Gene expression levels tune germline mutation rates through the compound effects of transcription-coupled repair and damage

Bo Xia
NYU Langone Health
Itai Yanai (itai.yanai@nyulangone.org)
NYU Langone Health https://orcid.org/0000-0002-8438-2741

Research Article

Keywords: TCR, population-wide, hypothesis, transcriptional scanning

Posted Date: May 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-540111/v1

Abstract

Of all mammalian organs, the testis has long been observed to have the most diverse gene expression profile. To account for this widespread gene expression, we have proposed a mechanism termed 'transcriptional scanning', which reduces germline mutation rates through transcription-coupled repair (TCR). Our hypothesis contrasts with an earlier observation that mutation rates are overall positively correlated with gene expression levels in yeast, implying that transcription is mutagenic due to transcription-coupled damage (TCD). Here we report evidence that the compound effects of both TCR and TCD during spermatogenesis tune germline mutation rates in human, with TCR dominating in most genes, thus supporting the transcriptional scanning hypothesis. Our analyses address potentially confounding factors, distinguish the differential mutagenic effects acting on the highly expressed genes and the low-to-moderately expressed genes, and resolve concerns relating to the validation of the results using a *de novo* mutation dataset. We also discuss the theoretical possibility of transcriptional scanning hypothesis from an evolutionary perspective. Together, these analyses support a model by which the coupling of transcription-coupled repair and damage establishes the pattern of germline mutation rates and provide an evolutionary explanation for widespread gene expression during spermatogenesis.

Introduction

It has been known for many years that gene expression in the testis accounts for the largest number of genes among mammalian organs (Schmidt and Schibler 1995; Soumillon et al. 2013; Melé et al. 2015). However, how this widespread gene expression pattern is established, and the biological role that it plays during spermatogenesis, has been a long standing question in molecular biology (Schmidt 1996). We recently proposed a mechanism – which we termed 'transcriptional scanning' – in which widespread gene expression during spermatogenesis reduces germline mutation rates through transcription-coupled DNA repair (TCR), thus safeguarding sequence integrity of spermatogenesis-expressed genes (Xia et al. 2020). By analyzing germline mutation rates from population-wide SNP database and a *de novo* mutation (DNM) dataset, we 1) observed that genes expressed in male germ cells have lower germline mutation rates than those of unexpressed genes; 2) inferred that the genes that benefit from transcriptional scanning have lower mutation rates on the template-strand than on the coding strand, which we attributed to the result of germline transcription-coupled repair (TCR); 3) hypothesized that the germline mutation rate for a gene is tuned by its level of expression during spermatogenesis through the compound effects of TCR and transcription-coupled damage (TCD); and 4) found that the observed pattern of biased mutation rates is also recapitulated by a comparison between human and apes, supporting the notion that the cumulative effect on germline mutation rates could be observed over evolutionary time scales (Xia et al. 2020). These observations indicate that widespread gene expression during spermatogenesis enables large-scale genome scanning by the DNA repair machinery, thus reducing germline mutation rates through TCR. Together, the transcriptional scanning hypothesis provides a compelling model to understand the biological meaning for widespread gene expression in the male germ cells (Xia et al. 2020).

In this manuscript, we dissect the compound effects of transcription-coupled repair and damage according to gene expression levels during spermatogenesis. We have performed correlation analyses between spermatogenesis gene expression levels and germline mutation rates following appropriate assumptions and incorporation of possible confounding factors (Liu and Zhang 2020). Contrary to a monotonic positive correlation between gene expression level and mutation rate predicted by the TAM model (Liu and Zhang 2020), our results are most consistent with a model that transcription during spermatogenesis in human tunes germline mutation rates in a compound effect: TCR dominates in the low-to-moderately expressed genes, representing most of genes, while TCD gradually overwhelms TCR in the small group of highly expressed genes. We also validated the results using the a *de novo* mutation (DNM) dataset. The results from our analysis of germline mutation rates – presented here with rectified assumptions and accounting for confounding factors – provide additional support for the transcriptional scanning hypothesis. We conclude that while the TAM model holds for the most highly expressed genes, transcriptional scanning provides a more comprehensive model that includes also the low-to-moderately expressed genes and an evolutionary explanation for widespread expression in the germline.

Results

The SNP and DNM datasets provide consistent mutational patterns

We previously reported consistent results of the predictions of transcriptional scanning across two datasets (Xia et al. 2020): the SNPs from the dbSNP150 database and *de novo* mutations from DNM datasets based on parents-child trio-genome sequencing (Jónsson et al. 2017; An et al. 2018; Xia et al. 2020). In preprocessing the DNM dataset for analysis, we excluded the zero-variant genes (genes with no detected mutations), which Liu and Zhang have argued as unjustified (Liu and Zhang 2020). However, there is a clear rationale for excluding the zero-variant genes as these introduce a strong bias into mutation rate inference. To demonstrate this effect, we first compared the mutation rates distribution and mutational signatures across spermatogenesis expression levels by including or excluding zero-variant genes, using both the SNP and DNM datasets (Fig. S1). We found that the inclusion or exclusion does not change the mutation rate distribution when working with SNP datasets (dbSNP150 or 1000 Genome project, Fig. S1A-D; See Methods), nor does it change the mutational signatures along expression levels (Fig. S1G-H). The salient point regarding the DNM dataset is that it is > 1000-fold sparser than the SNP dataset (Fig. 2A), because of the natural sparsity of *de novo* mutations when only sequencing a limited number of genomes. Consequently, there is a strong 'zero-inflated' mutation rate profile at the gene level (Fig. S1E-F), such that genes with no observed mutations are dragged down to a significant underestimate of their mutation rate (Fig. S1E-F).

To demonstrate how zero-variant genes introduce a bias into the analysis of the DNMs, we simulated down-samplings of the SNP dataset. When down-sampling the SNP dataset to the scale of the DNM dataset and then calculating the mutation rate per gene per kilobase, we found that removing the zero-

variant genes leads to a dramatically improved estimation of mutation rates, relative to their inclusion (Fig. 2B-C). When plotting the mutational signatures across gene expression levels for these down-sampled SNP datasets that include the zero-variant genes, we indeed found the same flip in the relationship that Liu and Zhang indicate for the DNM dataset (Fig. 2D and S1I, left) (Liu and Zhang 2020). However, removing the zero-variant genes in the down-sampled SNP dataset maintained the same mutation pattern as that of the original SNP dataset (Fig. 2E and S1I, right), consistent with the results from the original SNP dataset (Fig. S1G). Thus, we found that removing zero-variant genes reduces the overall biases when analyzing DNM mutation rate at the gene level.

