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

Vladimir Seplyarskiy[1,2,*], Daniel J. Lee[1,2,*], Evan M. Koch[1,2,*], Joshua S. Lichtman[3], Harding H. Luan[3], Shamil R. Sunyaev[1,2]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Brigham and Women's Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA
[3]NGM Biopharmaceuticals, South San Francisco, CA, USA
*Contributed equally

***De novo* mutations occur with substantially different rates depending on genomic location, sequence context and DNA strand[1–4]. The success of many human genetics techniques, especially when applied to large population sequencing datasets with numerous recurrent mutations[5,6], depends strongly on assumptions about the local mutation rate. Such techniques include estimation of selection intensity[7], inference of demographic history[8], and mapping of rare disease genes[9]. Here, we present Roulette, a genome-wide mutation rate model at the basepair resolution that incorporates known determinants of local mutation rate (http://genetics.bwh.harvard.edu/downloads/Vova/Roulette/). Roulette is shown to be more accurate than existing models[1,10]. Roulette has sufficient resolution at high mutation rate sites to model allele frequencies under recurrent mutation. We use Roulette to refine estimates of population growth within Europe by incorporating the full range of human mutation rates. The analysis of significant deviations from the model predictions revealed a 10-fold increase in mutation rate in nearly all genes transcribed by Polymerase III, suggesting a new mutagenic mechanism. We also detected an accelerated mutation rate within transcription factor binding sites restricted to sites actively utilized in testis and residing in promoters.**

The human single nucleotide mutation rate is known to vary along the genome at different scales[4,11,12]. Some of this variation is explained by the combination of mutation type and immediately adjacent nucleotides, conceptualized as the mutation spectra[10,13]. The CpG di-nucleotide context induces by far the largest spectrum effect because of the high rate caused by the mutagenic effect of methylation at cytosines that are followed by guanine[14]. Previous studies demonstrated that the extended sequence context, well beyond the two adjacent bases, exerts an additional effect on mutation rate[1,15,16]. Still, mutation spectra vary along the genome, indicating that rate differences are not fully explained by the surrounding DNA sequence[4,11]. Some of this variability tracks DNA properties like replication timing and gene expression[4]. Other effects, such as local spikes of clustered mutations in oocytes, lack obvious epigenetic correlates[4,17,18]. In addition to regional variation, the rates of many mutation types depend on the DNA strand[4,19–21]. The transcription direction alters the mutation spectra between transcribed and non-transcribed strands, and the replication direction gives rise to differences between leading and lagging strands.

We developed "Roulette" a mutation rate model that incorporates these factors and more (see Methods). Each nucleotide has three potential mutations, and we hereafter refer to each of these potential mutations as a site. The extended sequence context is included by estimating the effect of the 6 upstream and 6 downstream nucleotides adjacent to each site (Figure 1a). Due to sparsity, it is impossible to accurately estimate the effect of each unique 12-nucleotide context. To account for this, we estimated the effect of the central pentamer (two nucleotides on either side) separately from the individual effects of the 8 more distant nucleotides, which are included as covariates (Figure 1a,b). For epigenomic features, Roulette incorporates methylation level (for CpG transitions and CpG transversions), transcription direction, gene expression level in testis (for mutations within gene bodies), and quantitative estimates of replication direction (Figure 1a,c). The incorporation of transcription and replication directions makes the model strand-dependent with unequal rates for mutations of

the same type on the two DNA strands. To our knowledge, strand-dependency has not been incorporated into existing context-dependent and regional mutation models[1,10,22]. In addition to the known epigenomic features listed above, unexplained regional variation in the mutation rate has been observed[17,18]. The Roulette model accounts for this residual variation by including the observed mutability of each of tri-nucleotide context in 50KB windows (Figure 1d). Mutation probabilities were modeled using logistic regression with pairwise interactions between all covariates. We fit models for each pentamer separately, thereby allowing the effects of covariates to vary independently among pentamers (see Methods). Finally, we group the predicted rates into 100 bins to add utility to the estimates while losing minimal information.

