nature genetics

Article

https://doi.org/10.1038/s41588-023-01562-0

A mutation rate model at the basepair resolution identifies the mutagenic effect of polymerase III transcription

In the format provided by the authors and unedited

Supplementary methods

1. A log-link model for sparse compartments

A simpler log-link model was included for pentamer-compartment categories with few observed SNVs. For instance, certain sequence contexts that are rare in promoters. For this model we estimated the effect of each covariate independently (no multiple regression), and multiplied the odds ratios to obtain the probability of not observing an SNV:

$SNV^{\sim} \sim N_{-6} * N_{-5} * N_{-4} * N_{-3} * N_3 * N_4 * N_5 * N_6 *$ *local_mutation_rate*

The likelihood of this model was then also compared to the case of a constant mutation rate (one parameter) within that category of sites. The fraction of the genome fit best by each model in the held-out 50% of sites is shown in Supplementary Figure 11.

2. Application of an infinite sites model to single genomic positions

Assuming constant mutation rate μ along a particular lineage of length t, the infinite-sites model states that the number of mutations in that lineage will be $Pois(\mu t)$. Let S be the number of mutations in the genealogy of a sample of individuals. It follows that $S \sim Pois(\mu T_{tot})$, where T_{tot} is the total length of genealogy, because T_{tot} is the sum over all branches in the genealogy. However, T_{tot} varies along the genome due to the inherent stochasticity of the coalescent process, and is also affected by linkage to selected sites. For a non-recombining locus, a category applicable to single sites in the human genome, the variance of S is $Var(S) = \mu E(T_{tot}) + \mu^2 Var(T_{tot})$ (Watterson 1975)[7]. The Poisson approximation is therefore applicable when $\mu^2 Var(T_{tot})$ is small. In a large sample from a constant-size population $Var(S) = \theta \frac{\sum_{i=1}^{n-1} 1}{i} + \pi^2 \frac{\theta^2}{6}$, where $\theta = 4N\mu$, N is the effective population size, and n is the sample size. Since $\theta \ll 1$, for large n, $Var(S)$ approaches $E(S)$ and the distribution of S becomes approximately Poisson (Ewens 2004, pg. 299)[8]. Indeed, even the distributions of the number of low counts SNVs (singletons, doubletons, etc.) are approximately independent and Poisson distributed as the sample size becomes very large[9]. The variance in T_{tot} is therefore negligible.

The recent growth of the human population results in genealogies with a greater proportion of T_{tot} residing in branches with fewer descendants compared to constant size populations. Summing over a larger number of branches with fewer descendants decreases $Var(T_{tot})$ compared to a constant-sized population, increasing the accuracy of the Poisson approximation. Therefore, for a sample as large as gnomAD v3 (71,702 individuals), we can expect $SPois(\mu T_{tot})$. Please see Wakeley et al. (2022) for a detailed investigation into the applicability of Poisson models to single sites.

3. Filtering

We marked regions and individual sites as low quality based on quality criteria from gnomAD, abnormal density of SNV sites, and on suspicious patterns of recurrence.

More specifically, we classified sites as low quality if Umap100 mappability was below 0.5, if the site overlapped with a long (>50 nucleotide) simple repeat, if in a hundred nucleotide window the mean ReadPosRankSum was above 1, or if the number of segregating SNVs in a hundred nucleotide window was zero.

After we obtained predicted mutation rates for every site, we noticed that within some rate categories the site frequency spectra (SFS) has an abnormally high fraction of high frequency variants. We noticed that this problem is context specific. To separate problematic sites for each pair of pentamer and mutation rate we calculated fraction of high frequency SNVs [MAF >0.005 and MAF <=0.2] and compared it to the average

fraction of high frequency SNVs across all mutation types; the pentamer was labeled unreliable if the fraction of high frequency sites is 1.5-fold higher than the mutation rate-controlled average (See Supplementary Figure 9). Most sites masked this way belong to repetitive contexts like AAATT>T, TTAAA>T, AATAT>A or CTCTA>A (Table S3).

4. Correcting for mutational hotspots

Comparing the predicted and observed number of rare SNVs shows that some regions have much higher mutation densities than expected. While we attributed some of these mutational hotspots to transcription by polymerase III, for other regions the biological etiology was unclear. To recalculate the mutation rate in hypermutable regions we ran additional logistic regressions among hotpots – 100 nucleotide windows with more than 75 rare SNVs. For this regression we used previously estimated mutability and mutation type to predict an adjusted hypermutable mutation rate. We applied a similar procedure to adjust for higher mutation rates at transcription factor binding sites (TFBS), whereas as with additional variables we used the type of transcription factor, distance from the center of the CHIP-seq peak, overlap with a promoter and tissues where the factor is active. We provide both adjusted and initial mutation rate predictions. We used an analogous procedure to re-calibrate mutation rate in contexts that have atypical SFS.

5. Genomic features for downstream analyses

Chip-seq tracks were downloaded from the Vorontsov *et. al.*[11] and only category A data (highest quality) were used. Chip-seq signal from different overlapping TFBS were counted independently (Figure 4).

DHS tracks were downloaded from ENCODE (https://www.encodeproject.org/search/?type=Experiment&assay_title=DNase-seq). We aggregated DHS peaks from 4 adult tissues (lung, stomach, leg muscle, brain) to obtain the "adult" DHS track and from 4 embryonic/fetal tissues (fibroblast, placenta, large intestine, stomach) to obtain the "fetal" DHS track, finally we aggregated two tracks from testis to obtain the testis track. In order to obtain peaks private to "fetal" and "adult" tissues, we excluded peaks that overlapped between them or with "testis". Testis, in contrast, includes ubiquitous peaks, but using the private testis track does not change our results.

Annotations of active genes and pseudogenes that are transcribed by polymerase III (tRNA, snRNA, vault RNA, RNA component of 7SK nuclear ribonucleoprotein, ribonuclease P RNA component H1, ribosomal RNA, Ro60-associated Y RNA) were downloaded from HUGO Gene Nomenclature Committee (https://www.genenames.org/data/gene-symbol-report/). It has been reported that some ALU repeats are also transcribed by polymerase III[12], and we downloaded coordinates of such ALU elements from ref. [12].

We used Ensembl for genic annotations and definition of promoters. We defined promoters as a region 0 to 2 KB upstream of the CDS containing the gene. We do not exclude regions that match our definition of promoters for more than one transcript.

