Sex differences in transcription-associated mutagenesis in the human germline

Minyoung J. Wyman[1,*]

Ipsita Agarwal[1,2]

Marc de Manuel[1,3]

Natanael Spisak[1]

Molly Przeworski[1,4]

[1] Department of Biological Sciences, Columbia University, New York, NY 10027, USA

[2] current address: Department of Statistics, University of Oxford; Oxford OX1 3LB, United Kingdom

[3] current address: Institute of Evolutionary Biology (CSIC-UPF). Passeig Marítim de la Barceloneta 37-49; 08003 Barcelona, Spain

[4] Department of Systems Biology, Columbia University, New York, NY 10027, USA

[*] To whom correspondence should be addressed: mw2930@columbia.edu

Running Title: Sex differences in transcriptional mutagenesis

Keywords: germline mutation, gonadal expression, sexual dimorphism, gametogenesis

Abstract

In humans, germline mutation rates are three- to four-fold higher in males than females for largely unknown reasons. We investigated whether transcription, a well-documented source of both DNA damage and repair in somatic tissues, is associated with sex differences in germline mutations. We used single-cell germline expression data from different stages of gametogenesis in males and females and phased de novo mutations from pedigrees. Focusing on protein-coding genes, we found no relationship between the male mutation rate and averaged gene expression levels during spermatogenesis despite evidence for transcription-asymmetry. Individual stages of spermatogenesis appear to vary in their contribution to mutation, however: expression levels in spermatogonial stem cells and sperm are significantly positively associated with paternal mutation rates, while those in primary spermatocytes are significantly negatively associated. Thus, transcription may have varying effects over male gametogenesis that are not visible from its cumulative effect on the total germline mutation rate. By contrast, the female mutation rate increases significantly with transcription levels averaged across different stages of folliculogenesis, with corresponding evidence for transcription-asymmetry. The difference between the sexes is also observed when using a smaller set of phased mutations from three-generation pedigrees and when relying on expression data from bulk gonadal tissues. These results indicate that transcription has distinct effects on the mutation rate in the two sexes, leading to more mutations in females but not males, in contrast to what one might expect from the overall paternal bias in germline mutations.

Introduction

Mutations in the human male and female germlines differ in a number of regards. Notably, the rate of de novo mutation (DNM) is three to four times higher in males than in females (Jónsson et al. 2017), and the mutational spectrum differs measurably between the sexes (Goldmann et al. 2016; Jónsson et al. 2017; Agarwal and Przeworski 2019; Spisak et al. 2024). The basis for these sex-specific differences remains poorly understood, but multiple lines of evidence suggest that germline point mutations arise predominantly from DNA damage that is inaccurately repaired (Gao et al. 2016; Gao et al. 2019; Seplyarskiy et al. 2019; Seplyarskiy et al. 2021; de Manuel et al. 2022; Spisak et al. 2024). Therefore, the sex difference in mutations may reflect, at least in part, differences in damage rates or repair accuracies between sexes.

Transcription is a known source of both DNA damage and repair, shaping mutation rates in human somatic tissues (Haradhvala et al. 2016; Pleasance et al. 2010; Alexandrov et al. 2013; Alexandrov et al. 2020). High transcriptional levels can create supercoiled DNA in front of or behind the advancing RNA polymerase II (RNAPII). TOP1 relieves positive supercoiling by nicking the DNA backbone, but cleavage at sites of genome-embedded ribonucleotides can lead to short deletions (Takahashi et al. 2011), a signature detected in the human germline (Reijns et al. 2022). Negative supercoiling results in single stranded DNA, a chemically less stable conformation (Jinks-Robertson and Bhagwat 2014), which introduces opportunities for lesions on both strands (Gaillard and Aguilera 2016; Cho and Jinks-Robertson 2017).

On the other hand, transcription can recruit the repair machinery. Transcription-coupled repair (TCR), a sub-pathway of nucleotide excision repair, may intervene when a bulky lesion on the transcribed strand stalls RNAPII (Hanawalt and Spivak 2008; Vermeulen and Fousteri 2013; Spivak 2015; Nicholson et al. 2024). TCR also processes R-loops (Sollier et al. 2014; Sollier and Cimprich 2015). In the soma, TCR repairs insertions and deletions on the transcribed strand in conjunction with mismatch repair (MMR) (Georgakopoulos-Soares et al. 2020). TCR cross-talks with the components of base excision repair (BER) to repair smaller oxidative lesions encountered during transcription (Svejstrup 2002; Kim and Jinks-Robertson 2010; Chakraborty et al. 2021). In addition, a separate process known as domain-associated repair can repair either strand within transcribed regions (Nouspikel et al. 2006; Zheng et al. 2014). The effect of transcription levels on the overall mutation rate (i.e., across both strands) will therefore depend upon the combination of all of these processes.

In somatic cell types, the mutation rate appears to be negatively correlated with gene expression levels, which suggests that the effect of TCR dominates over transcription-related damage (Pleasance et al. 2010; Chapman et al. 2011; Chen et al. 2017). The patterns in the germline are less clear: previous studies have suggested a negative (Xia et al. 2020; Xia and Yanai 2022), a positive (Chen et al. 2017; Liu and

Zhang 2020), or no significant correlation (Moore et al. 2021) between mutation rates and male expression levels. Part of the discrepancy may stem from differences across studies in the male expression data used: e.g., bulk testis tissue (Chen et al. 2017), seminiferous tubules (Moore et al. 2021), or single-cell spermatogonia (Xia et al. 2020; Liu and Zhang 2020). Studies also differed in whether they considered only mutations in testes (Moore et al. 2021), all germline mutations (Xia et al. 2020; Liu and Zhang 2020), or paternally and maternally phased mutations separately (Chen et al. 2017), as well as in what covariates were included, and whether genes with no mutations were included (Xia et al. 2020; Xia and Yanai 2022; Liu and Zhang 2020). Finally, only one study (Chen et al. 2017) looked specifically at maternal germline mutations, reporting a positive correlation with bulk ovary tissue expression.

Despite the lack of clarity about the relationship of expression levels to germline mutation rates, indirect evidence points to an effect of transcription on human germline mutations, at least for some mutation types: there are fewer A>G substitutions on the transcribed strand than on the non-transcribed strand in mammalian comparisons (Green et al. 2003). Strand asymmetry in transcribed regions (T-asymmetry) indicates differential repair or damage rates between the DNA strands. Analyses of DNM and single nucleotide variant data from current human populations also support germline A>G T-asymmetry (Xia et al. 2020; Chen et al. 2017; Seplyarskiy et al. 2021; Seplyarskiy and Sunyaev 2021).

To clarify how transcription-related repair and damage affects human germline mutations, especially in females, we set out to characterize the relationship between mutation rates and gene expression levels in the germline cells of both sexes, using single-cell expression data and phased DNMs.

Materials and methods

Mutation rates in transcribed regions We compiled 451,005 single nucleotide DNMs from three datasets composed of 3792 trios (816 trios from (Goldmann et al. 2016; An et al. 2018; Halldorsson et al. 2019), 2976 trios from (Halldorsson et al. 2019) and 1902 quartets from (An et al. 2018)). From the total set of mutations, 145,999 were assigned to either the paternal (115,093) or maternal (30,906) genome through read-based phasing or third generation transmission. We removed 896 mutations with reference alleles that differed between hg19 and grch38.

To estimate the sex-specific mutation rate in transcribed regions, we required an estimate of the total number of base pairs surveyed. To this end, we targeted transcribed regions by selecting for “transcripts” in the feature column of the GENCODE comprehensive gene annotation GTF for chromosomes (grch38, release 41), which includes exons, introns, and UTRs. We took the coordinates of all the transcripts belonging to one gene and merged them to derive one contiguous transcribed region per gene (bedtools merge command), taking care not to merge overlapping transcripts from opposite strands.