Controlling for potential confounding factors in the relationship between germline gene expression levels and mutation rates

The transcriptional scanning hypothesis is based on the observation of reduced germline mutation rates in the spermatogenesis-expressed genes, as well as multiple mutational signatures derived from the transcriptional events during spermatogenesis (Xia et al. 2020). Liu and Zhang argue that our analysis did not control for potentially confounding factors including replication timing, GC content and nucleosome occupancy rate (Liu and Zhang 2020). It is indeed crucial to consider potentially confounding factors in studying the correlation between spermatogenesis gene expression levels and germline mutation rates. In this section, we discuss each of the factors introduced by in the TAM model (Liu and Zhang 2020), and in the next section we revisit the relationship between germline gene expression and mutation rates.

In particular, we note several errors and inappropriate assumptions made while Liu and Zhang reexamined our work (Liu and Zhang 2020; Xia et al. 2020). The authors mistakenly used the entire raw gene expression data as opposed to restricting analysis to germ cells, by averaging gene expression across the raw unique molecular identifier (UMI) matrix of 8,000 cell barcodes, which in fact included 5906 empty barcodes and non-germ cells (Liu and Zhang 2020; Xia et al. 2020). InDrop single-cell RNA-seq experimentally introduces far more cell-barcoding beads than cells to ensure individual barcoding, and subsequent computational pre-processing steps filter out such extra and low-quality cell barcodes (Klein et al. 2015). Our original analysis retained 2554 high-quality human testicular cells, of which 2094 cells were annotated as germ cells and their gene expression levels were used for the germline mutation rate analysis. Liu and Zhang did not describe this aspect of the analysis in their manuscript, nor did they publish their code, however, we confirmed this aspect by reanalyzing the dataset provided by the authors (See Methods & code).

Second, it has been argued that nucleosome occupancy should be controlled for when studying the correlation between gene expression and mutation rates. We argue that controlling for nucleosome occupancy in the mutation rates analysis at the whole-gene level is not appropriate. Nucleosome occupancy mainly influences mutation rate at the local genomic scale of a few hundred nucleotides (Gonzalez-Perez et al. 2019). Transcriptional scanning is proposed to affect mutation rate at full gene loci, with an average length of 67,537 nucleotides, which is far beyond the scale where nucleosome

occupancy could affect the mutation rates locally. More importantly, nucleosome occupancy is closely related to the molecular mechanism which contributes to the widespread gene expression present during spermatogenesis, while also being influenced by transcription which generates nucleosome-free intermediates (Lai and Pugh 2017). In other words, the biased mutation rates may indeed be (partially) influenced by such molecular events as nucleosome occupancy, as these would lead to differential transcription rates. Thus, since nucleosome occupancy may be a part of the mechanism by which transcriptional scanning occurs it is not appropriate to control for it as a possible confounding factor in the correlation analysis between expression level and gene level mutation rates.

Furthermore, Liu and Zhang did not use nucleosome occupancy data correctly. They invoked nucleosome occupancy profiles from mature sperm samples to control for germline mutation rate analysis (Liu and Zhang 2020), however the transcriptional scanning hypothesis pertains to spermatogenic cells where transcription is still active (Xia et al. 2020). Spermatids replace more than 95% of their histones to protamine during spermiogenesis, shutting down transcription almost completely (Johnson et al. 2016; Hao et al. 2019). The remaining < 5% of histone-occupied regions in mature sperms could not represent the nucleosome occupation pattern of transcription-active cells (Johnson et al. 2016; Zhang et al. 2019).

Third, Liu and Zhang tested replication timing (RT) as a possible confounder in the germline mutation rates using a RT profile from a cancer cell line (HeLa) (Liu and Zhang 2020). Yehuda *et al* reported the uniqueness of replication timing in mouse male germ cells, and found that RT profile from correct tissue-of-origin correlates better with recombination hotspots and mutation rates (Yehuda et al. 2018). While it may thus not be ideal to use the RT profile from HeLa cells, RT profiles from human spermatogenic cells are currently not available, and Liu and Zhang used the former as an approximate. In our analysis below, we included RT as a possible confounder.

Last, the authors introduced GC content as a possible confounding factor in the germline mutation rate analysis (Liu and Zhang 2020). Indeed, previous reports found that GC content positively correlates with mutation rates (Kiktev et al. 2018). Analyzing germline mutations from SNP dataset, we found that the GC content positively correlates with mutation rates of C > D/G > H types, but not with A > B/T > V types (Fig. S2 A-B). In addition, we observed the strongest positive correlation between GC content and C > T/G > A mutation rates among all 12 mutation types (Fig. S3). C > T/G > A mutations almost exclusively derive from cytosine deamination-induced mutagenesis, and occur most frequently at CpG dinucleotides (Sassa et al. 2016). The observed positive correlation between GC content and C > T/G > A mutation rates likely follows from the increased frequency of CpG dinucleotides (Fig. S2C). In contrast, spermatogenesis gene expression levels negatively correlate with GC content (Figure S1E in Liu and Zhang, 2020), contradicting an earlier report that gene expression levels increase with high GC content (Kudla et al. 2006). These results indicate that controlling for GC content for gene-level mutation rate analyses is questionable and requires further study. We thus also included it in our analyses as a possible confounder.

Genes across expression level categories exhibit mutation rates consistent with the TS hypothesis.

With the above points considered, we repeated the correlation analysis using the corrected gene expression profile (from spermatogenic cells only) and germline mutation rates, controlling for the replication timing and/or GC content as potential confounding factors. We observed a negative correlation between spermatogenesis gene expression levels and germline mutation rates inferred from both SNP and DNM datasets, with or without controlling replication timing (Fig. 3A-B). Controlling for GC content as a potential confounding factor led to a weak positive correlation when using the SNP dataset, while a negative correlation was maintained using the DNM dataset (Fig. 3A-B). Our results differ from those of others (Liu and Zhang 2020) because of our preprocessing of the DNM dataset (Fig. 2) and of our consideration of RT alone, rather than as combined with GC content and nucleosome occupancy.

In contrast with the TAM model which proposes a monotonic effect of transcription on mutation rate (Liu and Zhang 2020), a key tenet of the transcriptional scanning hypothesis is that transcription leads to both TCR and TCD, tuning gene mutation rates according to transcription levels (Fig. 1A-B). TCR, which reduces mutation rates, dominates the net effect in the low-to-moderately expressed genes (Xia et al. 2020). However, in the highly expressed genes, the effect of TCD overwhelms the TCR effect, leading to an overall increase in mutation rates (Xia et al. 2020). Although we repeated the calculation of Spearman's correlation coefficients between mutation rates and gene expression level across all genes (Fig. 3A-B), as performed in the previous analysis (Liu and Zhang 2020), this approach does not correctly distinguish between the effects of TCR and TCD across the range of gene expression levels (Xia et al. 2020). These results indicate that splitting the genes according to their expression levels is required to reveal the correlation between mutation rates and gene expression levels.

