense response to bacterium	1.56E-13	9.33E-11	2.79	15260	174	1793	57
GO:0006959	humoral immune response	3.06E-13	1.76E-10	3.05	15260	134	1793	48

GO:0009607	response to biotic stimulus	3.79E-13	2.10E-10	1.85	15260	627	1793	136
GO:0033141	positive regulation of peptidyl-serine phosphorylation of STAT protein	5.44E-13	2.90E-10	6.89	15260	21	1793	17
GO:0009617	response to bacterium	6.15E-13	3.16E-10	2.63	15260	194	1793	60
GO:0033139	regulation of peptidyl-serine phosphorylation of STAT protein	2.13E-12	1.06E-09	6.58	15260	22	1793	17
GO:0002376	immune system process	2.69E-12	1.29E-09	1.47	15260	1587	1793	275
GO:0050912	detection of chemical stimulus involved in sensory perception of taste	4.96E-12	2.30E-09	5.06	15260	37	1793	22
GO:0001580	detection of chemical stimulus involved in sensory perception of bitter taste	5.37E-12	2.41E-09	5.26	15260	34	1793	21
GO:0070268	cornification	3.83E-11	1.67E-08	3.07	15260	108	1793	39
GO:0030101	natural killer cell activation	4.22E-11	1.79E-08	4.68	15260	40	1793	22
GO:0006954	inflammatory response	6.47E-11	2.66E-08	2.05	15260	349	1793	84
GO:0009605	response to external stimulus	1.23E-10	4.90E-08	1.55	15260	1067	1793	194
GO:0007187	G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	3.17E-10	1.23E-07	2.57	15260	159	1793	48
GO:0007218	neuropeptide signaling pathway	3.78E-09	1.43E-06	2.96	15260	95	1793	33
GO:0042100	B cell proliferation	3.57E-08	1.32E-05	4.54	15260	30	1793	16
GO:0018149	peptide cross-linking	5.64E-08	2.03E-05	3.47	15260	54	1793	22
GO:0051607	defense response to virus	6.76E-08	2.37E-05	2.44	15260	136	1793	39
GO:0007200	phospholipase C-activating G-protein coupled receptor signaling pathway	9.40E-08	3.22E-05	3.03	15260	73	1793	26
GO:0002250	adaptive immune response	1.47E-07	4.91E-05	2.35	15260	145	1793	40
GO:0043330	response to exogenous dsRNA	2.34E-07	7.66E-05	3.74	15260	41	1793	18
GO:1904892	regulation of STAT cascade	2.75E-07	8.78E-05	2.39	15260	132	1793	37
GO:0002286	T cell activation involved in immune response	3.50E-07	1.09E-04	3.52	15260	46	1793	19
GO:0007188	adenylate cyclase-modulating G-protein coupled receptor signaling pathway	3.84E-07	1.17E-04	2.33	15260	139	1793	38
GO:0050832	defense response to fungus	5.06E-07	1.52E-04	4.12	15260	31	1793	15
GO:0043331	response to dsRNA	5.22E-07	1.53E-04	3.44	15260	47	1793	19
GO:0019221	cytokine-mediated signaling pathway	6.24E-07	1.79E-04	1.69	15260	434	1793	86
GO:0046425	regulation of JAK-STAT cascade	6.72E-07	1.89E-04	2.34	15260	131	1793	36
GO:0097746	regulation of blood vessel diameter	1.26E-06	3.50E-04	2.59	15260	92	1793	28

GO:0035296	regulation of tube diameter	1.26E-06	3.43E-04	2.59	15260	92	1793	28
GO:0051480	regulation of cytosolic calcium ion concentration	2.46E-06	6.55E-04	1.92	15260	231	1793	52
GO:0042110	T cell activation	2.49E-06	6.51E-04	2.17	15260	149	1793	38
GO:0050878	regulation of body fluid levels	2.66E-06	6.82E-04	1.78	15260	306	1793	64
GO:0001906	cell killing	2.68E-06	6.76E-04	3.04	15260	56	1793	20
GO:0035150	regulation of tube size	3.07E-06	7.60E-04	2.44	15260	101	1793	29
GO:0070098	chemokine-mediated signaling pathway	3.67E-06	8.95E-04	2.99	15260	57	1793	20
GO:0007204	positive regulation of cytosolic calcium ion concentration	3.78E-06	9.07E-04	1.96	15260	204	1793	47
GO:0007267	cell-cell signaling	3.82E-06	8.99E-04	1.57	15260	527	1793	97
GO:0052695	cellular glucuronidation	4.82E-06	1.12E-03	5.47	15260	14	1793	9
GO:0061844	antimicrobial humoral immune response mediated by antimicrobial peptide	5.28E-06	1.21E-03	3.55	15260	36	1793	15
GO:0060337	type I interferon signaling pathway	6.11E-06	1.37E-03	2.99	15260	54	1793	19
GO:0050880	regulation of blood vessel size	7.68E-06	1.70E-03	2.38	15260	100	1793	28
GO:0009620	response to fungus	7.92E-06	1.73E-03	3.45	15260	37	1793	15
GO:0052697	xenobiotic glucuronidation	8.88E-06	1.90E-03	6.62	15260	9	1793	7
GO:0052696	flavonoid glucuronidation	8.88E-06	1.88E-03	6.62	15260	9	1793	7
GO:0050830	defense response to Gram-positive bacterium	9.39E-06	1.96E-03	2.75	15260	65	1793	21
GO:0002252	immune effector process	9.87E-06	2.03E-03	1.47	15260	678	1793	117