Within the contiguous transcript, we tallied all base pairs that had at least 10x mean coverage, as reported in gnomAD v3. We reasoned that this low pass quality filter is a proxy for the sites that are consistently accessible by whole genome sequencing. We applied the same mean 10X coverage requirement to the mutations used for our calculations, leaving 70,238 paternal and 18,755 maternal mutations. We included genes with no mutations as long as the genes were listed in the two single-cell RNAseq (scRNA-seq) datasets (Guo et al. 2018; Zhang et al. 2018) and in the GTEx expression dataset (Xia et al. 2020; Xia and Yanai 2022; GTEx Consortium 2013; Liu and Zhang 2020). Only a subset of these mutations fell in 15,625 protein-coding genes with expression data available in both sexes: 48,511 paternal and 12,906 maternal mutations. Among the genes common to the three expression datasets, only 15,562 genes had replication timing values, a covariate in our regression models, leaving 48,157 male and 12,799 female mutations. Within this final set, 4895 paternal and 1144 maternal mutations were phased with a third generation.

We were also interested in the association of expression with different kinds of mutations. We consider transitions at CpG sites and all other mutation types, corresponding roughly to SBS1 and SBS5 respectively, the two main signatures prevalent among germline mutations (Rahbari et al. 2016; Moore et al. 2021). We furthermore partitioned transitions at CpGs into those methylated in gonadal tissues and those that are not (see *Methylomes* section).

In all regression models, we estimated the mutational opportunities (denominator used for offset term) for all types of mutations (numerator used for the response). We kept track of the total number of base pairs passing the 10X filter.

Gene expression levels We used three sets of expression data: single-cell RNA-seq for male (Guo et al. 2018) and female germlines (Zhang et al. 2018), and bulk tissue expression for male and female gonads from the GTEx Portal (v8) (GTEx Consortium 2013). All expression values used in multiple regression models were natural-log transformed and then re-scaled (to have mean = 0, standard deviation = 1).

Male expression data were collected from eight reported stages of spermatogenesis distinguished by cell marker expression (original stages from (Guo et al. 2018) indicated with numbers): spermatogonial stem cells (stage 1), differentiating spermatogonia (2), early primary spermatocytes (3), late primary spermatocytes (4), round spermatids (5), elongated spermatids (6), and finally, early (7) and later stages of sperm (8). Five types of testis somatic tissues (i.e., Leydig cells, macrophages, Sertoli cells, endothelial cells, myoid cells) were also identified (Guo et al. 2018). single-cell female expression data were collected from five temporally consecutive stages of folliculogenesis (primordial, primary, secondary, antral and preovulatory) for both oocytes and somatic granulosa cells collected in adult females; cell types were distinguished first morphologically and subsequently through marker expression (Zhang et al. 2018). There were no expression data for any of the fetal germline stages (e.g., primordial germ cell and proliferating germ cell for either sex) nor for oogonia in females.

The Zhang et al. (2018) and Guo et al. (2018) datasets used different sequencing technologies and expression units (FPKM and UMI, respectively). The male expression data consisted of the normalized UMI counts of 13 cell types from three donors. For each gene, we averaged within replicates of one cell type from an individual donor using the Seurat R package (AverageExpression() command); we then averaged expression across donors, to obtain 13 stage-specific expression values per gene. For the female expression data, we first converted the expression units from FPKM to TPM, as TPM is more suitable for the purpose of sample comparisons (Li et al. 2010; Pachter 2011; Wagner et al. 2012). Eight donors contributed unequally to the total sample of 80 oocytes and 71 granulosa cells (Zhang et al. 2018). For each gene, we averaged within cells of one type from an individual donor and then averaged expression across donors, resulting in 10 stage-specific expression values per gene.

To permit their inclusion, genes with zero expression values (after averaging across all donor samples) in the male data were arbitrarily reassigned a non-zero value (3.7×10^{-4} UMI) equal to half of the smallest non-zero expression value across all male germline stages. We performed the analogous procedure using the female mutation counts and the germline stage specific expression values. Genes with zero expression values in the female data (after averaging across all donor samples and across tissue types) were reassigned a non-zero value (3.3×10^{-9} TPM) equal to half of the smallest non-zero expression value.

We conducted analyses based on two summary calculations of the original single-cell expression datasets. For each gene, within each sex, we weighted each of the germline cell stages equally. Specifically, we took the stage-specific expression values and then averaged across all germline developmental stages within a sex (8 stages in males and 5 in females) to obtain one value per gene, which we call the germline averaged expression value. Again, to permit the inclusion of any gene with a germline averaged value of zero, we arbitrarily reassigned them a new non-zero value equal to half of the smallest non-zero germline averaged expression for the male (4.7×10^{-5} UMI) and female (7.3×10^{-9} TPM) single-cell data separately. All data were then ln-transformed and rescaled within each germline separately. We note that while in principle, it might be preferable to weight expression levels across the distinct germline stages according to their temporal duration (or the number of cell divisions within each stage), any such estimate would be highly imprecise, given limited information about gametogenesis in humans and variation in parental ages.

The GTEx expression data came from bulk testis and ovary tissue samples, so both male and female samples are an amalgam of somatic and germline cells. We took the mean TPM across all replicates within each of the two tissue types separately. If a gene had an expression value of zero after averaging across replicates, we replaced it with a non-zero value equal to half the smallest non-zero expression value (0.0083 TPM in the testis data and 0.00013 TPM in the ovary data); all data were ln-transformed and subsequently rescaled for the multiple regressions.

Methylomes We used methylation data from Gene Expression Omnibus for ovaries (GSM1010980) and testis (GSE127321, GSE127281) (ENCODE Project Consortium 2012). Because the ovary methylation data are in hg19, we re-generated the contiguous merged transcripts and 10x coverage positions for each gene with the hg19 equivalents of the GENCODE gtf (v19) and the gnomAD (v2) files. CpG sites were considered methylated if they had a methylation value equal to or greater than 0.75 (i.e., 75% of reads indicate methylation, (e.g., (Agarwal and Przeworski 2021)). This classification enabled us to build regression models for transitions at methylated and nonmethylated sites, as well as to calculate T-asymmetry. Not all CpG positions had methylation data, but this missing data should not bias our mutation rate estimates, since it affects the mutations in the numerator and the nucleotide counts in the denominator similarly.

Additional covariates We quantified the effect of expression levels on mutations by building a series of multiple regression models to estimate the effect of a predictor while holding all others constant and controlling for confounding effects. All dependent variables were re-scaled in the models for interpretability.

All regression models used individual genes as the data points. The predictor for expression level included one of three possibilities: the within-sex, unweighted averaged expression values of all germline stages, the original thirteen male or ten female stage-specific expression values (with progressive removal of non-significant stages), or the GTEx gonadal expression values averaged across replicates.

All regression models of the effect of expression levels on the mutation rate included three important covariates: GC content, replication timing, and the fraction of mutations that were phased to male or female genomes. GC content was included because it is known to affect the mutation rate, both directly (Kiktev et al. 2018), and indirectly, via the sequencing depth (Risso et al. 2011; Vallejos et al. 2017). Within each gene, we calculated GC percent for the regions passing the 10X filter. Phased fraction was not included in models using mutations that were phased by third generation transmission.

Replication timing is also known to affect the mutation rate (Stamatoyannopoulos et al. 2009; Koren et al. 2012). We averaged the replication timing values of 300 iPSC lines (Ding et al. 2021) at the provided genomic locations (i.e., zero-scaled with negative values denoting regions that replicate later than average and positive values earlier than average). We then averaged the positions that fell along the transcribed regions passing the 10X filter to obtain one replication timing value per gene. Of the 15,625 protein-coding genes present across all expression data sets, only 15,562 had replication timing information available.