6. Demographic inference

In order to determine whether mutation rate estimates from Roulette are sufficient to capture distortions to the site frequency spectrum due to recurrent mutation, we fitted a demographic model and assessed how well it matched the observed SFS in each mutation rate bin. We based our demographic model on those presented by Gao and Keinan (2016)[13] and Gazave *et. al.*, (2014)[14]. The Gazave model starts with a constant population size, followed by two bottlenecks (one of which is the out-of-Africa bottleneck), and ends with a recent population growth. Gao and Keinan (2016) used the parameters from Gazave et al. (2014), but re-estimates the parameters for the recent exponential growth phase. They conclude that the best demographic model has a faster-than-exponential growth, with growth speed parameter $b = 1.12$ ($b = 1$ is equivalent to exponential growth). We keep the model from Gao and Keinan (2016), but refit the parameters b and the initial growth rate g, which they estimated as 0.0055.

We use a Wright-Fisher simulator used in Weghorn *et. al.*, (2019)[15] to simulate the site-frequency spectrum a dense grid of $b = [1.1,1.3]$ and $g = [0.005,0.006]$. Though this may seem like a narrow range, the final population size varies from 0.7 million to 40 million. Here, the log-likelihood is computed as:

$$(b, g) = \sum_{i=0}^{56885} C_i * logE[b, g],$$

where C_i is the folded synonymous allele counts for non-Finnish Europeans (NFE) in gnomAD v.2.1.1 whole exomes dataset and $E[b, g]$ is the expected folded SFS given parameters *b* and *g*. Under this framework we find that maximizes the above likelihood.

We utilize the full mutation rate information to find the maximum log-likelihood. For each parameter, we simulate the SFS for all the mutation rate bins and calculate the log-likelihood for each mutation rate bin. Then, we sum up the log-likelihood over the bins and find the growth parameters that maximizes this summed likelihood.

Then, to compare how well the SFS fits within different possible mutation rate bins we re-fit μ using the maximum likelihood demographic parameters. We do this for all Roulette bins used in other analyses, as well as for low and high-rate defined as [1.3e-09, 3e-09] and [1e-07, 2.8e- 07]. The Wright-Fisher simulations allow for recurrent mutation, so the SFS changes shape as the mutation rate increases. We measure the fit to the shape of the SFS by calculating the likelihood conditional on sites being polymorphic by removing the zero bin and normalizing the remaining expected SFS. We evaluate the information added by Roulette's fine-scale mutation rate estimates by comparing the conditional likelihoods of the low and high-rate fits to the μ fit specifically to that bin.

To evaluate the information added to demographic modeling by high rate sites we compute the average per-polymorphism contribution to the log-likelihood. We calculate the likelihood using the first 40 entries of the SFS under a model of pure exponential growth with recurrent mutation (Wakeley et al. 2022). As an example, we use the best fitting parameter $\beta = N_0 \frac{r}{n} \approx 6$ to the rare SFS in the gnomAD v2 data, where $N_0$is the current population size, r is the per-generation growth rate, and n is the sample size. The parameter $\theta_0 = 4N_0\mu$ was matched to estimates in Roulette bins to provide comparison points at 3e-09, 3e-08, and 3e-07.

7. Estimating fraction of ancestral variants in training dataset

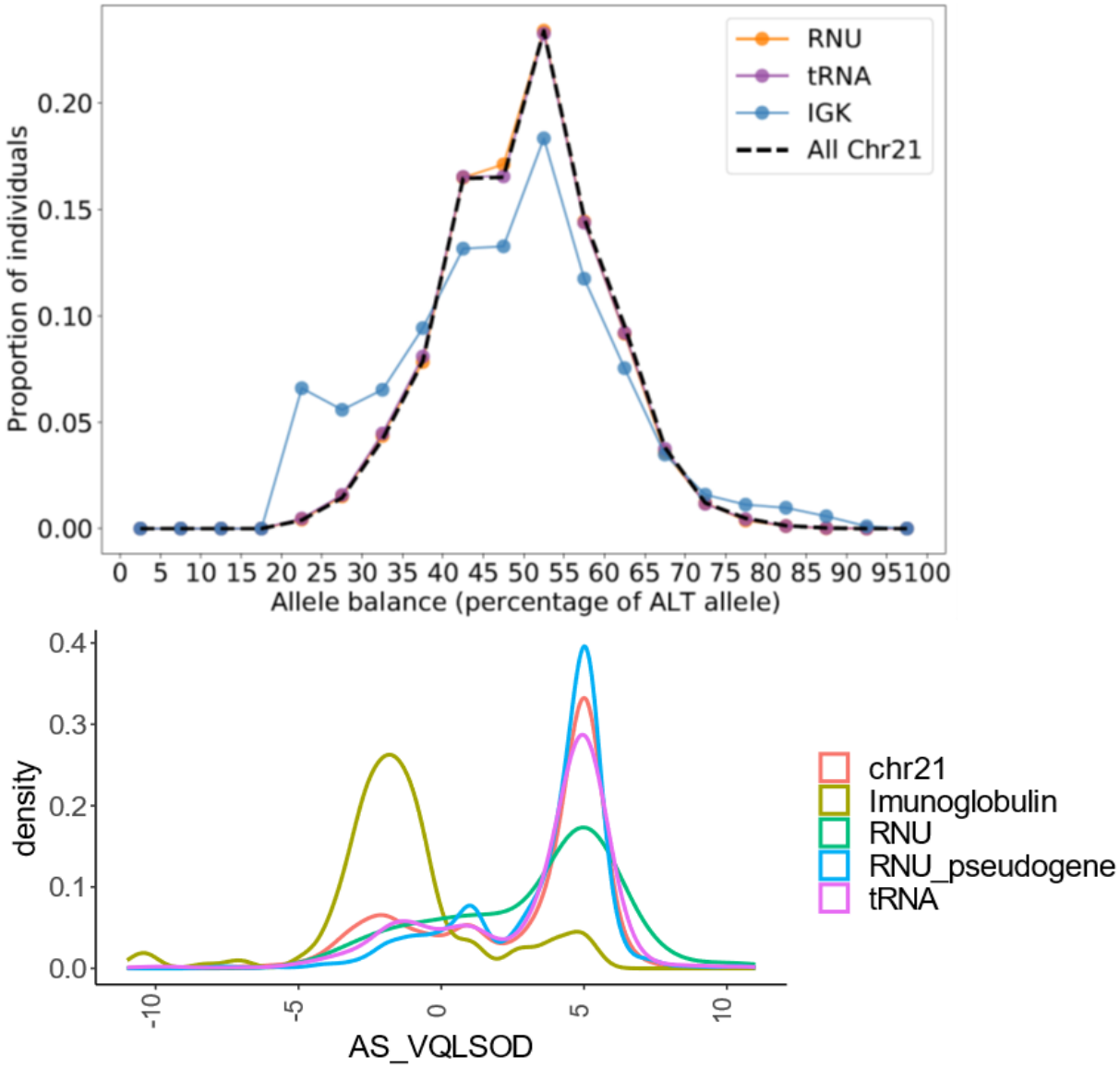
When fitting Roulette we assumed that alternative alleles with frequency < 0.001 were derived and therefore represented mutation events from the reference to the alternative state. This assumption can be violated if an alternative allele observed at frequency x is really an ancestral allele at frequency 1-x. In order to bound the fraction of rare variants (MAF < 0.001) used in training that are ancestral, we used the demographic fit of non-Finnish Europeans (NFE) mentioned above. We simulated the unfolded SFS of a high mutation rate class (3e-07 mutations per site per generation) using the Wright-Fisher simulator used for the demographic analysis. However, the gnomAD v3 dataset used to fit Roulette includes individuals with ancestry labels other than NFE, most of whom are of African descent. Since it was not feasible to fit another, more complex, demographic history, we used simulations of an equilibrium population to capture a greater range of potential SFS. The probability the minor allele is ancestral will be greatest for the highest frequency under consideration, so we report values for 0.001. For the high mutation rate class, we get that $1.25*10^{-5}$ and $3.09*10^{-5}$ of the rare variants are ancestral in NFEs and in equilibrium population, respectively.