We thus divided the genes into two major groups, low-to-moderately expressed genes (70% of all genes), and highly expressed genes (30% of all genes) (Fig. 3C-D), and calculated their correlation coefficients between expression levels and mutation rates, respectively. In contrast to the monotonic positive correlation predicted by the TAM model (Liu and Zhang 2020), we observed a strong negative correlation between germline gene expression level and mutation rate in the low-to-moderately expressed genes, and a positive correlation in the highly expressed genes (Fig. 3E). Controlling for replication timing and/or GC ratio does not change the trend of this correlation (Fig. 3E). We also repeated this analysis using the DNM dataset (with removal of zero-variant genes as discussed above), and observed consistent results (Fig. 3F). Together, these results highlight the dual effects of transcription on gene mutation rates, that TCR dominates the effect in the low-to-moderately expressed genes, representing most genes, while TCD gradually overwhelms TCR effect in the highly expressed genes, leading to a positive correlation between mutation rate and expression level.

Comparison between coding and template strand mutation rates

Another piece of evidence supporting the transcriptional scanning hypothesis is the inference of a lower germline mutation rate in the template strand (transcribed strand) relative to the coding strand (non-transcribed strand) (Xia et al. 2020). This observation is further supported by a bidirectional transcription-

induced asymmetric signature, as well as a fine-tuning of mutation rate asymmetry by transcription levels (Xia et al. 2020). While it is usually not possible to determine the exact damage/mutation that induced an observed germline mutation/SNP, we adapted an indirect way to infer the mutation rate asymmetry between coding and template strands (Haradhvala et al. 2016). Using A-to-G/T-to-C mutation as example, the A > G mutations in a gene region is actually a mixture of two types of mutations (Fig. S4A): A-to-G mutation/damage occurring on the coding strand ($A > G^{coding}$) and T-to-C mutation/damage occurring on the transcribed strand ($T > C^{template}$). Through DNA replication, the two types of unrepaired DNA damage will introduce the same A > G mutations (with reference to the coding strand). Similarly, the observed T > C mutations in a gene region originate from both A-to-G mutations occurring on the template strand ($A > G^{template}$) and T-to-C mutations occurring on the coding strand ($T > C^{coding}$) (Fig. S4B). Because of this uncertainty, the analysis of strand-specific mutation rate is performed indirectly as 6 pairs, instead of 12 mutation types. The null expectation is that the A-to-G/T-to-C mutation rates are equal between the coding strand and template strands, meaning that the observed A-to-G mutation rate ($A > G^{coding} + T > C^{template}$) should equal the observed T-to-C mutation rate ($A > G^{template} + T > C^{coding}$) (Fig. S4C). A significant difference between these two would indicate an asymmetric mutation rate between the strands, presumably dues to transcription-associated activities (Fig. S4C).

To further test for the dependency of asymmetric mutation rates between strands and transcription-associated activities, we predict that of the two rates, the higher would exhibit a positive correlation with gene expression level, accounting for the A-to-G mutation rates on coding strand in the example, while the lower would exhibit a negative correlation, accounting for the A-to-G mutation rates on template strand (Fig. S4C). This prediction also relies on prior studies that TCR would reduce mutation rates (Hanawalt and Spivak 2008), while transcription itself could increase mutation rates due to TCD (Gaillard and Aguilera 2016). Together, this analytic approach has been well-accepted for analyzing somatic mutations (Haradhvala et al. 2016), and was adapted in our analysis of the transcriptional scanning hypothesis (Xia et al. 2020).

The authors of the TAM model misrepresented the logic of the mutation rate asymmetry analysis by noting that the analysis in the transcriptional scanning manuscript assumed the mutation type with a higher rate is the original mutation and the complementary type is the subsequent change (Liu and Zhang 2020). This approach (Haradhvala et al. 2016), however, does not make this assumption and instead attempts to test whether the two mutation types (for example A > G/T > C) occur together at equal rates on the two strands.

The analysis into the mutation rates between coding strand and template strand further supported the dual effect of TCR and TCD on gene mutation rates. Consistent with the results in the TAM model (Liu and Zhang 2020), the inferred coding strand mutation rate (Fig. S4) positively correlates with spermatogenesis gene expression level using SNP dataset (Fig. 4A, left). In addition, we observed a consistent positive correlation in both low-to-moderately expressed genes and highly expressed genes, supporting a dominant effect of TCD on the coding strand throughout the expression level spectrum

(Fig. 4B). In the template strand, however, we observed a stronger negative correlation between the inferred mutation rates and expression level (Fig. 4A, right; see also Fig. S4). Splitting the genes into low-to-moderately expressed group and highly expressed group, we observed the strongest negative correlation between mutation rate and expression level in the low-to-moderately expressed genes (Fig. 4C, left), indicating more efficient TCR effects on this group. In contrast, a positive correlation was also observed between template strand mutation rate and expression level in the highly expressed gene group (Fig. 4C, right). These results are consistent with the notion that the TCD effect gradually overwhelms that of TCR in the highly expressed genes (Fig. 1B), as also observed from the mutation rate distribution profile in Fig. 5 of Xia et al (2020). Additionally, we observed overall consistent results when repeating these analyses using DNMs, though the correlation coefficients in the coding strand or highly expressed gene group were not significant, possibly because of the sparsity of *de novo* mutations (Fig. 4D-F). Together, these results again indicate the compound effect of TCD and TCR in the tuning the germline mutation rates on coding strand and template strand across gene expression levels, and refute the claim of a monotonic effect of gene expression increasing mutation rate (Liu and Zhang 2020).

Genes unexpressed during spermatogenesis and cross-species evolutionary rate analysis

The transcriptional scanning hypothesis proposes that most genes expressed during spermatogenesis incur an overall TCR effect, which acts to reduce their germline mutation rate. In contrast, genes unexpressed during spermatogenesis (~ 10% of all protein-coding genes) would not benefit from such a transcription-dependent process, thus leaving any existing DNA damages unrepaired in these genes as a source for new germline mutations. Studying the genes unexpressed during spermatogenesis revealed enrichment in functions related to environmental sensing, immune system, defense responses, and signal transduction (Xia et al. 2020). The tuning of germline mutation rates may also contribute to differential rates of gene sequence evolution, according to the neutral theory of evolution (Nei 2005; Wagner 2008).