In addition, we included the fraction of phased mutations as a covariate. We observed that the fraction of phased mutations (maternal + paternal) declines with increasing male germline expression across deciles. This negative correlation is potentially problematic because the sex-specific mutation rate may appear to decline with expression simply through its correlation with phasing rate. Since the phased fraction per gene is too imprecisely estimated, we assigned each gene a phased fraction value by using the phased fraction value from the expression decile to which it belonged. We calculated the decile-specific phased fraction by summing the number of all male and female phased mutations and dividing by the total number of mutations within that expression decile.

As CpG transitions occur at a greater rate at methylated dinucleotides, we included methylation level as a covariate in the regression models when CpG C>T mutation rate was the response. Average methylation values for each gene were calculated by summing the methylation values for each CpG site in the 10X filtered regions and dividing by the total number of CpG sites included. This calculation was performed in the ovary and testis datasets separately for the individual male and female regression models. The regressions excluded genes without CpG dinucleotides or genes without any methylation data. Genes with zero CpG C>T mutations were retained as long as CpG methylation data were available.

Modeling approach We implemented a Poisson glm multiple regression model (R command glm(family=poisson(link=log)) on the gene-specific data. If the data suggested overdispersion with respect to a Poisson model, we instead applied a quasipoisson (glm(family=quasipoisson(link=log)) model. All models used the log of the number of surveyed base pairs as an offset term. All predictors were scaled for interpretability. For the regression models including expression from multiple somatic and germline cell stages, we conducted a backwards elimination that dropped non-significant stages to only a pared down set of stages that had a significant effect on the mutation rate.

Asymmetry measure By choosing the frame of reference (e.g., mutation type or strand), it is possible to separate changes of paired bases and tally the changes on each strand individually. For a given mutation type, T-asymmetry is measured by the difference in the number of base changes on the transcribed versus non-transcribed strands. If the transcribed strand is 5'AAATT in both parents and 5'AAGTC in the offspring, one third of the A bases on the transcribed strand were changed to G; one half of the A bases were changed to G on the non-transcribed strand. Picking a reference base (e.g., A instead of T) to count on each strand is arbitrary. We analyzed strand asymmetry in the six basic mutation types: A>G, A>C, A>T, G>T, C>G, and C>T. We parsed the C>T mutation type by nonCpG C>T, CpG C>T, methylated CpG C>T and nonmethylated CpG C>T sites.

We tested for differences in T-asymmetry through a proportions test (chi-square test with df=1). We used the FDR method for multiple testing comparisons (done once across all mutation types across both sexes). For the proportions test, we counted the total number of DNMs of a given type on the transcribed strand (e.g., A>G) versus the total number of mutations of the complementary type (T>C), as well as the total number of relevant nucleotides on the transcribed strand (i.e., A), versus the total number of complementary base pairs (T), to account for mutational opportunity. We estimated the degree of strand asymmetry using the same set of genes for the transcribed strand and non-transcribed strands. Strand asymmetry was also always calculated in the same set of genes for which male and female data entries are available across the three expression datasets and the replication timing dataset.

We compared the extent of asymmetry between the two sexes by performing a two-tailed test on the sex difference in the $\log_2$ (non-transcribed strand / transcribed strand) for a given type of mutation; p-values were FDR adjusted.

DNA repair genes To examine expression levels of DNA tolerance and repair genes across gametogenesis, we used the list maintained by the Wood laboratory at the MD Anderson Cancer Center (Wood et al. 2001) https://www.mdanderson.org/documents/Labs/Wood-Laboratory/human-dna-repair-gen

es.html. Using the logged expression values, we constructed heatmaps (heatmap.2() in R) that scaled and clustered across rows (i.e., genes). To include data points with zero expression, we applied our usual regularization procedure and added a small amount equal to half of the smallest non-zero expression value across the entire RNAseq dataset within each sex.

Results

Since transcription may differentially affect mutagenesis in the germline at many stages of gamete development, we first analyzed the cumulative effects of transcription by examining the effect of averaged germline expression levels in each sex, and then sought to identify the contribution of expression levels from individual stages of gametogenesis.

Averaged germline transcription and mutation rate We regressed the rate of DNMs on expression levels in the male and female germlines separately. To this end, we used the average of single-cell expression levels from adult spermatogenesis (Guo et al. 2018) and adult folliculogenesis (Zhang et al. 2018), weighing each stage equally ("germline averaged expression"). We note that the expression data are on different measurement scales in the two sexes, UMI (unique molecular identifiers) in males and TPM (transcripts per million) in females, precluding a direct sex comparison in expression levels with these data. We considered only protein-coding genes, which are transcribed by RNAPII, the enzyme integral to initiating TCR. We also used only genes with expression information recorded in both sexes across the two single-cell expression datasets, as well as in GTEx.

We considered 48,157 paternal and 12,799 maternal mutations in protein-coding regions from pedigree sequencing (Goldmann et al. 2016; An et al. 2018; Halldorsson et al. 2019). These mutations were phased, i.e., assigned to male versus female germline, by read-backed phasing and third generation transmission. The mutation rate was calculated as the number of DNMs in a transcribed region, adjusted by the length of the region, after removing sites with potentially low sequencing depth. A complication is that using read-back phasing, we expect more mutations to be phased in regions with greater nucleotide diversity. Since genes with higher levels of transcription tend to be under stronger purifying selection (Zhang and Yang 2015), and thus have less diversity, the rate of phasing may itself depend on expression. To take this confounding into account, we included the fraction of mutations that were phased per decile as a covariate when regressing mutation rates on expression levels. We also added GC content and replication timing from induced pluripotent stem cells (iPSC) as covariates, since both are known to affect mutation rates (Stamatoyannopoulos et al. 2009; Michaelson et al. 2012; Liu and Zhang 2020)

Using this approach, we found no significant relationship between the male mutation rate and averaged expression level across adult spermatogenesis (Figure 1A; Supp. Table 1), after adjusting for the covariates of GC content, replication timing, and phased fraction. By contrast, the association between the female mutation rate and the averaged expression level across adult folliculogenesis was significantly positive (Figure 1B; Supp. Table 2), after adjusting for the same covariates. GC content and phased

fraction were positively associated with the mutation rate, and replication timing was negatively associated with mutations in both of the sex-specific regression models.

To verify our finding of a null association in males and a positive association in females, we conducted a number of checks. First, as an alternative to including the phasing rate as a covariate, we limited the analysis to the (much smaller) subset of mutations that were phased using transmission to a third generation (Jónsson et al. 2017), an approach that does not rely on the presence of nearby SNPs. Using these 4895 paternal and 1144 maternal mutations, we again found that the germline averaged expression level has no effect on mutations in males (P=0.28; Table S3) and a positive effect in females (P=0.0023; Table S4).

As a second way to gauge the consistency of the results, we used the total number of available mutations, whether or not phased and regardless of parental origin, as our response variable. We note that ~80% of these mutations are expected to have occurred in the male germline (Jónsson et al. 2017); if there is truly no effect of male germline expression levels on the male mutation rate, the inclusion of all paternal mutations should preserve a null association. Conversely, if only female expression levels have an effect on female germline mutations, the male mutations should contribute noise but not systematically alter an underlying positive relationship. Accordingly, we recovered a positive association between the total mutation rate and female germline averaged expression levels ($P<2x10^{-16}$; Table S5), but no significant relationship with male germline averaged expression levels (P=0.15; Table S6).