8. The distribution of quality metrics in hypermutable gene classes

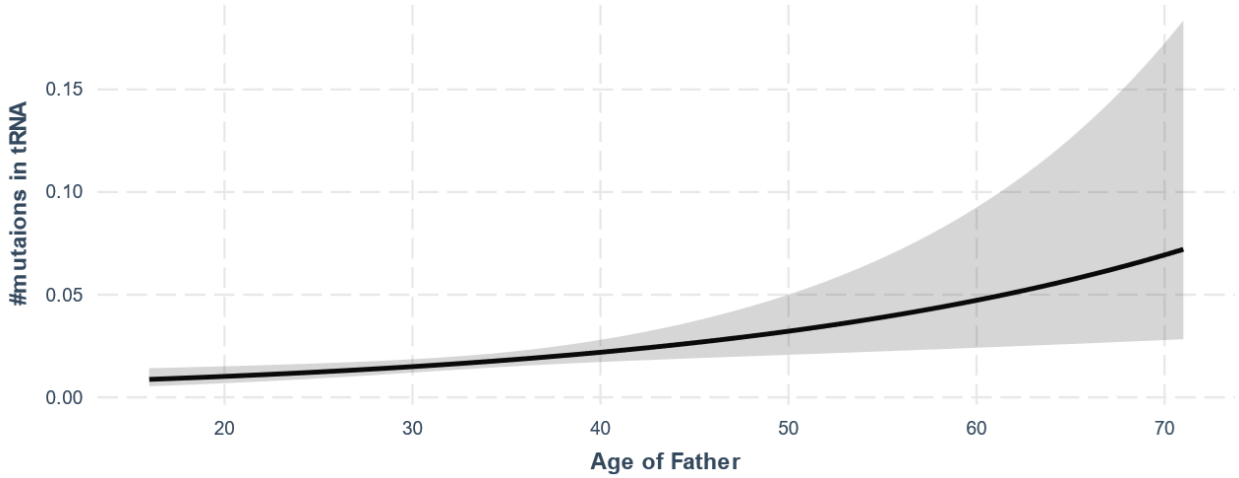
To evaluate variants within hypermutable regions we used three quality metrics provided by gnomAD. Allelic balance is determined by the relative count of reads with the alternative and reference alleles. This should be concentrated around 50% for true germline variants, while variants representing somatic mutations will show up

at frequencies <50% for the derived allele. We also examined mapping quality scores, which reflect the relative mapping likelihoods of alternative versus reference reads, in order to diagnose mis-mapping artifacts. These are shown in Supplementary Figure 2 and indicate that mapping quality is better in RNU and tRNA genes compared to the genomic background on chromosome 21, while scores in IGK are worse. Finally, we look at an overall allele-specific variant quality score (AS_VQSLOD) to capture any other factors. No *de novo* mutations were reported in IGK genes, consistent with little hypermutability.