It is impossible to fit parameter-rich mutation models to currently available *de novo* mutation datasets because of data sparsity. To train Roulette, we collected all non-coding SNVs with frequency below 0.001 from gnomAD v3 whole genomes[10]. The distribution of very rare non-coding SNVs along the genome is primarily driven by mutation rate differences with the effects of biased gene conversion, direct and background selection being negligible[1]. Due to the sample size of contemporary human sequencing data, many rare SNVs represent recurrent mutations that have been introduced into the population multiple times. To correct for recurrence, Roulette fits the probability that a site remains monomorphic and uses a simple population genetics approximation to transform estimates to the mutation rate scale (see Methods).

After estimation and rescaling, we found that Roulette captures expected genomic mutation rate variation in test sites not used in model training. For instance, nearly two-fold rate differences the transcribed and non-transcribed strands are predicted accurately (Figure 1c). Additionally, the importance of Roulette's regional correction is illustrated by DNA segments that are hypermutable in oocytes (sometimes called regions of maternal mutagenesis)[17,18,23,24]. Maternal mutagenesis is responsible for a localized increase in C>G mutations on the left arm of the chromosome 8 (Figure 1d, Supplementary Figure 1a). The direct inclusion of mutation density at the 50KB windows is sufficient to account for spatial genomic features acting on long scales. Therefore, despite not using replication timing, histone modifications, or recombination rate[25] as covariates, Roulette is able to capture associations between mutability and these epigenetic factors (Supplementary Figure 1b).

As a second point of validation, we tested whether Roulette estimates resolve the old riddle of "cryptic variation." Early comparative genomics literature[12,26–28] observed that the frequency of triallelic SNVs is higher than expected based on the probability of pairs of biallelic SNVs assuming independence and a three nucleotide mutational model. Roulette accurately predicts the probability of triallelic SNVs (Supplementary Figure 2), suggesting that previous observations of "cryptic variation" reflected residual mutation rate variance in earlier models associated with extended nucleotide context and local genomic factors.

To further validate the performance of Roulette, we compared it with two existing mutation rate models. One uses trinucleotide context and methylation levels was applied by Karczewski et al. (2020)[10] to the study of the gnomAD and the other developed by Carlson et al. (2018)[1] uses heptamer context with several epigenetic features including methylation levels (see Supplementary Table 1 for a more detailed description of model differences). Because the Karczewski et al. (2020)[10] mutation rate model was not publicly available, we recreated it using gnomAD v3 and the authors' descriptions. We hereafter refer to these models as gnomAD and Carlson.

While previous studies evaluated goodness of fit of mutation rate models[1], none to our knowledge have attempted to estimate the remaining residual variance. We used two novel site-by-site metrics to analyze each model's ability to predict the rate and location of synonymous SNVs from gnomAD v2 whole exomes (not used in model fitting) as well as of *de novo* mutations in one whole exome family sequencing study[29] and whole genome family sequencing from two studies[18,30]. The first metric is an adjusted version of Nagelkerke's pseudo-R^2 for logistic models[31] that measures the residual variance between the observed and expected likelihood

given the inherently stochastic nature of mutational processes. Pseudo-R^2 assumes that there is no variance among sites with the same predicted mutation rate, so that errors result solely from misclassification among mutation rate bins. We therefore developed a second per-site metric that estimates this additional variance within bins using observations of multiple mutations occurring at the same site. We compare the rate of *de novo* mutations at sites where an SNV was observed to the rate at sites without SNVs. If mutation rates are estimated without error, the rate of *de novo* mutations in both groups should be equal. This SNV-conditional method calculates the mean of the conditional mutation rate distribution depending on whether an SNV is observed or not and uses this to estimate the overall within-bin variance. Both methods necessarily require assumptions about the true distribution of mutation rates. For pseudo-R^2, this is that the full distribution is well-captured by the model even if per-site estimates are subject to error, and for the SNV-conditional this is that the distribution of true mutation rates within each category of sites predicted to have the same rate is log-normal.