In addition, we performed the same set of multiple regressions but replaced the averaged single-cell expression data with bulk ovary and testis tissue expression data from GTEx. Bulk tissue expression data, being comprised of both somatic and germline cells, provide a noisier estimate of transcription levels in gametogenesis. Nonetheless, expression levels in somatic tissues should not cause changes in germline mutation rates (see the *Stage-specific gene expression* section). Moreover, because the GTEx data are on the same measurement scale in both ovaries and testes, it is possible to combine the two datasets in the same model and test directly for a statistical interaction between the effects of gonadal expression and sex on the mutation rates. After including the covariates of GC content, replication timing, and phased fraction, we again found that expression levels have no effect on the mutation rate in males (P=0.51), but a significantly positive effect in females ($P=1.4 \times 10^{-8}$) that differs from the male slope (P=0.0003; Table S7). In other words, the relationship between expression level and mutation rate depends on sex. Altogether, these analyses confirm that there is no discernible effect of averaged germline expression levels on mutation rates in males but a positive effect in females.

The differential association of expression levels and mutation rates in the two sexes ought to impact the extent of paternal bias along the genome. To date, estimates of paternal bias, whether from DNM calls in pedigrees or comparisons of X and

autosomal divergence rate have been genome-wide (e.g., (Wilson Sayres and Makova 2011; Jónsson et al. 2017; de Manuel et al. 2022)). Our findings imply instead that, at least for some mutation types in protein-coding genes, the paternal bias should decrease with expression levels. Indeed, when comparing the number of third generation phased mutations predicted by our model (after adjusting for covariates of GC content and replication timing), we observe a 25% reduction in the amount of male-biased mutation between the first and fourth quartiles of sex-specific expression, from a paternal bias of 5.06 to 3.8 respectively. Therefore, the degree of sex bias in mutation varies along the genome.

Effects for CpG sites versus other mutation types Next, we sought to understand whether the sex difference may pertain to only a subset of mutations. COSMIC signatures (Alexandrov et al. 2013; Alexandrov et al. 2020) classify single base substitutions (SBS) into 96 types by recovering mutational signatures that correspond to one or a few underlying mutagenic processes (Koh et al. 2021). The vast majority of germline mutations in males and females are assigned to SBS5 and to a lesser extent SBS1 (Rahbari et al. 2016; Moore et al. 2021). However, given the low numbers of mutations in the germline, especially when stratified by expression levels, signature assignment is unreliable (Serrano Colome et al. 2023). Instead, we took the approach of considering only transitions at CpG sites, which constitute the bulk of SBS1 mutations (Alexandrov et al. 2013) and are thought to arise from deamination of methylated CpGs (Guo et al. 2022) or errors in replication (Seplyarskiy et al. 2021; Tomkova et al. 2024). Then, we considered all other mutations jointly, most of which contribute to the diffuse SBS5 signature, whose etiology is unknown but may reflect the error profile of repair (Spisak et al. 2024).

Among transitions at CpG dinucleotides, we found that germline averaged expression levels have a positive effect in females (P=0.014; Table S8), but a negative effect in males ($P=9.4 \times 10^{-5}$; Table S9). These associations appear to be due specifically to transitions at methylated CpG sites, for which the association is positive in females (P=0.005; Table S10) and negative in males (P=0.0002; Table S11). By contrast, for nonmethylated CpGs specifically, expression levels show no association with CpG transitions in either sex (Supplementary Tables 12 & 13).

Among all other mutation types (i.e., excluding transitions at CpGs), germline averaged expression levels are positively associated with the mutation rate in females ($P<2\times10^{-16}$; Table S14) and not significantly associated in males (P=0.99; Table S15).

Stage-specific gene expression and mutation rate The germline mutation rate is a sum of mutations that arise at different stages of gametogenesis. Transcription levels differ across stages and may influence mutagenesis in distinct ways. To study stage-specific effects, we used scRNA-seq data from several stages of male and female

gametogenesis (Guo et al. 2018; Zhang et al. 2018). For paternal mutations, we performed a Poisson multiple regression using expression from each of the 13 testis cell types as predictors (n=15,562 genes), which included eight germline and five somatic cell types (Guo et al. 2018). We included the three covariates of GC content, replication timing, and phased fraction, as well as a logged offset term to account for denominator length (i.e., mutational opportunity). By including all cell stages, we can estimate the effect of a particular cell stage while holding the other remaining cell stages and predictors constant.

We performed a backwards elimination on the full model, progressively dropping cell types that had no significant effect on the male mutation rate (at $P > 0.05$). This stepwise procedure removed all five somatic stages, as expected. The resulting model retained four spermatogenic cell stages (numbers indicate position in the temporal order): spermatogonial stem cells (stage 1), early primary spermatocyte (3), late primary spermatocytes (4), and early sperm (7). However, the two primary spermatocyte stages demonstrated high collinearity (Variation inflation factor, VIF > 4), as might be expected by their ontogenetic similarity. We therefore kept only the late primary spermatocyte stage, which lowered the VIFs of the remaining predictors (Table 1). Male mutation rates were positively associated with gene expression levels in the spermatogonial stem cell stage and the sperm cell stage, but negatively associated in the primary spermatocyte stage. The sum of the scaled coefficients of the three cell stages was only very slightly positive (Table 1). GC content and phased fraction were positively associated with the mutation rate, and replication timing was negatively associated.

Changes in the sign of the slopes across temporally ordered cell stages can indicate that the coefficient estimation was unstable due to multicollinearity among the cell stages. To explore this possibility, we also performed LASSO and ridge regressions, two regularization approaches that reduce overfitting and the effect of predictor correlations. We obtained coefficients with similar values as in the Poisson regression (Table S16), suggesting that the estimated coefficients are robust.

For the female data, we performed a multiple regression model using expression from each of 10 folliculogenesis cell types as predictors (Zhang et al. 2018). We included the three covariates of GC content, replication timing, and phased fraction, as well as a logged offset term to account for denominator length. We used a quasipoisson model since the Poisson model did not adequately capture the variability (dispersion parameter > 1.3). Stepwise elimination of non-significant predictors and a predictor with high VIF (>2.7) removed all somatic cell stages from the model. The final model retained only the first stage of folliculogenesis, primordial oocyte (Table 2), from which expression levels have a positive effect on the female mutation rate. GC content and phased fraction were positively associated with the mutation rate and replication timing was negatively associated in the regression model including the female germline stages.

One biological explanation for differences in the sign of the association of cell stages with the mutation rate may be differential repair efficacy among stages. To explore this possibility, we examined the expression levels of genes from various repair pathways across these stages. Overall, expression levels did not correspond with the results from the stage-specific regression model (Table 1 and Figure S1). For example, the spermatogonial stem cell stage has comparatively high expression levels of repair genes, yet mutation rates increase with expression in this stage. We also examined the expression pattern of genes involved in repair across the five stages of adult folliculogenesis. Primordial oocytes appeared to have the lowest amount of expression of repair genes, consistent with a positive effect on the mutation rate (Table 2 and Figure S2). However, in the four stages after the primordial oocyte stage, the expression levels of repair genes remain high, and yet there is no statistically significant association of expression levels with the mutation rate. Therefore, in neither sex is there an obvious correspondence between repair expression levels and the stages identified as affecting the mutation rate. The discordance is perhaps not unexpected, as expression levels of repair genes may be a weak proxy for their enzymatic repair potential (Buccitelli and Selbach 2020).

Sex-specific T-asymmetry T-asymmetry arises because the balance of repair and damage differs between the transcribed and non-transcribed strands. A negative association between expression levels and mutation rate, together with appreciable T-asymmetry, may indicate a balance in favor of TCR. By contrast, a positive association combined with strong T-asymmetry may indicate that the effects of TCD dominate. Finally, the lack of an association between expression levels and mutation rate, coupled with the absence of T-asymmetry, suggests the lack of both TCR and TCD, or that the two effects exactly cancel out.