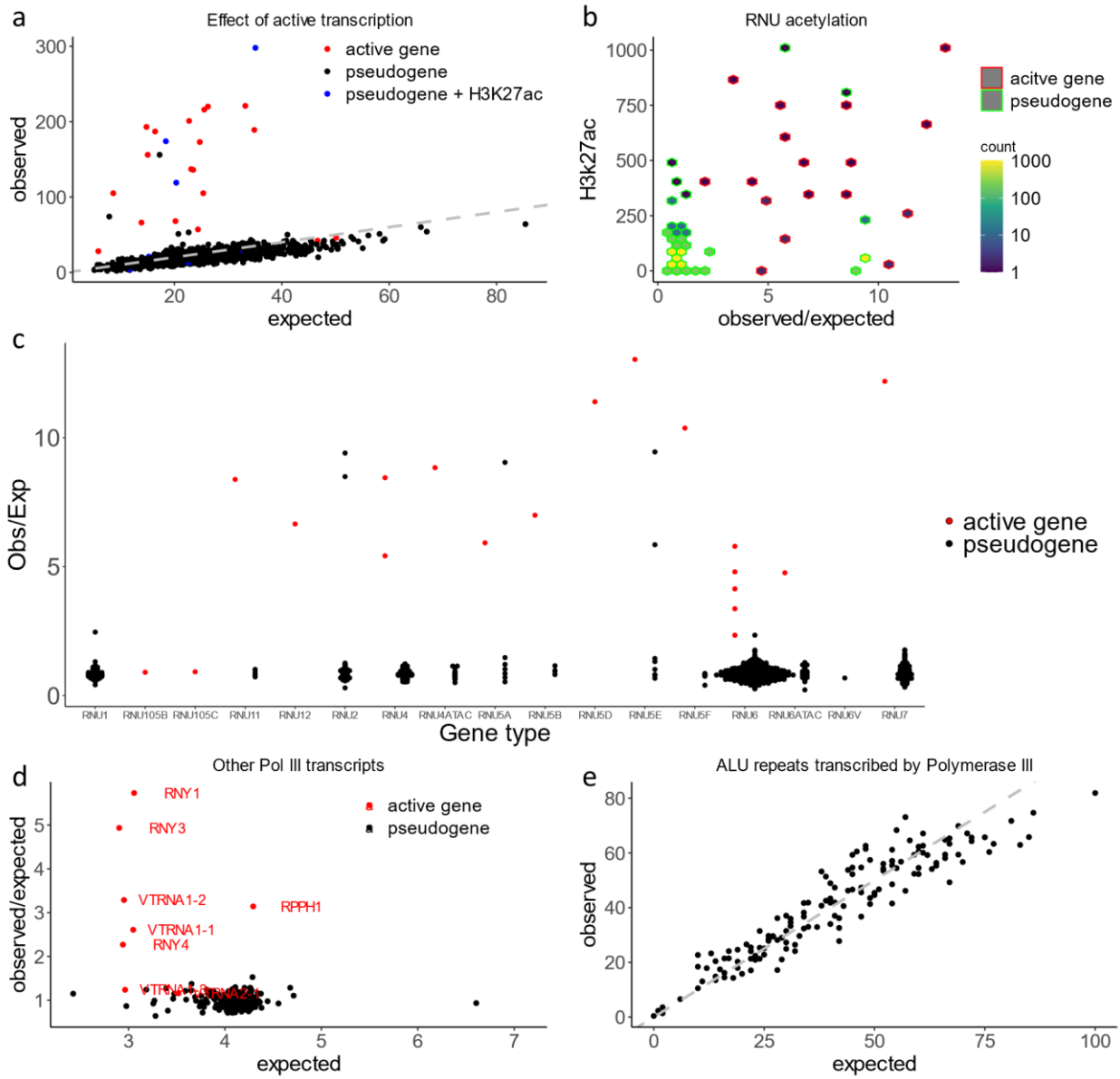
Supplementary Figures



Supplementary Figure 1. Quality metrics for hypermutable regions. a) The allelic balance within RNU and tRNA genes match the control curve for all sites on chromosome 21, while IGK genes deviate substantially from the background. b) AS_VQSLOD which is the main metric for the quality of the variant is dramatically decreased for IGK.

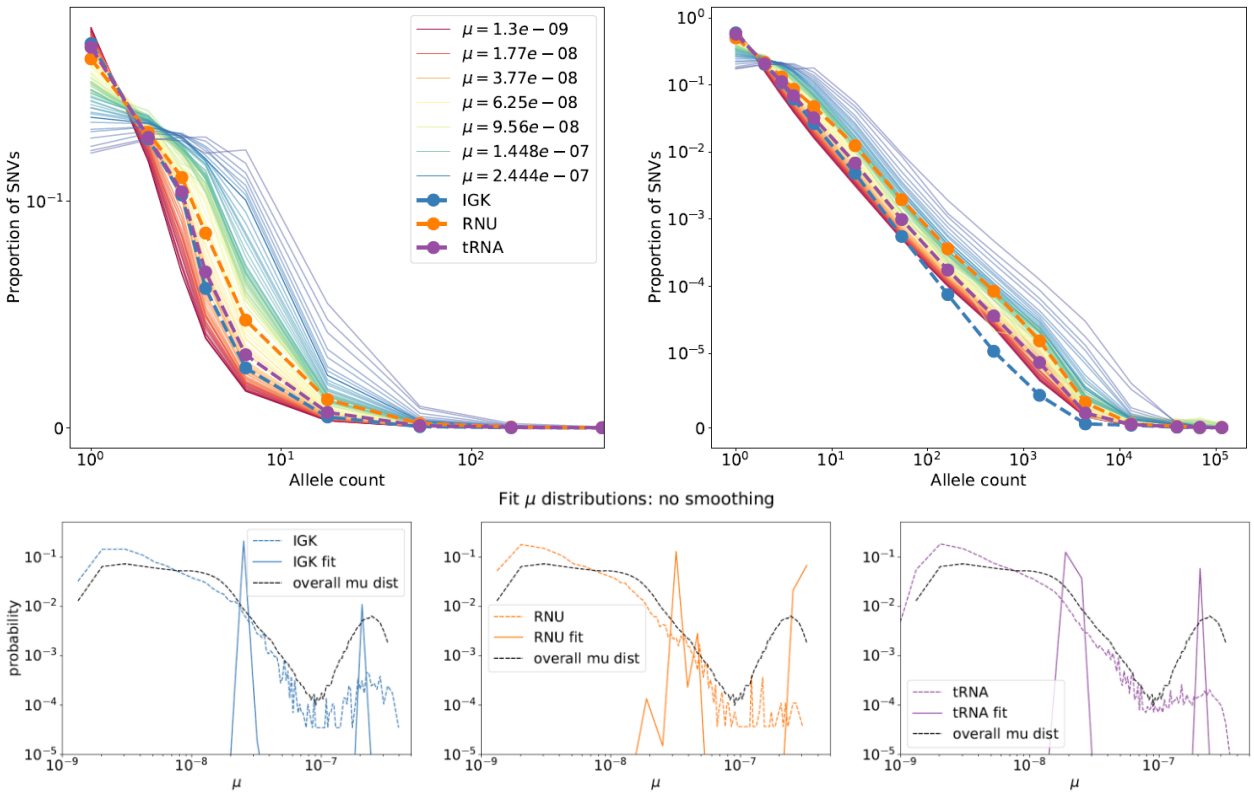


Supplementary Figure 2. Parental age effect. The best fit for the relationship between the number of mutations occurring in tRNA sites and the parental age. The fit is for the exponential dependency on age using Poisson regression. While only 104 *de novo* mutations occurred in tRNA genes, the association between parental age and the number of mutations is highly significant ($p=9.9*10^{-5}$, results remain significant for the linear relationship between age and mutation count: $p=5.6*10^{-3}$). Only 18 mutations happened in RNU genes making analysis of the age effect unreliable. Error bars are 95% confidence intervals for t-statistic.