The authors of the TAM model apparently misunderstood the working model of the transcriptional scanning hypothesis in this regard as evidenced by their note that there is ‘no need to invoke selection for high mutagenesis’ (Liu and Zhang 2020). It is true that close to half of the spermatogenesis-unexpressed genes are also not expressed in other adult tissues. However, this does not interfere with the interpretation that spermatogenesis-unexpressed genes have higher germline mutation rates than the expressed counterpart, potentially because of a lack of TCR effect for these unexpressed genes in the germ cells. To test this, in our original article we had divided the genes into four groups depending on their expression or not in male germ cells and in somatic tissues (Xia et al. 2020). We found that the spermatogenesis-unexpressed genes have higher germline mutation rates than the expressed genes, regardless of their expression or lack of expression in the somatic tissues (referring to Supplementary Fig. 3H-I of Xia et al. 2020). In addition, we found that only gene expression in the testis stands out among other tissues regarding predicting a higher germline mutation rate in the unexpressed genes than the expressed counterparts (referring to Fig. 2E of Xia et al. 2020).

Liu and Zhang additionally studied the dN/dS values of two particular groups of genes, those that are uniquely expressed in the testis but not in other tissues (Group I), and those that are unexpressed in the testis but expressed in other tissues (Group II) (Liu and Zhang 2020). They reported that Group I genes have higher dN/dS values than the Group II genes, leading them to infer less purifying selection on the Group I genes (Liu and Zhang 2020). The Group I genes mostly correspond to male reproduction genes and a group of testis-expressed olfactory receptor genes (Supplementary file 1). Previous studies have found that male reproduction-related genes are fast-evolving, likely due sexual selection (Wyckoff et al. 2000). Thus, it is not surprising that Group I genes have higher dN/dS values since this can at least partially be explained by the fast evolution of genes related to male reproduction. Moreover, dN/dS values are not directly relevant to the discussion of biased germline mutation rates. In fact, as we showed, we observed that similar Group I genes have lower germline mutation rates than the Group II genes (refer to Supplementary Fig. 3H-I of Xia et al. 2020). Thus, the analysis of the dN/dS values of the unexpressed genes during spermatogenesis in the TAM model manuscript is not a valid argument against the transcriptional scanning hypothesis.

Transcriptional scanning is theoretically tenable with a mechanism affecting many genes

In the TAM model, the authors implicitly assumed that the mechanism modifying mutation rates is independent across the majority of genes (Liu and Zhang 2020). In other words, a modifier (e.g. a specific mutation) may exclusively affect the expression level of a particular gene thus enabling the tuning of its potential for TAM. The authors extended this assumption to the transcriptional scanning hypothesis by expecting the modifier of gene expression levels to be unlinked across the vast majority of genes. Additionally, the authors argued that if the modifier for regulating transcription during spermatogenesis is gene-specific, then its coefficient of selection would not be strong enough to fix in the population (Liu and Zhang 2020). In our formulation of transcriptional scanning, however, we did not make the assumption of an independent-modifier. Liu and Zhang do concede that if regulation is not independent and occurred at the level of 10^3 or more genes, the mechanism is plausible given its sufficient selective strength (Liu and Zhang 2020). In other words, the transcriptional scanning hypothesis is theoretically tenable if the modifier influences the majority of genes.

While we did not propose a specific mechanism, we speculate that a common regulatory mechanism would allow for widespread transcriptional scanning, possibly by a dynamic chromatin remodeling during spermatogenesis. Instead of functioning at the individual gene level, we proposed that there are two broad classes of genes in spermatogenesis: the expressed (~ 18k genes) and unexpressed (~ 1.8k genes). Since both groups are large it is plausible that a simple regulatory mechanism accounts for their expression pattern – for example, expression of the vast majority of genes, and a special mechanism to repress the expression of the remaining set of 1.8k genes according to a common regulatory mechanism. A possible dynamic chromatin remodeling mechanism may influence gene expression of thousands of genes during spermatogenesis by spreading across the mitosis stage, and throughout the meiosis and spermiogenesis stages. These chromatin changes across the different stages possibly allow for many

genes (at the scale of 10^3) to be transcribed temporally, though their function may not be required for spermatogenesis. Because of the selective advantage for reducing germline mutation rates, such chromatin changes may have evolved to allow the observed widespread gene expression during spermatogenesis. As Liu and Zhang describe, $Ne*s$ is > 1 for regulatory mechanisms influencing groups of genes at the scale of thousands of genes (Liu and Zhang 2020), thus rendering transcriptional scanning evolutionarily plausible.

Conclusion

Here we have presented analyses on the relationship between gene expression during spermatogenesis and germline mutation rates, dissecting the compound effects of TCR and TCD. Unlike the TAM model of a monotonic positive correlation between gene expression and mutation rates (Park et al. 2012; Liu and Zhang 2020), we provide compelling evidence that the germline mutation rate is influenced by the compound effect of TCR and TCD. The net effect of these two leads to a decrease of mutation rates in most genes expressed during spermatogenesis. In particular, we found that controlling for potential confounding factors including replication timing and GC content does not change the overall trend of correlations.

Our results are consistent with the observation that somatic mutation rates negatively correlate with gene expression levels across multiple tumor types (Pleasance et al. 2010; Chapman et al. 2011; Lawrence et al. 2013; Supek and Lehner 2015). In addition, we showed that TCD is evident mostly in the highly expressed genes, gradually overwhelming the effect of TCR, thus leading to a positive correlation between expression level and mutation rates. The transcriptional scanning model thus provides a comprehensive view that transcription modulates gene mutation rates through a compound effect of TCR and TCD, in contrast to the TAM model which posits that gene expression monotonically leads to an increase in mutation rates (Park et al. 2012; Liu and Zhang 2020). We note that the support for the TAM model which comprises less than 200 mutations detected in the highly expressed genes in yeast (Park et al. 2012), may suffer from observation bias. Additionally, the difference between the evidence for the TAM model and the transcriptional scanning model may be partially explained by a difference in mutation rate/pattern between yeast and mammalian genomes, since the former contains mostly intron-less genes (Rodriguez-Medina and Rymond 1994).

Our transcriptional scanning hypothesis is premised on the notion that the TCR is related to the widespread transcription during spermatogenesis and functions to reduce germline mutation rates for most genes. This hypothesis is theoretically tenable since a single mechanism may function on large groups of genes. Early epigenetic analysis results have found broad open chromatin in the testis (Soumillon et al. 2013; Hammoud et al. 2014). These open chromatin states during the mitosis and meiosis stages may provide the chromatin basis for establishing the observed widespread gene expression. Throughout evolution, maintaining such chromatin dynamics could be selected for, thus enabling this pattern of expression and ultimately maintaining genomic sequence integrity for most

genes. This hypothesis could also explain the observation that many spermatogenesis-expressed genes are not functionally required for fertility (Miyata et al. 2016).