Table S2. Gene Ontology terms showing enrichment in the set of genes that are highly-expressed throughout spermatogenesis. The GO term analysis was done as described in Table S1.

GO Term	Description	P-value	FDR q-value	Enrichment	N	B	n	b
GO:0022414	reproductive process	4.58E-10	6.77E-06	3.58	16863	1249	113	30
GO:0022900	electron transport chain	7.20E-09	5.32E-05	10.59	16863	155	113	11
GO:0022412	cellular process involved in reproduction in multicellular organism	1.22E-08	6.01E-05	7.01	16863	298	113	14
GO:0022904	respiratory electron transport chain	1.43E-08	5.29E-05	14.14	16863	95	113	9
GO:0006091	generation of precursor metabolites and energy	2.46E-08	7.26E-05	6.63	16863	315	113	14
GO:0003006	developmental process involved in reproduction	2.97E-08	7.31E-05	4.83	16863	556	113	18
GO:0048609	multicellular organismal reproductive process	1.26E-06	2.65E-03	4.47	16863	501	113	15
GO:0007276	gamete generation	1.77E-06	3.28E-03	5.05	16863	384	113	13
GO:0007283	spermatogenesis	4.11E-06	6.75E-03	5.1	16863	351	113	12
GO:0048232	male gamete generation	4.36E-06	6.44E-03	5.07	16863	353	113	12

References

1. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–263 (2009).

2. Sonawane, A. R. *et al.* Understanding Tissue-Specific Gene Regulation. *Cell Rep.* **21,** 1077–1088 (2017).

3. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).

4. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (80-.).* **348,** 648–660 (2015).

5. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science (80-.).* **348,** 660–665 (2015).

6. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science (80-.).* **347,** 1260419 (2015).

7. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550,** 451–453 (2017).

8. Spiller, C., Koopman, P. & Bowles, J. Sex determination in the mammalian germline. *Annu. Rev. Genet.* **51,** 265–285 (2017).

9. Khaitovich, P., Enard, W., Lachmann, M. & Pääbo, S. Evolution of primate gene expression. *Nat. Rev. Genet.* **7,** 693–702 (2006).

10. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3,** 2179–2190 (2013).

11. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell Proteomics* **13,** 397–406 (2014).

12. Schmidt, E. E. Transcriptional promiscuity in testes. *Curr. Biol.* **6,** 768–769 (1996).

13. Naro, C. *et al.* An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. *Dev. Cell* **41,** 82–93.e4 (2017).

14. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3,** 346–360.e4 (2016).

15. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14,** 955–958 (2017).

16. Lake, B. B. *et al.* Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science (80-.).* **352,** 1586–1590 (2016).

17. Miyata, H. *et al.* Genome engineering uncovers 54 evolutionarily conserved and testis-enriched genes that are not required for male fertility in mice. *Proc. Natl. Acad. Sci. USA* **113,** 7704–7710 (2016).

18. Rathke, C., Baarends, W. M., Awe, S. & Renkawitz-Pohl, R. Chromatin dynamics during spermiogenesis. *Biochim. Biophys. Acta* **1839,** 155–168 (2014).

19. Necsulea, A. & Kaessmann, H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* **15,** 734–748 (2014).

20. Sassone-Corsi, P. Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science (80-.).* **296,** 2176–2178 (2002).

21. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: two decades of progress and surprises. *Nat. Rev. Mol. Cell Biol.* **9,** 958–970 (2008).

22. Vermeulen, W. & Fousteri, M. Mammalian transcription-coupled excision repair. *Cold Spring Harb. Perspect. Biol.* **5,** a012625 (2013).

23. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11,** 47–59 (2010).

24. Boehm, T. Evolution of vertebrate immunity. *Curr. Biol.* **22,** R722–32 (2012).

25. Singh, R. S., Xu, J. & Kulathinal, R. J. *Rapidly evolving genes and genetic systems*. (books.google.com, 2012).

26. Kanatsu-Shinohara, M. & Shinohara, T. Spermatogonial stem cell self-renewal and development. *Annu. Rev. Cell Dev. Biol.* **29,** 163–187 (2013).

27. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).

28. La Manno, G. *et al.* RNA velocity in single cells. *BioRxiv* (2017). doi:10.1101/206052

29. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14,** 979–982 (2017).

30. Jinks-Robertson, S. & Bhagwat, A. S. Transcription-associated mutagenesis. *Annu. Rev. Genet.* **48,** 341–359 (2014).

31. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46,** D754–D761 (2018).

32. Haradhvala, N. J. *et al.* Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164,** 538–549 (2016).

33. Makova, K. D. & Li, W.-H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416,** 624–626 (2002).

34. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet.* **29,** 575–584 (2013).

35. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* **17,** 241 (2016).

36. Tubbs, A. & Nussenzweig, A. Endogenous DNA damage as a source of genomic instability in cancer. *Cell* **168,** 644–656 (2017).

37. Xu, G. *et al.* Nucleotide excision repair activity varies among murine spermatogenic cell types. *Biol. Reprod.* **73,** 123–130 (2005).

38. Chen, C., Qi, H., Shen, Y., Pickrell, J. & Przeworski, M. Contrasting determinants of mutation rates in germline and soma. *Genetics* **207,** 255–267 (2017).

39. Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.* **33,** 514–517 (2003).

40. Mugal, C. F., von Grünberg, H.-H. & Peifer, M. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol. Biol. Evol.* **26,** 131–142 (2009).

41. McVicker, G. & Green, P. Genomic signatures of germline gene expression. *Genome Res.* **20,** 1503–1511 (2010).

42. Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat. Rev. Genet.* **14,** 880–893 (2013).

43. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (80-.).* **322,** 1845–1848 (2008).

44. Duttke, S. H. C. *et al.* Human promoters are intrinsically directional. *Mol. Cell* **57,** 674–684 (2015).

45. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (80-.).* **352,** aad9926 (2016).

46. Park, C., Qian, W. & Zhang, J. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep.* **13,** 1123–1129 (2012).

47. Roos, W. P. & Kaina, B. DNA damage-induced cell death by apoptosis. *Trends Mol. Med.* **12,** 440–450 (2006).

48. Barnes, D. E. & Lindahl, T. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu. Rev. Genet.* **38,** 445–476 (2004).

49. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549,** 519–522 (2017).

50. Hansen, R. S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. USA* **107,** 139–144 (2010).

51. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482 (2011).

52. Cleaver, J. E. Transcription coupled repair deficiency protects against human mutagenesis and carcinogenesis: Personal Reflections on the 50th anniversary of the discovery of xeroderma pigmentosum. *DNA Repair (Amst)* **58,** 21–28 (2017).

53. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science (80-.).* **359,** 555–559 (2018).

54. Valli, H. *et al.* Fluorescence- and magnetic-activated cell sorting strategies to isolate and enrich human spermatogonial stem cells. *Fertil. Steril.* **102,** 566–580.e7 (2014).

55. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17,** 77 (2016).

56. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

57. Wagner, F., Yan, Y. & Yanai, I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *BioRxiv* (2017). doi:10.1101/217737

58. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11,** 637–640 (2014).

59. Mali, P. *et al.* Stage-specific expression of nucleoprotein mRNAs during rat and mouse spermiogenesis. *Reprod Fertil Dev* **1,** 369–382 (1989).

60. Potter, S. J. & DeFalco, T. Role of the testis interstitial compartment in spermatogonial stem cell function. *Reproduction* **153,** R151–R162 (2017).

61. Chang, Y.-F., Lee-Chang, J. S., Panneerdoss, S., MacLean, J. A. & Rao, M. K. Isolation of Sertoli, Leydig, and spermatogenic cells from the mouse testis. *BioTechniques* **51,** 341–2, 344 (2011).

62. Ye, L., Li, X., Li, L., Chen, H. & Ge, R.-S. Insights into the Development of the Adult Leydig Cell Lineage from Stem Leydig Cells. *Front. Physiol.* **8,** 430 (2017).

63. Buganim, Y. *et al.* Direct reprogramming of fibroblasts into embryonic Sertoli-like cells by defined factors. *Cell Stem Cell* **11,** 373–386 (2012).

64. Von Kopylow, K. & Spiess, A.-N. Human spermatogonial markers. *Stem Cell Res.* **25,** 300–309 (2017).

65. Djureinovic, D. *et al.* The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Mol. Hum. Reprod.* **20,** 476–488 (2014).

66. Yan, W., Ma, L., Burns, K. H. & Matzuk, M. M. HILS1 is a spermatid-specific linker histone H1-like protein implicated in chromatin remodeling during mammalian spermiogenesis. *Proc. Natl. Acad. Sci. USA* **100,** 10546–10551 (2003).

67. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10,** 48 (2009).

68. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.