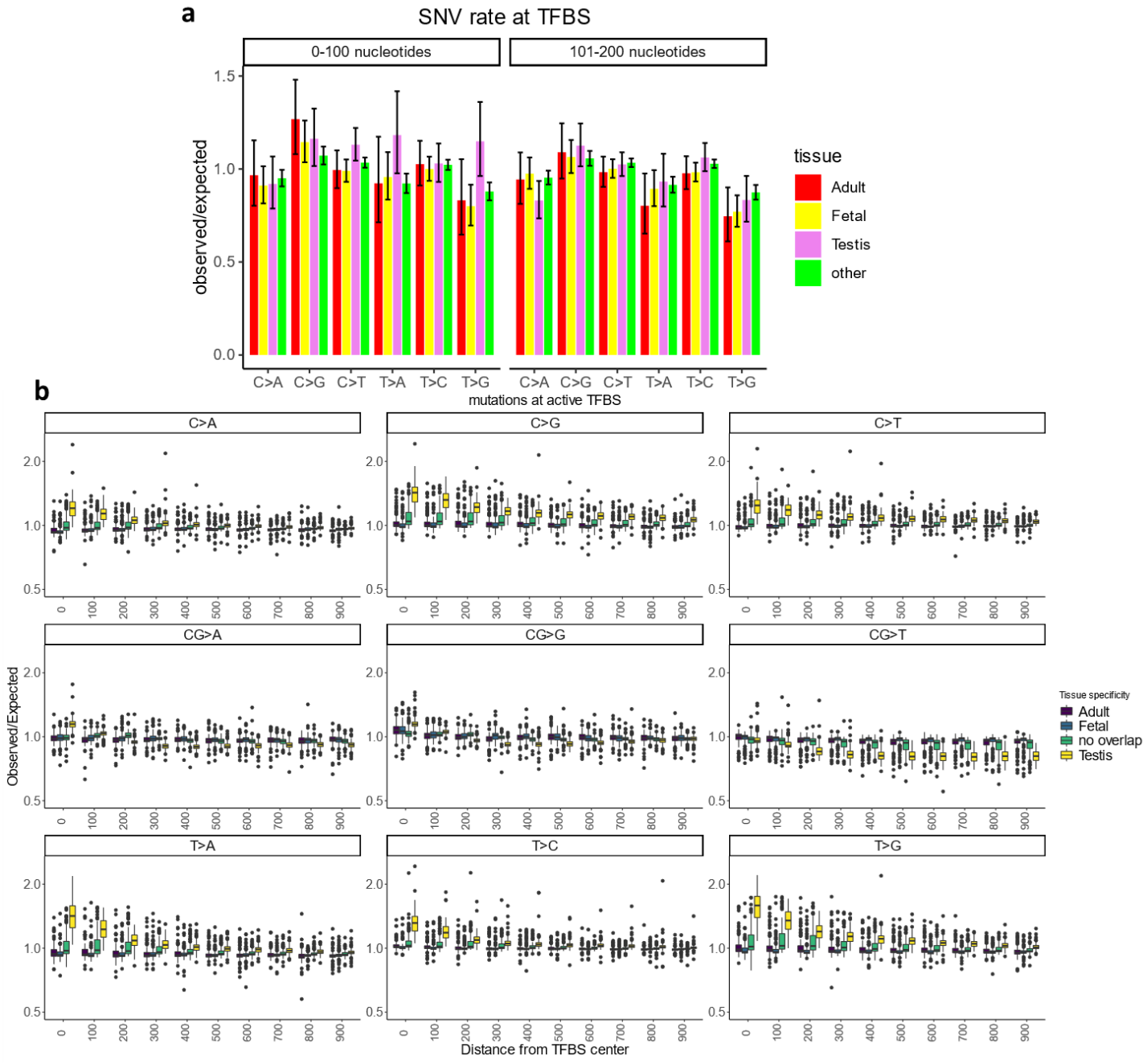


Supplementary Figure 3. Effect of polymerase III transcription

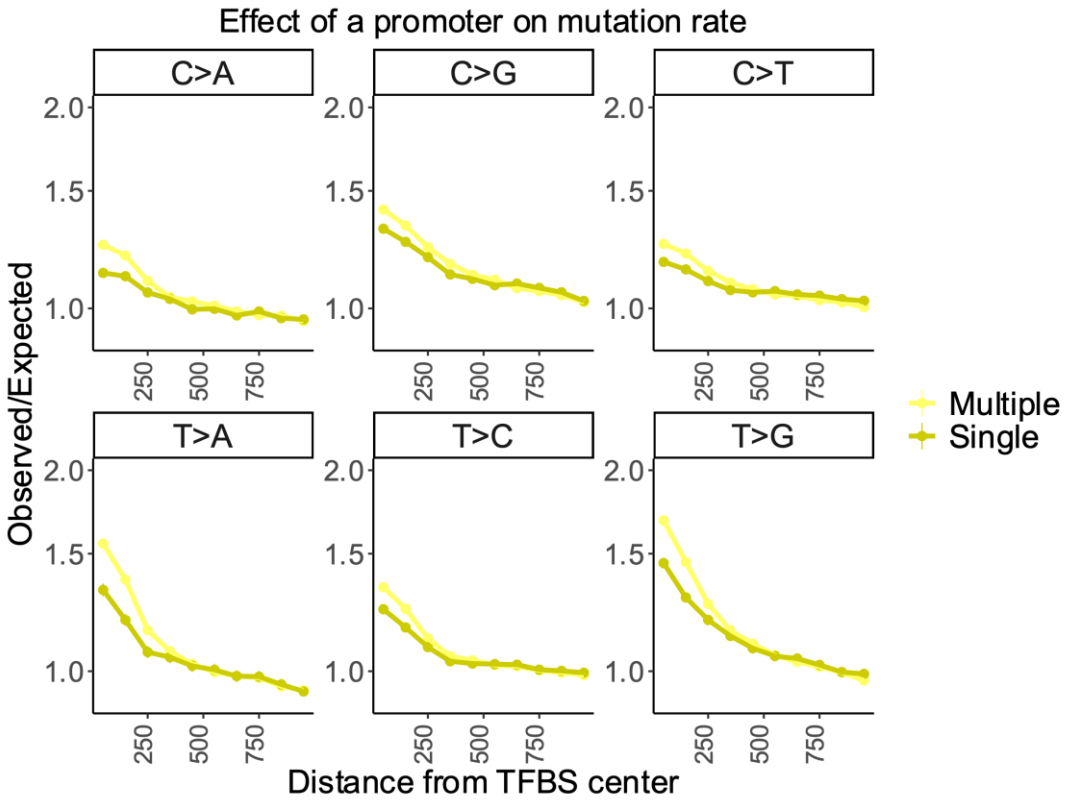
a,b) The effect of acetylation and pseudo/active gene stratification (according to the HUGO annotation) on the SNV number in RNU genes. b) While the vast majority of RNU pseudogenes do not overlap with H3k27ac peaks, four out of five RNU pseudogenes with high mutation rate overlap H3k27ac. c) Active transcription increases mutation rate across different classes of RNU genes. d) Other classes of Pol III transcript, represented by a small number of genes also have elevated mutation rate e) The density of rare SNVs is in line with the Roulette predictions in ALU elements that have been predicted to be transcribed by polymerase III.