Leaky transcription is a competing hypothesis to account for widespread gene expression during spermatogenesis (Schmidt 1996; Necsulea and Kaessmann 2014; Rathke et al. 2014). According to this this hypothesis, spermatogenic cells contain widespread open chromatin regions, which may thus enable random transcriptional events (Soumillon et al. 2013; Hammoud et al. 2014). The main difference between the transcriptional scanning hypothesis and the leaky transcription hypothesis is that the latter only stands as a passive model. Widespread leaky transcription during spermatogenesis would increase the energy cost in the germ cells and would also be expected to increase the mutation rate, following the TAM model. If not for the beneficial effect of reducing mutation rate from gene transcription, we suggest that leaky expression would have been eliminated over evolutionary timescales. Thus, the notion of 'leaky transcription' is not as well supported as the transcriptional scanning hypothesis.

In summary, we have provided here further support for the observation that the compound effects of TCR and TCD tune germline mutation rates of genes according to their expression levels during spermatogenesis, with the former effect dominating for most genes and thus resulting in reduced germline mutation rates. A coordinated mechanism regulating the transcription at the scale of 10^3 genes could be evolved from the dramatic chromatin reorganization during spermatogenesis, indicating the theoretical plausibility of the transcriptional scanning hypothesis.

Materials And Methods

The code and corresponding datasets are available at GitHub:

https://github.com/xiabo821/TS_Reanalysis_scripts .

The gene expression levels of male germ cells and the clustering of genes are the same as used in Xia et al (2020). Specifically, we first extracted the germ cells out of the QC-qualified cells. Then single cells were clustered according to spermatogenesis stages, followed by averaging the UMI counts of cells by their corresponding stage (Xia et al. 2020). The final gene expression level of each gene was averaged from the mean expression across all spermatogenic stages.

DNA replication timing profile was provided by Liu and Zhang, and was originally measured in HeLa cells by Johnson et al. (2016). GC content was calculated directly from the base frequencies in the intron region of genes, as we previously defined in Xia et al. (2020). CpG dinucleotide rate of a gene was calculated as CpG counts per kilobase.

The germline variants dataset from dbSNP150, 1000Genome project or DNM datasets were previously processed (Xia et al. 2020). The mutation rate of a gene was defined as count of single-nucleotide variants (SNV) per kilobase, calculated by dividing SNV count by the reference base count and then

multiplied by 1000. Mutation rates on the coding strand and template strand were inferred as illustrated in Figure S4, and the approach was originally adapted from Haradhvala et al. (2016).

The down-sampling experiment of dbSNP150 variants to the scale of DNM was performed by first calculating the fold difference between genic variants in dbSNP150 dataset (~ 149M) and DNM dataset (~ 102K), resulting in a 1450 fold (1450X) difference. We first downgraded the dbSNP150 variants 145 folds, resulting in a 10X variant dataset. Then we randomly sampled 10% of the 10X dataset to give a down-sampled SNP dataset at the same scale of DNM. The Spearman's rank correlation coefficient between inferred mutation rates from down-sampled SNP dataset and the original SNP mutation rates was calculated, with or without the zero-variant genes, to estimate the mutation rate similarity between down-sampled SNP dataset and original SNP dataset.

While we explained the rationale for excluding zero-variant genes to Liu and Zhang, we regret not having made our reasoning more explicit in the original publication (Xia et al, 2020). We did mention a similar exclusion in the asymmetry score analysis which used mutation rates on both coding and template strands, but not the gene-level mutation rate analysis (see Methods section of Xia et al.). The complete code for analysis is available on Github since the initial submission of the original manuscript, but we do apologize for this oversight in properly describing the exclusion of genes with no detected *de novo* mutations.

Gene ontology (GO) term analysis were performed using GOrilla (Gene Ontology enRIchment anaLysis and visuaLizAtion) online tool (http://cbl-gorilla.cs.technion.ac.il/) (Eden et al. 2009). The target gene set was generated by its unique expression in the male germ cells but not in other tissues as determined from GTEx gene expression profile. The output GO terms and statistics were collected in the Supplementary File 1.

Declarations

ACKNOWLEDGEMENTS

We thank the authors of Liu and Zhang (2020) for providing their datasets used in the manuscript, as well as clarifying their method of analysis. We also thank Dalia Barkley of the Yanai lab for the constructive comments and suggestions to the manuscript. This work is supported by the NYU Grossman School of Medicine with funding to I.Y. B.X. is partially supported by a NYSTEM institutional training grant predoctoral fellowship (Contract #C322560GG).

ACKNOWLEDGEMENTS

We thank the authors of Liu and Zhang (2020) for providing their datasets used in the manuscript, as well as clarifying their method of analysis. We also thank Dalia Barkley of the Yanai lab for the constructive comments and suggestions to the manuscript. This work is supported by the NYU Grossman School of

Medicine with funding to I.Y. B.X. is partially supported by a NYSTEM institutional training grant predoctoral fellowship (Contract #C322560GG).

Funding: NYU Langone Health Start-up funds

Conflicts of interest/Competing interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Availability of data and material: Not applicable.

Code availability: https://github.com/xiabo821/TS_Reanalysis_scripts.

Authors' contributions: Conceptualization: Bo Xia and Itai Yanai; Methodology: Bo Xia and Itai Yanai; Formal analysis and investigation: Bo Xia; Writing - original draft preparation: Bo Xia and Itai Yanai; Writing - review and editing: Bo Xia and Itai Yanai; Funding acquisition: Itai Yanai; Supervision: Itai Yanai

Ethics approval: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

References

An J-Y, Lin K, Zhu L, Werling DM, Dong S, Brand H, Wang HZ, Zhao X, Schwartz GB, Collins RL, et al. 2018. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science 362.

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet J-P, Ahmann GJ, Adli M, et al. 2011. Initial genome sequencing and analysis of multiple myeloma. Nature 471:467–472.

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

Gaillard H, Aguilera A. 2016. Transcription as a threat to genome integrity. Annu. Rev. Biochem. 85:291–317.

Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. 2019. Local determinants of the mutational landscape of the human genome. Cell 177:101–114.

Hammoud SS, Low DHP, Yi C, Carrell DT, Guccione E, Cairns BR. 2014. Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. Cell Stem Cell 15:239–253.

Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and surprises. Nat. Rev. Mol. Cell Biol. 9:958–970.

Hao S-L, Ni F-D, Yang W-X. 2019. The dynamics and regulation of chromatin remodeling during spermiogenesis. Gene 706:201–210.

Haradhvala NJ, Polak P, Stojanov P, Covington KR, Shinbrot E, Hess JM, Rheinbay E, Kim J, Maruvka YE, Braunstein LZ, et al. 2016. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell 164:538–549.