Several mutation types demonstrate significant T-asymmetry in one or both sexes after FDR adjustment across both sexes (n=15,562 genes; significance indicated with darker bars in Figure 2). In general, there is stronger maternal T-asymmetry, (two-tailed test $P<0.05$ after FDR adjustment). Consistent with previous reports (Seplyarskiy and Sunyaev 2021), A>G and C>G asymmetry is stronger in females than in males. We also found stronger T-asymmetry for A>C, nonCpG C>T, and C>T at CpG dinucleotides in females. T-asymmetry in G>T is stronger in males than females.

At methylated CpGs specifically, we found evidence for C>T T-asymmetry in both sexes using ovary and testis methylation data (Figure 2). This finding is consistent with the presence of a positive association between transcriptional levels and methylated CpG C>T mutations in females (Supp. Tabl. 10) and the presence of a negative association in males (Supp. Tabl. 11). There is no sex difference in the extent of T-asymmetry at methylated CpG C>T mutations when testing the subset of genes that have methylation data available for both ovaries and testes (n=15,285 genes; adjusted $P=0.27$).

525 Finally, transitions occurring at nonmethylated CpGs have no T-asymmetry in
526 either sex (n=15,510 genes), consistent with the lack of association with expression in
527 either sex (Supplementary Tables 12 and 13). The sex difference in T-asymmetry is
528 marginally significant (P=0.04; Figure 2).

Discussion

We found consistent evidence for a null effect of male germline transcription on male germline mutations across different datasets, whether using all paternally phased mutations or only the paternal mutations identified through third generation transmission. Moreover, the null association was found using both single-cell and bulk tissue expression data. The lack of an effect of transcription suggests either that transcription-related damage and repair are both weak or non-existent throughout the male germline or that, summed over all stages of adult spermatogenesis, the effects of repair approximately cancels out the effects of transcription-related damage.

In support of the latter, expression levels from some stages of spermatogenesis have individually significant associations with the male mutation rate (Table 1). Expression levels from spermatogonial stem cells and sperm are positively associated with male germline mutations whereas they are negatively associated in primary spermatocytes; the combined effect of the three stages is close to zero, paralleling the null effect of male averaged germline expression. Consistent with significant associations across individual stages is the presence of significant A>G, A>T, G>T C>G, and methylated CpG C>T T-asymmetry in the male germline. A caveat is that T-asymmetry and R-asymmetry are somewhat correlated (Seplyarskiy et al. 2019; Koyanagi et al. 2022), so that the T-asymmetry could actually be indicative of R-asymmetry. However, the presence of significant associations of mutation rates and stage-specific expression levels, it seems more likely that the male T-asymmetry emerged from individual germline stages experiencing net mutagenic or net reparative transcription.

By contrast, female germline mutations increase with expression levels in the female germline, whether considering maternally phased mutations or only the maternal mutations identified from third generation transmission, and whether considering single-cell or bulk ovary expression. Thus, in females, adult germline transcription may incur more damage than is corrected accurately by repair. The stage-specific regression suggests that the positive association may arise specifically from the adult primordial oocytes. The overall positive association of transcription with the female germline mutations may explain the significantly larger degree of A>G, A>C, C>G, and nonCpG C>T, and CpG>TpG T-asymmetry observed in females relative to males. Overall, our findings establish a difference between sexes in the impact of transcription on mutation.

Curiously, we observed an association between the rate of CpG transitions and expression levels in both sexes. In females, expression levels show a positive association with both CpG transition rates in general (Table S8) and at methylated CpG sites in particular (Table S10). Likewise, in males, expression levels have a negative association with both CpG transitions (Table S9) and methylated CpG transitions (Table S11). A similar pattern has recently been reported among somatic SBS1 mutations,

albeit without covariate adjustment: SBS1 mutations are negatively correlated with expression levels (Silveira et al. 2024), but in this case with no T-asymmetry (Alexandrov et al. 2020; Silveira et al. 2024). A negative association between expression and methylated CpG C>T mutations suggests the possibility that transcription can repair deamination of methylated cytosines, a BER-specific lesion, possibly by TCR itself or by TCR-mediated BER recruitment (Svejstrup 2002; Kim and Jinks-Robertson 2010; Chakraborty et al. 2021). Alternatively, the association between expression levels and methylated CpG mutation rates may not be causal; in other words, the models may be lacking an important confounder for this mutation type in both the somatic and germline analyses.

Regardless, the finding that the relationship of expression levels and mutation rates differs among the individual stages within males and between sexes highlights that the germline ought to be viewed as a succession of distinct cell types, rather than as a homogeneous one. The effect of transcription on mutations in protein-coding genes likely differs across cell types in gametogenesis, and further insight will be gained by studying these mutagenic effects separately, rather than by making inferences from their cumulative effects.

Data availability

This paper relied on publicly available data from previous publications (see Materials and Methods for references). Data tables and accompanying analyses are available through https://github.com/minyoungWyman/sexbasedTCR

Acknowledgments

We thank S. Sanders for sharing unpublished phased mutational data, as well as Z. Yan and J. Guo for sharing unpublished data. We thank the High Performance Cluster Computing at Columbia University for computational support. We are grateful to M.C. Wyman as well as Przeworski, Sella, and Andolfatto labs for helpful discussions.

Study funding

The work was supported by RISE funding from Columbia University to M.P and K. Baldwin, and R35 GM083098 to MP.

References

Agarwal I, Przeworski M. 2019. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. Proc Natl Acad Sci U S A. 116(36):17916–17924.

Agarwal I, Przeworski M. 2021. Mutation saturation for fitness effects at human CpG sites. Elife. 10. doi:10.7554/eLife.71513. http://dx.doi.org/10.7554/eLife.71513.

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN, et al. 2020. The repertoire of mutational signatures in human cancer. Nature. 578(7793):94–101.

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Børresen-Dale A-L, et al. 2013. Signatures of mutational processes in human cancer. Nature. 500(7463):415–421.

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(6420):eaat6576.

Buccitelli C, Selbach M. 2020. mRNAs, proteins and the emerging principles of gene expression control. Nat Rev Genet. 21(10):630–644.

Chakraborty A, Tapryal N, Islam A, Mitra S, Hazra T. 2021. Transcription coupled base excision repair in mammalian cells: So little is known and so much to uncover. DNA Repair . 107:103204.

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(7339):467–472.

Chen C, Qi H, Shen Y, Pickrell J, Przeworski M. 2017. Contrasting Determinants of Mutation Rates in Germline and Soma. Genetics. 207(1):255–267.

Cho J-E, Jinks-Robertson S. 2017. Ribonucleotides and Transcription-Associated Mutagenesis in Yeast. J Mol Biol. 429(21):3156–3167.

Ding Q, Edwards MM, Wang N, Zhu X, Bracci AN, Hulke ML, Hu Y, Tong Y, Hsiao J, Charvet CJ, et al. 2021. The genetic architecture of DNA replication timing in human pluripotent stem cells. Nat Commun. 12(1):6746.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. 489(7414):57–74.

Gaillard H, Aguilera A. 2016. Transcription as a Threat to Genome Integrity. Annu Rev Biochem. 85(1):291–317.

Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, Amster G, Przeworski M. 2019. Overlooked roles of DNA damage and maternal age in generating human germline mutations. Proc Natl Acad Sci U S A. 116(19):9491–9500.

Gao Z, Wyman MJ, Sella G, Przeworski M. 2016. Interpreting the Dependence of Mutation Rates on Age and Time. PLoS Biol. 14(1):e1002355.

Georgakopoulos-Soares I, Koh G, Momen SE, Jiricny J, Hemberg M, Nik-Zainal S. 2020. Transcription-coupled repair and mismatch repair contribute towards preserving genome integrity at mononucleotide repeat tracts. Nat Commun. 11(1):1980.

Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LELM, Hoischen A, Roach JC, et al. 2016. Parent-of-origin-specific signatures of de novo mutations. Nat Genet. 48(8):935–939.

Green P, Ewing B, Miller W, Thomas PJ, NISC Comparative Sequencing Program, Green Eric D. 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet. 33(4):514–517.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45(6):580–585.

Guo J, Grow EJ, Mlcochova H, Maher GJ, Lindskog C, Nie X, Guo Y, Takei Y, Yun J, Cai L, et al. 2018. The adult human testis transcriptional cell atlas. Cell Res. 28(12):1141–1157.

Guo Q, Lakatos E, Bakir IA, Curtius K, Graham TA, Mustonen V. 2022. The mutational signatures of formalin fixation on the human genome. Nat Commun. 13(1):4487.

Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, Gunnarsson B, Oddsson A, Halldorsson GH, Zink F, et al. 2019. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 363(6425):eaau1043.

Hanawalt PC, Spivak G. 2008. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol. 9(12):958–970.

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(3):538–549.

Jinks-Robertson S, Bhagwat AS. 2014. Transcription-Associated Mutagenesis. Annu Rev Genet. 48(1):341–359.

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(7673):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 U S A. 115(30):E7109–E7118.

Kim N, Jinks-Robertson S. 2010. Abasic sites in the transcribed strand of yeast DNA are removed by transcription-coupled nucleotide excision repair. Mol Cell Biol. 30(13):3206–3215.

Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. 2021. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer. 21(10):619–637.

Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, McCarroll SA. 2012. Differential relationship of DNA replication timing to different forms of human mutation and variation. Am J Hum Genet. 91(6):1033–1040.

Koyanagi E, Kakimoto Y, Minamisawa T, Yoshifuji F, Natsume T, Higashitani A, Ogi T, Carr AM, Kanemaki MT, Daigaku Y. 2022. Global landscape of replicative DNA polymerase usage in the human genome. Nat Commun. 13(1):7221.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 26(4):493–500.

Liu H, Zhang J. 2020. Higher Germline Mutagenesis of Genes with Stronger Testis Expressions Refutes the Transcriptional Scanning Hypothesis. Liu L, editor. Mol Biol Evol. 37(11):3225–3231.

de Manuel M, Wu FL, Przeworski M. 2022. A paternal bias in germline mutation is widespread in amniotes and can arise independently of cell division numbers. Elife. 11. doi:10.7554/eLife.80008. http://dx.doi.org/10.7554/eLife.80008.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell. 151(7):1431–1442.

Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, Oliver TRW, Leongamornlert D, Ellis P, Noorani A, et al. 2021. The mutational landscape of human somatic and germline cells. Nature. 597(7876):381–386.

Nicholson MD, Anderson CJ, Odom DT, Aitken SJ, Taylor MS. 2024. DNA lesion bypass and the stochastic dynamics of transcription-coupled repair. Proc Natl Acad Sci U S A. 121:e2403871121.

Nouspikel TP, Hyka-Nouspikel N, Hanawalt PC. 2006. Transcription domain-associated repair in human cells. Mol Cell Biol. 26(23):8722–8730.

Pachter L. 2011. Models for transcript quantification from RNA-Seq. arXiv [q-bioGN].

711 http://arxiv.org/abs/1104.3889.

712 Pleasance ED, Cheetham RK, Stephens PJ, McBride DJ, Humphray SJ, Greenman
713 CD, Varela I, Lin M-L, Ordóñez GR, Bignell GR, et al. 2010. A comprehensive catalogue
714 of somatic mutations from a human cancer genome. Nature. 463(7278):191–196.

715 Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA, Dominiczak
716 A, Morris A, Porteous D, Smith B, et al. 2016. Timing, rates and spectra of human
717 germline mutation. Nat Genet. 48(2):126–133.

718 Reijns MAM, Parry DA, Williams TC, Nadeu F, Hindshaw RL, Rios Szwed DO,
719 Nicholson MD, Carroll P, Boyle S, Royo R, et al. 2022. Signatures of TOP1
720 transcription-associated mutagenesis in cancer and germline. Nature.
721 602(7898):623–631.

722 Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-content normalization for
723 RNA-Seq data. BMC Bioinformatics. 12:480.

724 Seplyarskiy VB, Akkuratov EE, Akkuratova N, Andrianova MA, Nikolaev SI, Bazykin GA,
725 Adameyko I, Sunyaev SR. 2019. Error-prone bypass of DNA lesions during
726 lagging-strand replication is a common source of germline and cancer mutations. Nat
727 Genet. 51(1):36–41.

728 Seplyarskiy VB, Soldatov RA, Koch E, McGinty RJ, Goldmann JM, Hernandez RD,
729 Barnes K, Correa A, Burchard EG, Ellinor PT, et al. 2021. Population sequencing data
730 reveal a compendium of mutational processes in the human germ line. Science.
731 373(6558):1030–1035.

732 Seplyarskiy VB, Sunyaev S. 2021. The origin of human mutation in light of genomic
733 data. Nat Rev Genet. doi:10.1038/s41576-021-00376-2. [accessed 2021 Aug 18].
734 https://www.nature.com/articles/s41576-021-00376-2.

735 Serrano Colome C, Canal Anton O, Seplyarskiy V, Weghorn D. 2023. Mutational
736 signature decomposition with deep neural networks reveals origins of clock-like
737 processes and hypoxia dependencies. bioRxiv. doi:10.1101/2023.12.06.570467.
738 http://dx.doi.org/10.1101/2023.12.06.570467.

739 Silveira AB, Houy A, Ganier O, Özemek B, Vanhuele S, Vincent-Salomon A, Cassoux
740 N, Mariani P, Pierron G, Leyvraz S, et al. 2024. Base-excision repair pathway shapes
741 5-methylcytosine deamination signatures in pan-cancer genomes. Nat Commun.
742 15(1):9864.

743 Sollier J, Cimprich KA. 2015. Breaking bad: R-loops and genome integrity. Trends Cell
744 Biol. 25(9):514–522.

745 Sollier J, Stork CT, García-Rubio ML, Paulsen RD, Aguilera A, Cimprich KA. 2014.
746 Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced
747 Genome Instability. Mol Cell. 56(6):777–785.

Spisak N, de Manuel M, Milligan W, Sella G, Przeworski M. 2024. The clock-like accumulation of germline and somatic mutations can arise from the interplay of DNA damage and repair. PLoS Biol. 22(6):e3002678.

Spivak G. 2015. Nucleotide excision repair in humans. DNA Repair . 36:13–18.

Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov GV, Mirkin SM, Sunyaev SR. 2009. Human mutation rate associated with DNA replication timing. Nat Genet. 41(4):393–395.

Svejstrup JQ. 2002. Mechanisms of transcription-coupled DNA repair. Nat Rev Mol Cell Biol. 3(1):21–29.

Takahashi DT, Burguiere-Slezak G, Van der Kemp PA, Boiteux S. 2011. Topoisomerase 1 provokes the formation of short deletions in repeated sequences upon high transcription in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 108(2):692–697.

Tomkova M, McClellan MJ, Crevel G, Shahid AM, Mozumdar N, Tomek J, Shepherd E, Cotterill S, Schuster-Böckler B, Kriaucionis S. 2024. Human DNA polymerase ε is a source of C>T mutations at CpG dinucleotides. Nat Genet. 56(11):2506–2516.

Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. 2017. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 14(6):565–571.

Vermeulen W, Fousteri M. 2013. Mammalian transcription-coupled excision repair. Cold Spring Harb Perspect Biol. 5(8):a012625.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 131(4):281–285.

Wilson Sayres MA, Makova KD. 2011. Genome analyses substantiate male mutation bias in many species. Bioessays. 33(12):938–945.