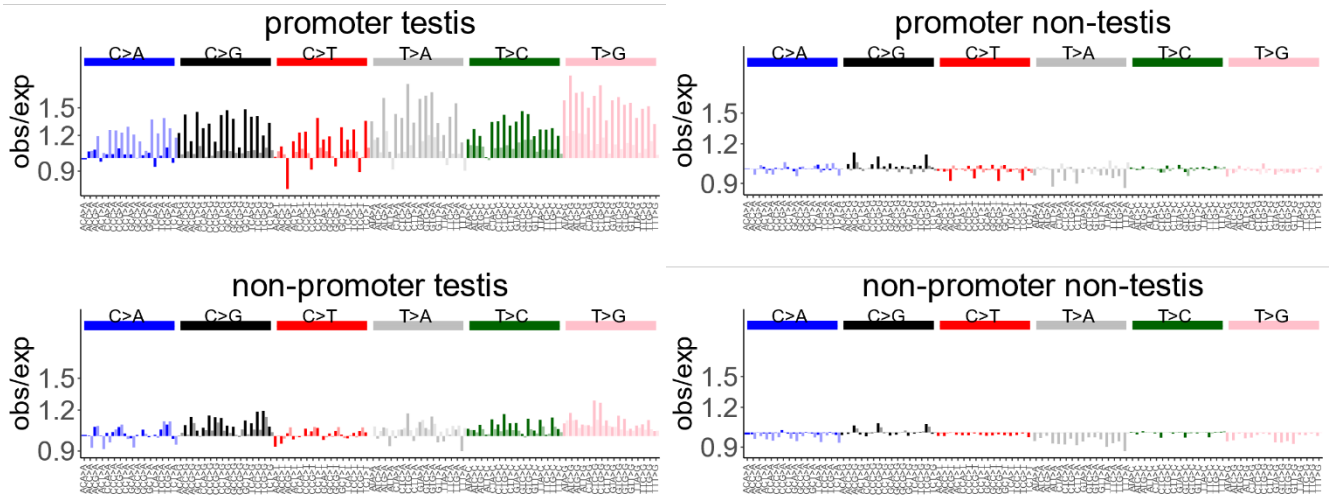
Supplementary Figure 4. Site frequency spectra (SFS) for sites with different mutation rates and for different gene categories. Rare variants on the top left and full SFS to the top right. On the bottom the mutation rate distributions for observed SNVs for different gene categories was estimated by fitting the SFS in these genes as a mixture of SFS shapes observed in Roulette bins. In contrast to Figure 4 c, d we did not use genome-wide mutation rate distribution as a prior.



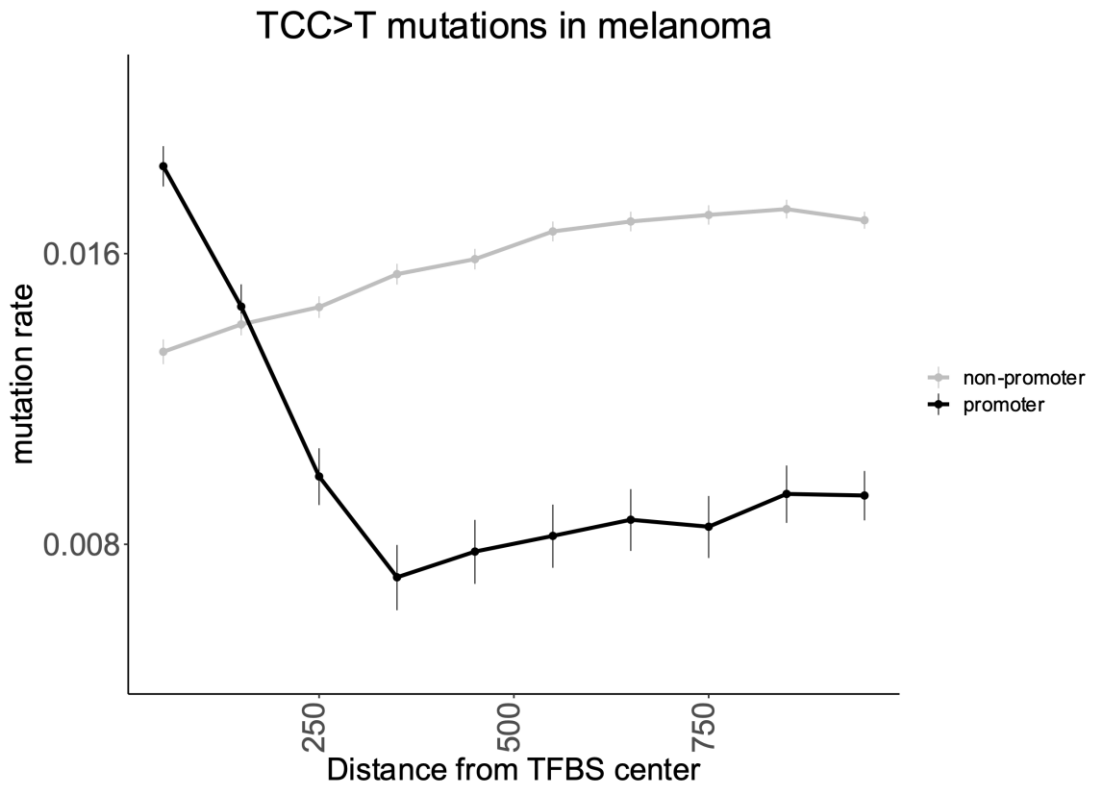
Supplementary Figure 5. Mutation rate is accelerated at TFBS active in testis
a) Observed to expected ratio of *de novo* mutations at TFBS active in different tissues. Distance from the TFBS center is shown on top. Error bars show 95% confidence intervals for the ratio of two Poisson variables. b) Observed to expected ratio of rare SNVs at TFBS active in different tissues. Panels are stratified by mutation type.