Johnson GD, Jodar M, Pique-Regi R, Krawetz SA. 2016. Nuclease footprints in sperm project past and future chromatin regulatory events. Sci. Rep. 6:25864.

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, et al. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature 549:519–522.

Kiktev DA, Sheng Z, Lobachev KS, Petes TD. 2018. GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA 115:E7109–E7118.

Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. 4:e180.

Lai WKM, Pugh BF. 2017. Understanding nucleosome dynamics and their links to gene expression and DNA replication. Nat. Rev. Mol. Cell Biol. 18:548–562.

Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499:214–218.

Liu H, Zhang J. 2020. Higher Germline Mutagenesis of Genes with Stronger Testis Expressions Refutes the Transcriptional Scanning Hypothesis. Mol. Biol. Evol. 37:3225–3231.

Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, et al. 2015. The human transcriptome across tissues and individuals. Science 348:660–665.

Miyata H, Castaneda JM, Fujihara Y, Yu Z, Archambeault DR, Isotani A, Kiyozumi D, Kriseman ML, Mashiko D, Matsumura T, et al. 2016. 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.

Necsulea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding transcriptomes. Nat. Rev. Genet. 15:734–748.

Nei M. 2005. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22:2318–2342.

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

Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, et al. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463:191–196.

Rathke C, Baarends WM, Awe S, Renkawitz-Pohl R. 2014. Chromatin dynamics during spermiogenesis. Biochim. Biophys. Acta 1839:155–168.

Rodriguez-Medina JR, Rymond BC. 1994. Prevalence and distribution of introns in non-ribosomal protein genes of yeast. Mol. Gen. Genet. 243:532–539.

Sassa A, Kanemaru Y, Kamoshita N, Honma M, Yasui M. 2016. Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. Genes Environ. 38:17.

Schmidt EE, Schibler U. 1995. High accumulation of components of the RNA polymerase II transcription machinery in rodent spermatids. Development 121:2373–2383.

Schmidt EE. 1996. Transcriptional promiscuity in testes. Curr. Biol. 6:768–769.

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

Supek F, Lehner B. 2015. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature 521:81–84.

Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nat. Rev. Genet. 9:965–974.

Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature 403:304–309.

Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, Kim SY, Keefe DL, Alukal JP, Boeke JD, et al. 2020. Widespread transcriptional scanning in the testis modulates gene evolution rates. Cell 180:248–262.e21.

Yehuda Y, Blumenfeld B, Mayorek N, Makedonski K, Vardi O, Cohen-Daniel L, Mansour Y, Baror-Sebban S, Masika H, Farago M, et al. 2018. Germline DNA replication timing shapes mammalian genome composition. Nucleic Acids Res. 46:8299–8310.

Zhang M-Z, Cao X-M, Xu F-Q, Liang X-W, Fu L-L, Li B, Liu W-G, Li S-G, Sun F-Z, Huang X-Y, et al. 2019. In the human sperm nucleus, nucleosomes form spatially restricted domains consistent with programmed

nucleosome positioning. Biol. Open 8.

Supplemental Files

Supplementary file 1 is not available with this version

Figures

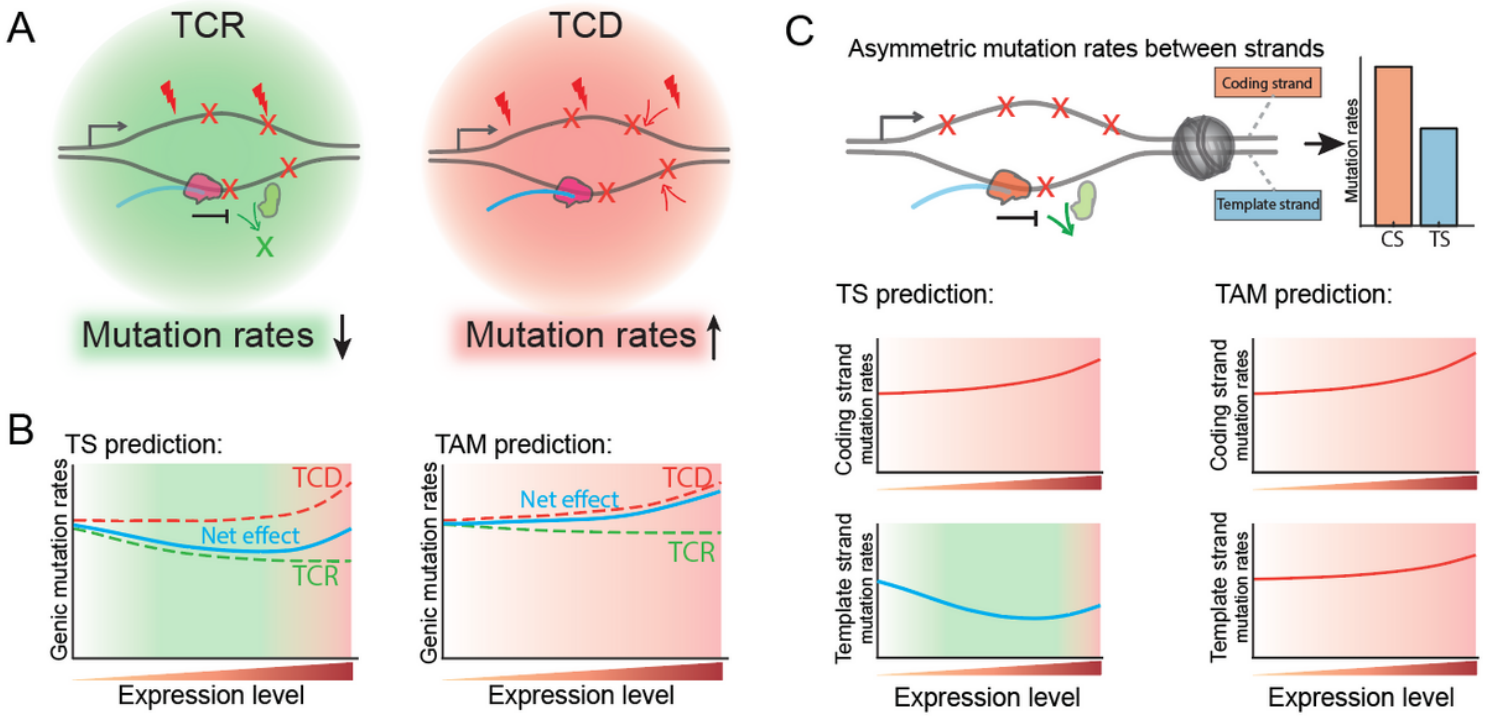

Figure 1

The distinct predictions of the transcriptional scanning (TS) hypothesis and the transcription-associated mutagenesis (TAM) model. (A) Schematic of TCR and TCD effects during transcription. TCR machinery detects and repairs existing DNA damages on the transcribed strand, thus reducing gene mutation rates; TCD results in increased mutagenesis when transcription machinery unwinds the DNA strands, making them relatively more venerable to cellular mutagens. (B) The TS hypothesis predicts a compound effect of TCR and TCD on mutation rates, tuned by expression level, while the TAM model predicts a monotonic positive correlation between gene expression level and mutation rates. (C) Schematic of asymmetric mutation rates between strands, and predictions from the TS and TAM models. The TS hypothesis predicts a positive correlation between gene expression levels and coding strand mutation rates, together with a negative correlation on the template strand. The TAM model predicts a positive correlation between gene expression levels and mutation rates on both strands. The effect of transcription on gene mutation rates has long been a topic of research in studies of disease and evolution (Pleasance et al. 2010; Chapman et al. 2011; Lawrence et al. 2013; Supek and Lehner 2015). Recently, Liu and Zhang published a critique of our work on germline mutation rates and gene expression during spermatogenesis,