Wood RD, Mitchell M, Sgouros JG, Lindahl T. 2001. Human DNA Repair Genes. Science. 291:1284–1289.

Xia B, Yanai I. 2022. Gene expression levels modulate germline mutation rates through the compound effects of transcription-coupled repair and damage. Hum Genet. 141(6):1211–1222.

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(2):248–262.e21.

Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 16(7):409–420.

Zhang Y, Yan Z, Qin Q, Nisenblat V, Chang H-M, Yu Y, Wang T, Lu C, Yang M, Yang S,

et al. 2018. Transcriptome Landscape of Human Folliculogenesis Reveals Oocyte and Granulosa Cell Interactions. Mol Cell. 72(6):1021–1034.e4.

Zheng CL, Wang NJ, Chung J, Moslehi H, Sanborn JZ, Hur JS, Collisson EA, Vemula SS, Naujokas A, Chiotti KE, et al. 2014. Transcription restores DNA repair to heterochromatin, determining regional mutation rates in cancer genomes. Cell Rep. 9(4):1228–1234.

Table 1. Significant predictors remaining from the Poisson multiple regression of gene expression from 13 testis cell types, replication timing, GC content and phased fraction on male mutation rate (n=15,562 genes). All predictors were re-scaled and non-significant cell stages were dropped; a logged offset term was included to account for mutational opportunities per gene. See Materials and Methods for details.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-10.057249	0.005816	-1729.250	< 2e-16	***
Spermatogonial stem cell	0.030820	0.008406	3.666	0.000246	***
Late primary spermatocyte	-0.066079	0.008747	-7.554	4.22e-14	***
Early sperm	0.036732	0.007419	4.951	7.38e-07	***
Replication timing	-0.066396	0.004258	-15.591	< 2e-16	***
GC content	0.129151	0.006474	19.948	< 2e-16	***
Phased fraction	0.034065	0.010635	3.203	0.001359	**

Table 2. Significant predictors remaining from the quasipoisson multiple regression of gene expression from the 10 folliculogenesis cell types, replication timing, GC content, and phased fraction on female mutation rate (n=15,562 genes). All predictors were re-scaled and non-significant cell stages were dropped from the model reported here. See Materials and Methods for details.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-11.429684	0.013995	-816.681	< 2e-16	***
Primordial oocyte	0.194260	0.018677	10.401	< 2e-16	***
Replication timing	-0.198322	0.009374	-21.157	< 2e-16	***
GC content	0.187022	0.015033	12.440	< 2e-16	***
Phased fraction	0.067055	0.012717	5.273	1.36e-07	***

Figure 1. Raw sex-specific mutation rates and germline averaged expression levels. A) Male germline mutation rates are not significantly associated with male germline averaged expression levels after adjusting for replication timing, GC content, and phased fraction (see Table S1 for full details of the regression model). B) Female germline mutation rates are significantly positively associated with female germline averaged expression levels after adjusting for replication timing, GC content, and phased fraction (Table S2). Grey points represent individual genes (n=15,562) and the loess lines (span=0.75) are plotted for the male (blue) and female (red) data. P-values are for the significance of the slope of expression in the full regression models.

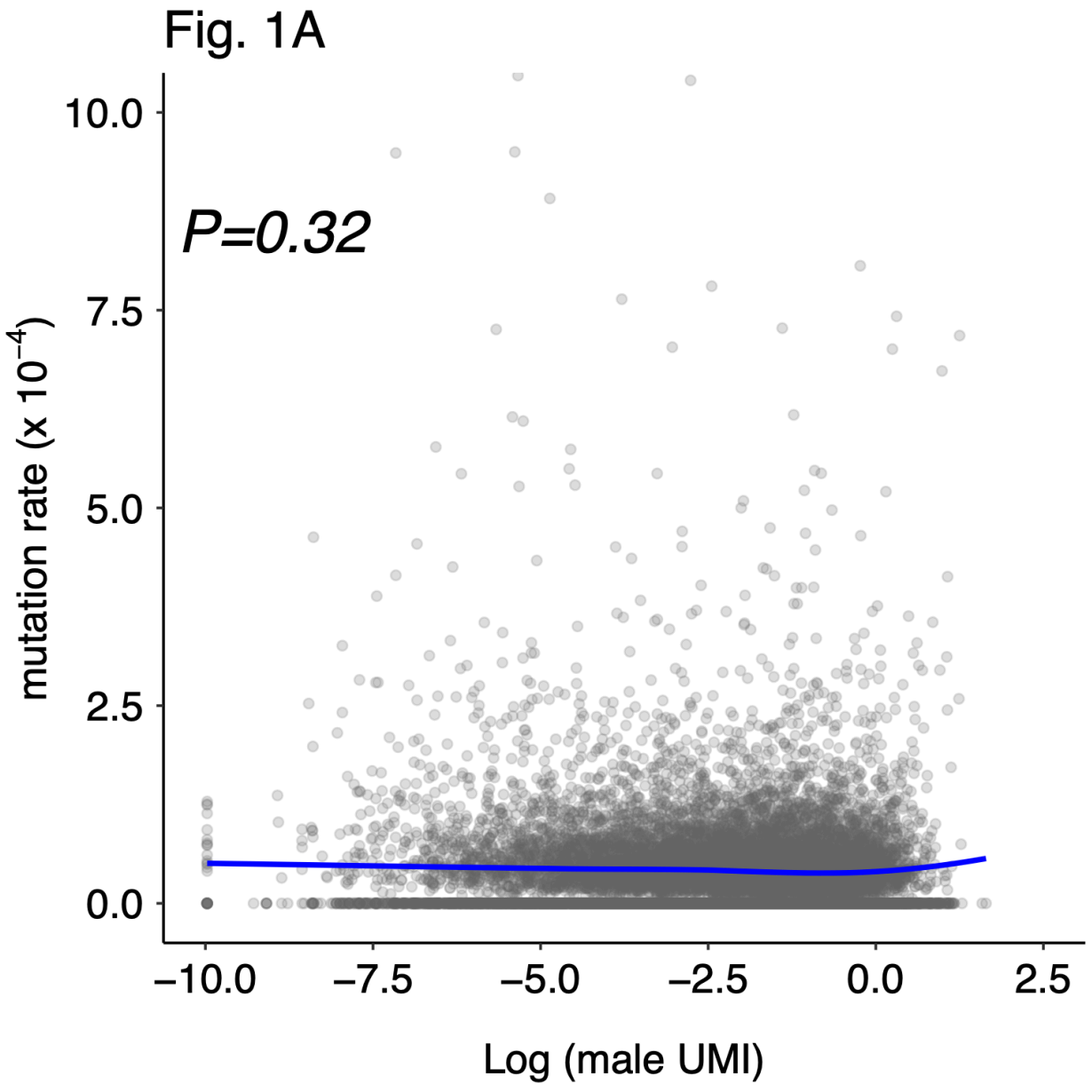
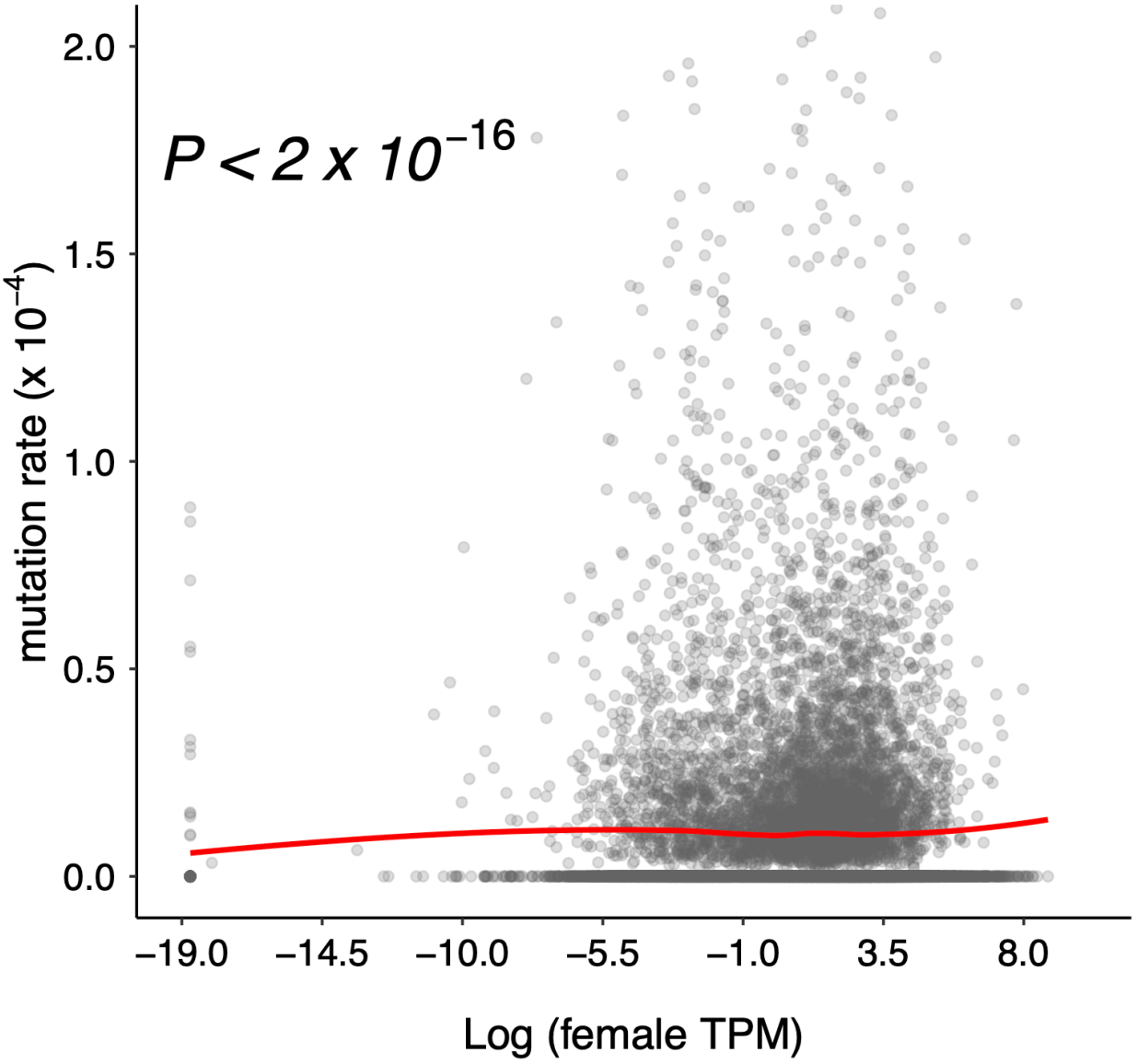