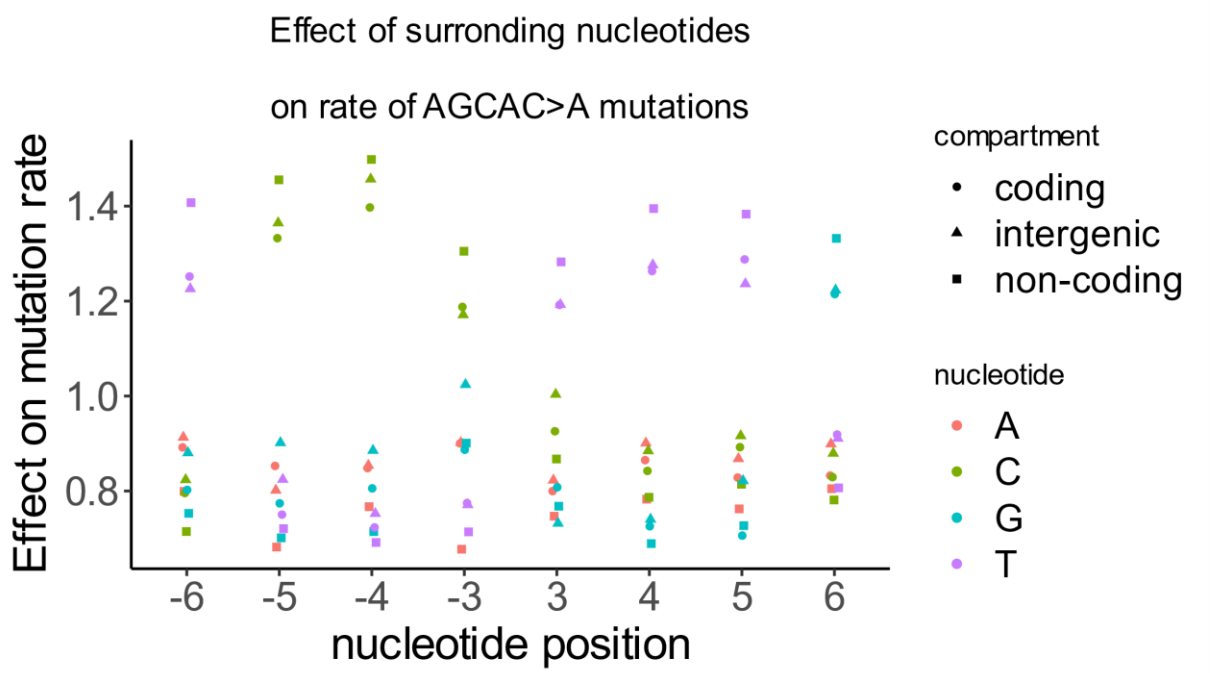
Supplementary Figure 6. Mutation rate is accelerated at TFBS active and overlapping multiple promoters. We compared observed/expected mutation rate for non- CpG mutations overlapping DHS in testis. These TFBS have higher mutation rate if they also overlap multiple promoters (light yellow) instead of a single promoter (dark yellow).



Supplementary Figure 7. Correction of mutation rate at TFBS
Panels show the observed to expected density of rare SNVs at TFBS active in Testis or in other tissues as well as whether TFBS do or do not overlap a promoter. Bright colors are reflecting deviation from the model before the correction for higher mutability at TFBS, pale colors correspond for corrected values.

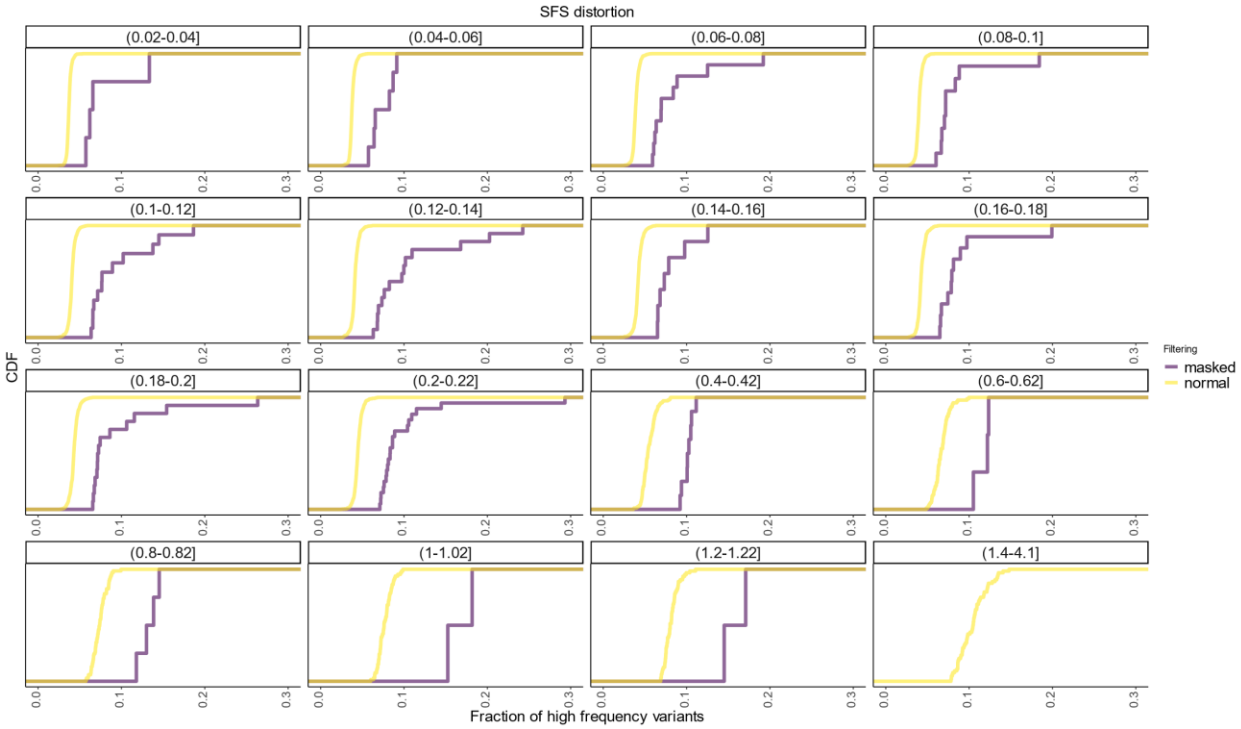


Supplementary Figure 8. UV-induced mutations in melanoma. We analyzed TFBS overlapping DHS of foreskin melanocytes and measured the rate of TCC>T mutations (major UV-induced mutation type). Mutation rates in melanoma samples (ICGC data) differ between TFBS sites overlapping and not overlapping promoters. Error bars show 95% Poisson confidence intervals. Results are in line with Mao et al, Nature Comms. 2018

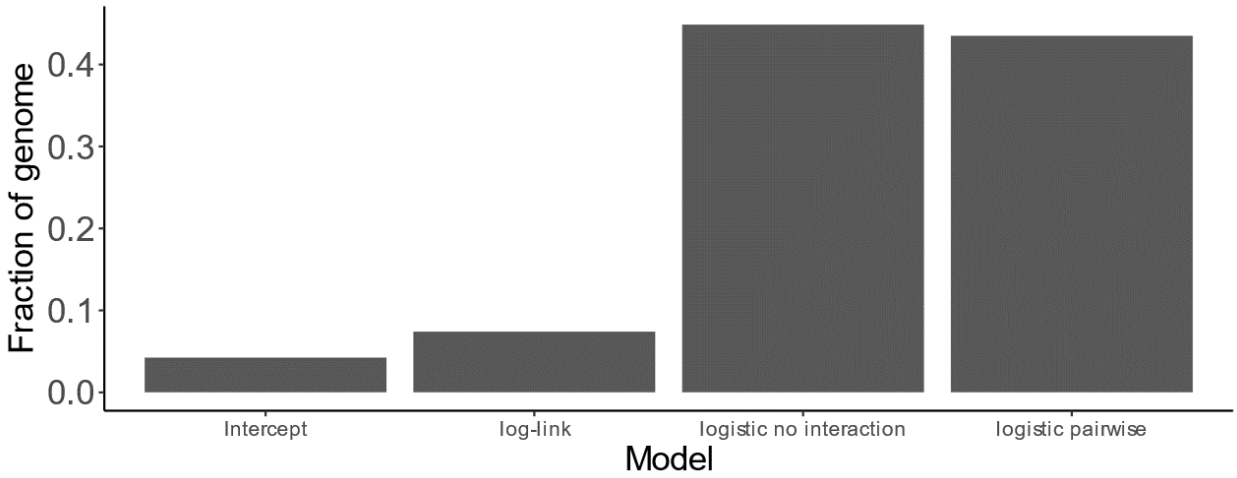


Supplementary Figure 9. Effect of surrounding nucleotides is varying between the coding strand, the non-coding strand within genes and intergenic regions