arguing that TCD is the overall dominant effect of transcription on germline mutation rates in human and reaffirming their support of a transcription-associated mutagenesis (TAM) model (Liu and Zhang 2020). The TAM model posits a monotonic positive correlation between gene expression levels and mutation rates (Fig. 1A, B). The key aspect of the TAM model is that transcription is categorically mutagenic, such that TCD overwhelms any effect of TCR. The TAM model thus predicts a positive correlation between expression level and mutation rates on both the coding and template strands (Fig. 1C). It is noteworthy that this model is not supported by data from somatic mutation rates in mammalian systems, which in fact show a negative correlation with gene expression levels across multiple tumor types (Pleasance et al. 2010; Chapman et al. 2011; Lawrence et al. 2013; Supek and Lehner 2015). In further contrast to the TAM model, the transcriptional scanning hypothesis recognizes that the net effect of TCR and TCD is expression-level and strand-dependent. Specifically, the transcriptional scanning model predicts a negative correlation between the template-strand mutation rates and expression levels (due to TCR) and a positive correlation between the coding-strand mutation rate and expression levels (due to TCD, Fig. 1C). Moreover, in the highly expressed genes, the TCD is predicted to dominate, such that – for this class of genes - a strong positive correlation would be evidence between mutation rate and expression levels (Fig. 1C).

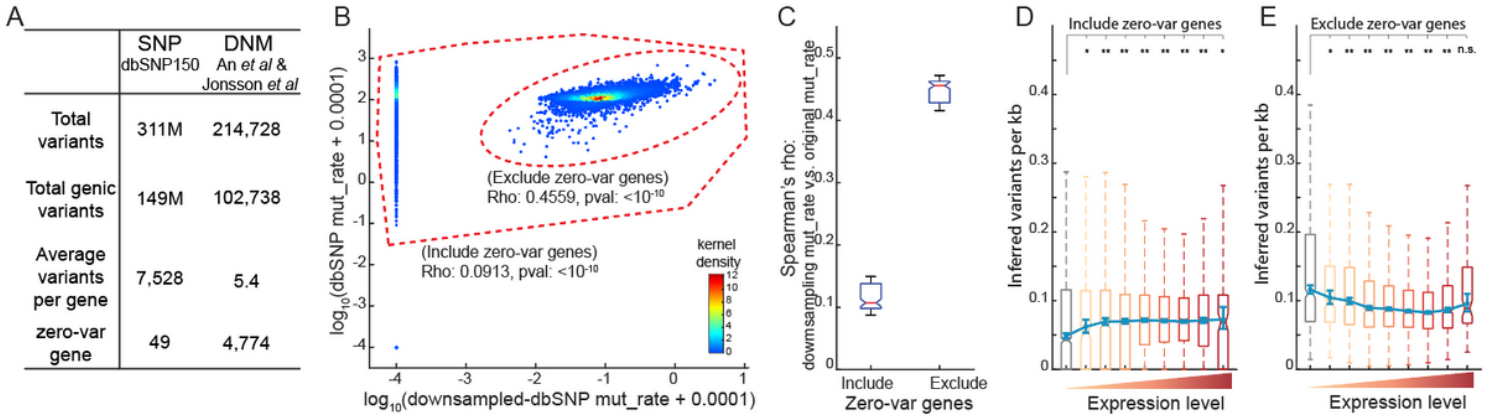

Figure 2

Excluding zero-variant genes in sparse mutation datasets reduces biases in mutation rate inference. (A) The number of variants in the SNP and DNM datasets. (B) A representative scatter view of mutation rates inferred from randomly down-sampled SNP dataset and the original SNP dataset. Spearman's rank correlation coefficients are indicated with or without zero-variant genes. (C) Spearman's rank correlation coefficients between mutation rates inferred from down-sampled SNP dataset and the original dataset. The random sampling was repeated for 100 times, and the Spearman's correlation coefficients were calculated in pairs, including or excluding zero-variant genes. (D,E) Down-sampled SNP dataset-inferred mutation rates across gene expression level categories including (D) or excluding (E) the zero-variant genes. Significance in D-E is computed by the Mann-Whitney test with Bonferroni correction for multiple tests. *, P<0.01; **, P<1.0-10; n.s., not significant.