Roulette predicts the rate of synonymous SNVs with higher accuracy than the gnomAD and Carlson models, reaching a pseudo-R^2 of 0.91 compared to 0.83 and 0.86 respectively (Figure 2a). The same was true for *de novo* variants, where Roulette reached 0.92 and 0.99 on the exome *de novo* and genome *de novo* datasets respectively, compared to gnomAD (0.85, 0.92) and Carlson (0.86, 0.96). At SNV positions we found an excess of *de novo* mutations, as expected in the presence of residual mutation rate variation. The mean excess was 34% within Roulette bins, 47% within Carlson bins, and 94% within gnomAD bins (Supplementary Figure 3). These result in estimated residual variances of 19%, 25%, and 51% for the Roulette, Carlson, and gnomAD models (Figure 2b). While overall residual variances are larger for the SNV-conditional method, Roulette still explains around 5% more of the variance in human mutation rates than the Carlson model. The increased performance of Roulette relative to Carlson is most pronounced at high-rate CpG transitions.

Many population genetics applications rely on aggregated mutation rate estimates by gene or within a genomic window. We tested the relevance of Roulette for these applications by aggregating synonymous sites by gene for gnomAD.v2 and predicting the number of SNVs. Aggregate estimates generated using Roulette are more accurate than those for gnomAD or Carlson (Figure 2c). There are 1758 genes with a Z-score greater than 2 or less than -2 for Roulette rates, which is significantly fewer outlier genes than 2468 for Carlson or 2295 for the gnomAD model. An area of population genetics inference with important applications in human disease genetics is the estimation of selective constraints for protein truncating variants (PTVs). All methods to infer strong selection rely on estimates of local mutation rate. We recomputed estimates of two measures of strong heterozygous selection, s_{het} and LOEUF[7,10], using Roulette mutation rates. The new estimates (available at http://genetics.bwh.harvard.edu/genescores/selection.html) show a slight but notable improvement in detection of autosomal dominant disease genes annotated in DD2G and ClinVar (Supplementary Table 2).

We next evaluated the importance of precise mutation rate estimates for the inference of demographic history (specifically, historical changes in effective population size) from the site frequency spectrum (SFS, the number of observed alleles at each frequency)[8,32]. Most studies rely on the "infinite number of sites" model and assume that only single mutation events contribute to each segregating site[33]. As a result, under neutrality, the relative distribution of allele frequencies only depends on genealogies and is independent of mutation rate while the overall level of variation is linearly dependent on mutation rate. The presence of recent recurrent mutations breaks this assumption and induces a dependency between the shape of site frequency spectrum and mutation rate[5,6] (Figure 3a, Supplementary Figure 4). As a result, site frequency spectra differ among mutation rate classes making inference challenging. On the other hand, a set of SFS curves at different mutation rates provides rich additional information about demographic history that increases power and eliminates biases due to recurrent mutations.

We evaluated the ability of Roulette to model the shape of the SFS across the range of mutation rates by re-fitting a model of European demographic history from Gao and Keinan (2016)[8] using simulations that allowed

for recurrent mutations[34]. We revise the demographic model by fitting to the whole range of mutation rates. The inclusion of high mutation rate sites is meaningful because these are more informative about population growth than low-rate sites (Figure 3c). We revise demography by updating the growth rate acceleration parameter from 1.120 to 1.122 and initial growth rate from 0.0050 to 0.0057 with a final population size estimated at 8.1 million compared to 2.5 million. This model fits the shape of the SFS well even as the mutation rate becomes large enough that recurrent mutation substantially skews to shape towards less rare variants[5,6] (Figure 3a). The fine-scale mutation rate bins defined by Roulette provide a much better fit to the SFS shape than can be achieved by dividing mutations into only two bins, one for low rates and one for high (Figure 3b). This is due to sufficient recurrent mutation within the higher end of the low-rate bin and sufficient rate variation within the high-rate bin to make single-rate summaries inadequate to capture the shape of the SFS. While one solution is to filter sites with high mutation rates so that the infinite sites assumption remains reasonable, this comes with a loss of more informative sites on a per-SNV level (Figure 3c). This utility extends to selection inference where it is possible to identify individual strongly constrained sites when mutation rates are in the neighborhood of 1e-07 per generation[27].