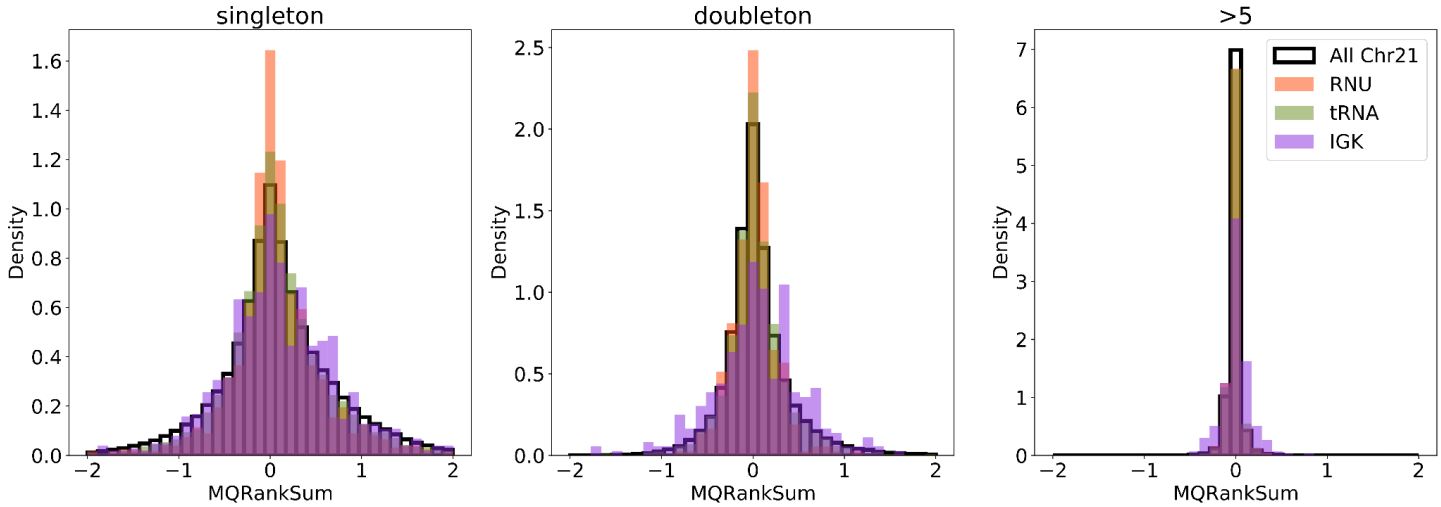
The figure shows the effect of nucleotides beyond the pentamer on the rate of AGCAC>A mutations. The effect differs between genes and intergenic regions and is strand-dependent within genes. The AGCAC>A mutation type is shown as an example.



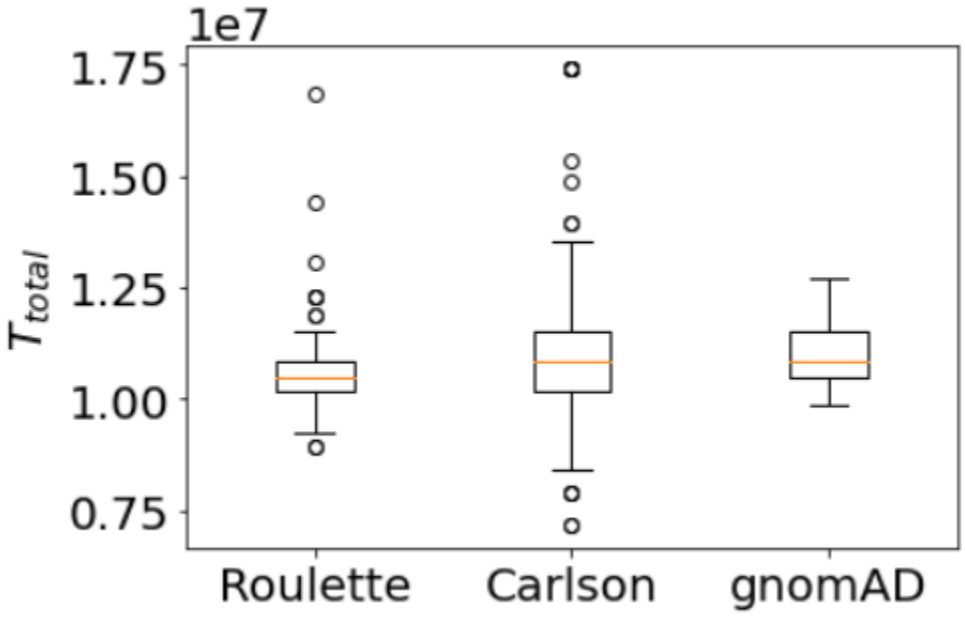
Supplementary Figure 10. Filtering pentameric contexts with abnormal patterns of site frequency spectra
Site frequency spectrum (SFS) is dependent on mutation rate for rare SNVs. For each pair of pentamer and predicted mutation rate (shown on top of the panels) we calculated the fraction of high frequency variants (MAF >0.005 and MAF <=0.2). We masked the context if it has a proportion of high frequency variants exceeding mean for the same mutation rate multiplied by 1.5. We show empirical cumulative distributions for the portion of high frequency variants for masked (purple) and non-masked pentamers.



Supplementary Figure 11. Fraction of the genome where each of four tested models fits best. We selected the best of four models on a 50% hold out test for each genomic compartment.



Supplementary Figure 12. Distributions of mapping quality scores for apparent hypermutable gene classes. We show the distribution of MQRankSum scores for SNVs in RNU, tRNA, and IGK genes compared to the background distribution from chromosome 21. Distributions were computed from the 1K genomes subset of gnomAD v3.



Supplementary Figure 13. Distributions of estimated the total coalescent time (T_{total}) across mutation rate bins in gnomadV3 for three mutation rate models