Fig. 1B


$P < 2 \times 10^{-16}$
mutation rate (x 10^{-4})
2.0
1.5
1.0
0.5
0.0
−19.0
−14.5
−10.0
−5.5
−1.0
3.5
8.0
Log (female TPM)

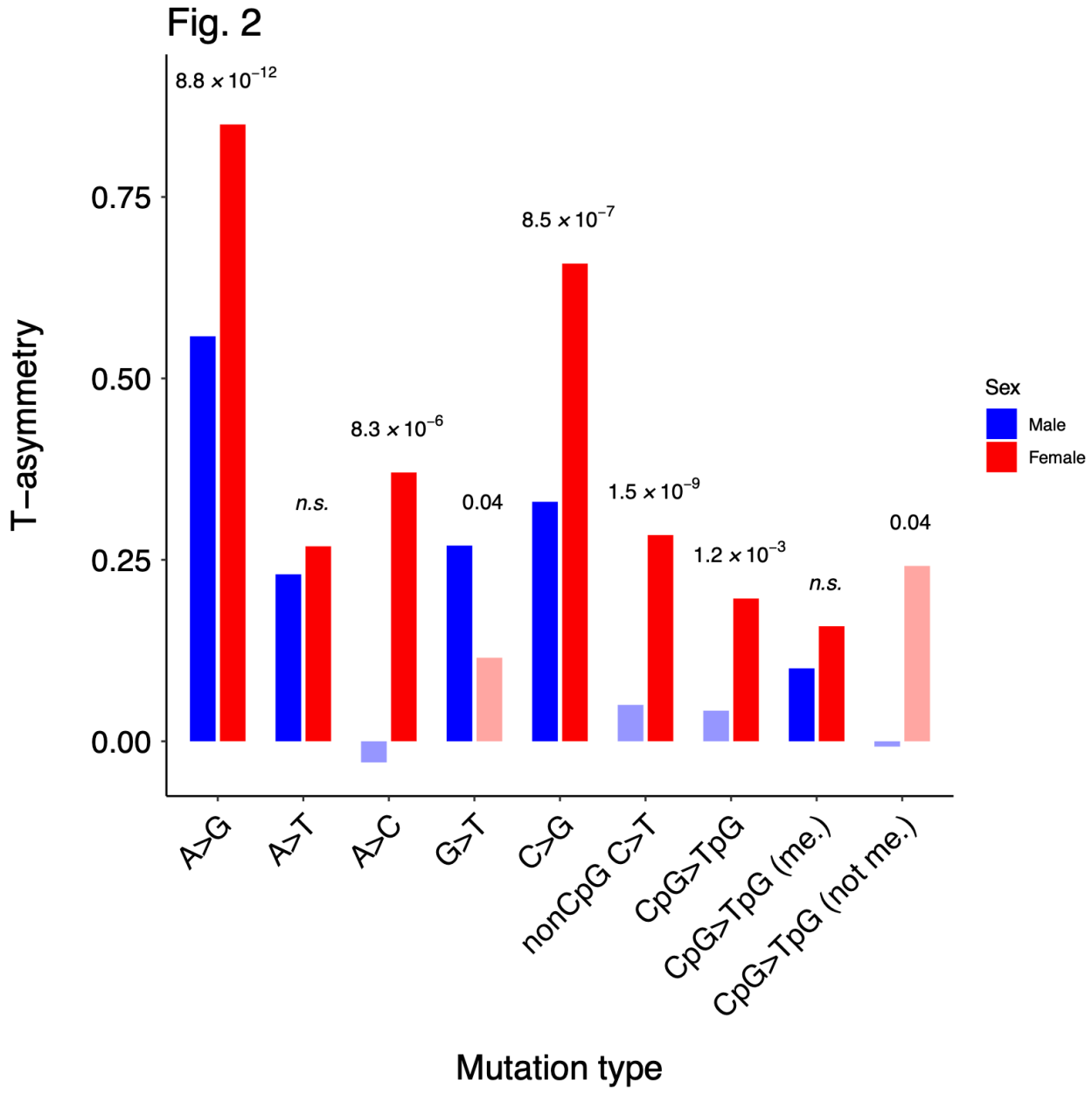



Figure 2. T-asymmetry across mutation types. Darker bars indicate sex-specific T-asymmetry (log base 2 of the ratio of the non-transcribed to transcribed strand mutation rates) that is statistically different from zero after multiple testing corrections across male and female values (FDR adjusted P<0.05). Lighter bars indicate the lack of sex-specific T-asymmetry. A second set of FDR adjusted P-values are provided above a pair of bars for a two-tailed test of difference between the male and female T-asymmetry values. See Results for details regarding methylation data for the last two pairs of bars (n.s., not significantly different; me., methylated).