While much of mutation rate variation is adequately captured by Roulette (Figure 1, 2) including various epigenetically active sites like enhancers and promoters (Supplementary Figure 5), strong local deviations can be used to identify new mutagenic mechanisms in humans. Regional variation in mutation rates and spectra have previously been characterized and biologically interpreted at scales exceeding 10kb. However, many mutagenic mechanisms arise due to epigenetic factors acting at much shorter scales. Data sparsity prevents the application of unsupervised statistical techniques to characterize variation at short scales[4]. We analyzed extreme deviations from Roulette predictions at the 100bp scale genome-wide (Figure 4a). The choice of the scale is determined by the need to balance the need for high resolution and statistical power.

While many extreme deviations likely represent unfiltered artifacts, the most striking observation is that a third of 100bp genomic windows with extremely high mutation rate unexplained by the Roulette features lie within RNA genes transcribed by polymerase III (Pol III) or within immunoglobulins kappa. These outlier windows harbor over 70 SNVs per 100bp, and some up to 100 SNVs. The two most prominent gene classes transcribed by Pol III are tRNA and small nuclear RNA genes (RNU) (Figure 4a, b). Elevated mutation rates in tRNA genes have been recently noted by a comparative genomics study[35], although unaccounted for recurrent mutations masked the magnitude of the effect. Similarly, despite huge 7-fold increase in SNV rate of RNU, recurrent mutations lead to underestimation of hypermutability (Figure 4a,b). However, analysis of *de novo* mutations observed in parent-child trio sequencing studies showed a 32-fold (19-50, 95% Poisson CI) higher mutation rate. The much higher mutation rate at Pol III transcripts masks the effect of purifying selection and leads to an unrealistic selection inference[36].

To further validate the link between Pol III transcription and accelerated mutation rate, we compared mutation rates between active RNU genes and pseudogenes. Indeed, the increased mutation rate is almost exclusively limited to active genes suggesting that active transcription rather than genomic location or sequence context is responsible for the mutation rate increase (Figure 4C). The few exceptions are pseudogenes that show H3K27ac chromatin marks associated with active transcription (Supplementary Figure 6a), suggesting that apparent hypermutable RNU pseudogenes are misannotated active genes. The association between Pol III transcription and high SNV density extends to all other classes of non-coding RNAs (Figure 4D) but not to SINE repeats (Supplementary Figure 6b), which may also be transcribed by Pol III[37].

There are non-mutually exclusive explanations as to why transcription by Pol III is strongly mutagenic. First, unlike RNA polymerase II (Pol II), Pol III does not have the ability to recruit transcription coupled repair (TCR). However, TCR removes only mutations on one of the two strands and cannot reduce the mutation rate by more than a half, thus it alone cannot explain the magnitude of the effect. Second, transcription associated mutagenesis (TAM) is a well-described phenomenon in yeasts[38] and is attributed primarily to ribonucleotide

incorporation into DNA during transcription. The third possibility would involve a previously uncharacterized mechanism associated with transcription specific to Pol III that is highly different from Pol II in protein composition[39]. Interestingly, it was recently shown that damage-induced mutations can accumulate on the non-transcribed strand outside of replication[4,40]. However, this mechanism creates a very strong mutational asymmetry that we were not able detect for Pol III transcripts, implying other mechanisms. Last, transcription initiation by the transcription factor (TF) IIIB triggers restructuring of the DNA-bound Pol III. This restructuring can be mutagenic and by itself create mutational hotspots upstream of RNU genes (Figure 4b).

Immunoglobulin kappa genes also exhibit long stretches of extreme hypermutability along with slightly different spectra (Fig 4e, f). Blood sample sequencing could be affected by clonal hematopoiesis (CH) raising the possibility that mutations in immunoglobulins are somatic rather than germline. Indeed, we observed an excess of SNVs in a few genes associated with CH (BCL6, LPP promoter, TCL1A, SNX29). However, there is little evidence that CH is associated with hypermutability of immunoglobulin kappa.