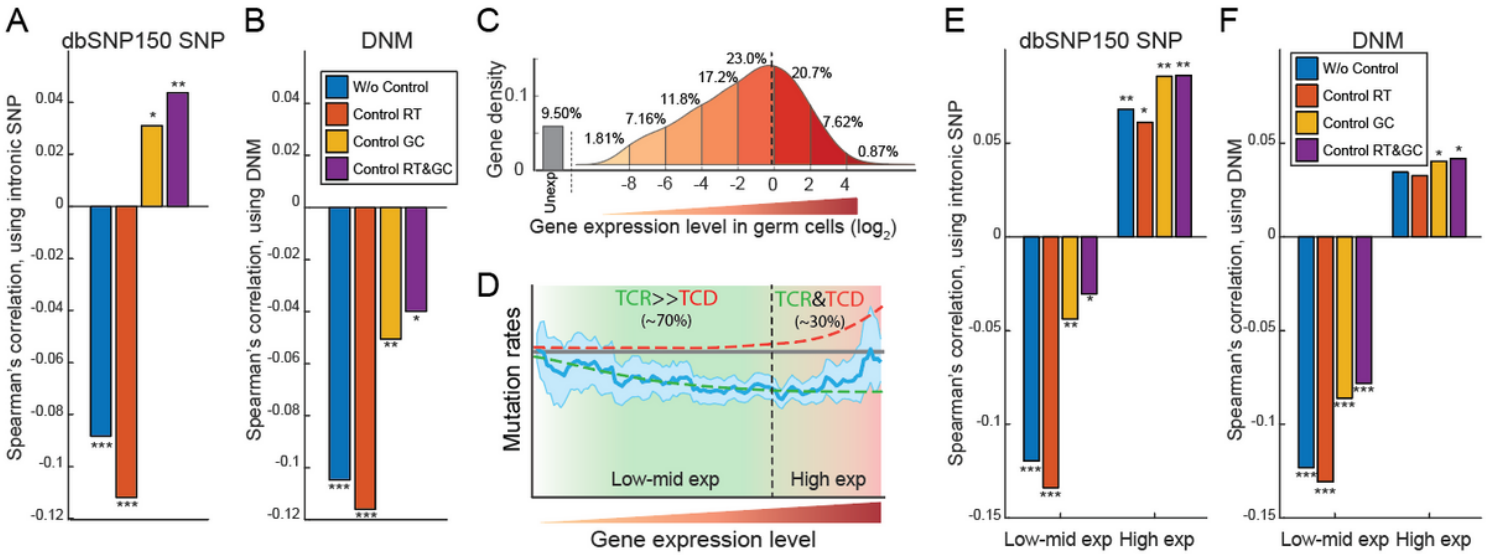

Figure 3

Gene expression level tunes mutation rates by the compound effects of TCR and TCD. (A-B) Spearman's correlation coefficients between gene expression level during spermatogenesis and mutation rates inferred from SNP dataset (A) or DNM dataset (B). (C) Distribution of gene expression levels in the human male germ cells. The plot is adapted from Xia, et al (2020). (D) Schematic of the compound effects of TCR and TCD. TCR dominates in the ~70% of low-to-moderately expressed genes, while TCD gradually overwhelms TCR in the top ~30% highly expressed genes. The schematic is adapted from Xia, et al (2020). (E-F) Spearman's correlation coefficients between gene expression level and mutation rates inferred from SNP dataset (E) or DNM dataset (F). Genes in E-F are divided into low-to-moderately expressed group and highly expressed group and their correlation coefficients between mutation rates and expression levels are plotted, respectively. *, P<0.01; **, P<1.0-5; ***, P<1.0-10. GC: GC content. RT: replication timing.

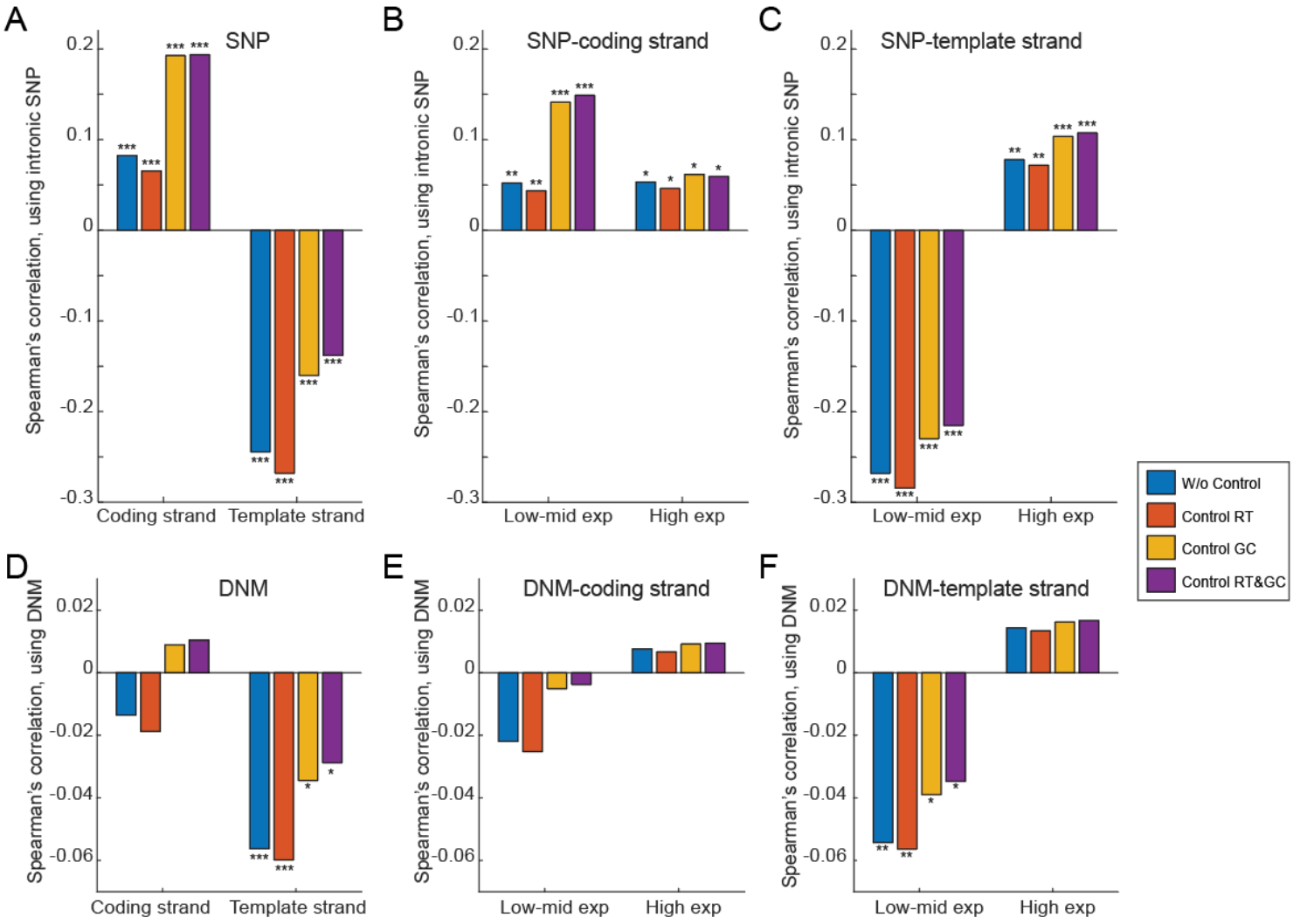

Figure 4

Gene expression level tunes mutation rates of coding strand and template strand differently. (A) Spearman's correlation coefficients between gene expression level and mutation rates inferred from coding strand (left) or template strand (right). (B-C) Spearman's correlation coefficients between gene expression level and coding strand mutation rate (B) or template strand mutation rate (C). Genes in B-C are divided into low-to-moderately expressed group and highly expressed group and their correlation coefficients between mutation rates and expression levels are plotted, respectively. (D-F) Same as in A-C, but used de novo mutations. *, P<0.01; **, P<1.0-5; ***, P<1.0-10. GC: GC content. RT: replication timing.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SFig1.png
- SFig2.png
- SFig3.png

- SFig4.png