Transcription factor binding occurs at short scales and has been shown to be highly mutagenic in yeast and human cancers either because of blocked resection of ribonucleotide primers introduced by polymerase alpha, interference with the access of nucleotide excision repair, or altered DNA conformation[41–44]. First, we attributed TFBS activity to specific tissues by overlapping ChIP-seq signals with regions of open chromatin measured by DNase I hypersensitivity. In the majority of TFBS, Roulette predicts mutation rates accurately, confirming that the observed mutation rate elevation within TFBS is due to sequence context and regional features[45] included in Roulette (Figure 5a). TFBS active in testis are a notable exception characterized by increases in the germline mutation rate over the background for most mutation types (Supplementary Figure 7a, b), with the strongest effect for T>G mutations (median increase across TFs is 1.59-fold, Figure 6a). This observation strongly suggests a direct mutagenic effect of transcription factor binding. Interestingly, binding of SNPC4, the factor responsible for RNU transcription, has the strongest (6-fold) impact on mutation rate.

Furthermore, we found that the higher mutation rates are almost exclusively restricted to TFBS in promoters (Figure 5b). Moreover, TFBS overlapping multiple promoters have higher mutation rate than TFBS overlapping a single promoter (Supplementary Figure 8). To expand the utility of Roulette we also provide mutation rates corrected for this TFBS effect (see methods, Supplementary Figure 9). Interestingly, UV-induced mutations at TFBS in melanoma also have different rates in and outside of promoters (Supplementary Figure 10). We found no significant deviation of the observed to expected ratio between leading and lagging strands, and between transcribed and non-transcribed strands (Figure 5c,d). This suggests that the effects of co-replicative ribonucleotide incorporation by polymerase alpha and hindrance of transcription coupled excision repair are unlikely explanations for elevated mutation rates at TFBS.

As shown above, Roulette is the most accurate human mutational model and has utility across different biological fields. Mutation rate estimates from the three analyzed models are made available here: http://genetics.bwh.harvard.edu/downloads/Vova/Roulette/. Future work may explain the sources of the demonstrated residual mutation rate variation, some of which may derive from evolving rates through time and variability between populations[46–48].

Code availability

Code used to preform the analysis could be accessed with https://github.com/vseplyarskiy/Roulette link

Data availability

Mutation rate estimates for autosomes http://genetics.bwh.harvard.edu/downloads/Vova/Roulette/

Shet values re-calculated with the help of Roulette could be found here http://genetics.bwh.harvard.edu/genescores/selection.html

Acknowledgements

We thank J. Wakeley and L. Fan for helpful suggestions on population genetics theory. We thank D.J. Balick for providing with a forward Wright-Fisher simulator. This research was supported by National Institutes of Health grants R35-GM127131, R01-MH101244, U01-HG012009, R01-HG010372, and R01-HG010372 along with funding from NGM Biopharmaceuticals.

References

1. Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *Nature Communications* **9**, 3753 (2018).
2. Carlson, J., DeWitt, W. S. & Harris, K. Inferring evolutionary dynamics of mutation rates through the lens of mutation spectrum variation. *Current Opinion in Genetics & Development* **62**, 50–57 (2020).
3. Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat Rev Genet* (2021) doi:10.1038/s41576-021-00376-2.
4. Seplyarskiy, V. B. *et al.* Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373**, 1030–1035 (2021).
5. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
6. Harpak, A., Bhaskar, A. & Pritchard, J. K. Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans. *PLOS Genetics* **12**, e1006489 (2016).
7. Cassa, C. A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat Genet* **49**, 806–810 (2017).
8. Gao, F. & Keinan, A. Explosive genetic evidence for explosive human population growth. *Curr Opin Genet Dev* **41**, 130–139 (2016).
9. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944–950 (2014).
10. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
11. Terekhanova, N. V., Seplyarskiy, V. B., Soldatov, R. A. & Bazykin, G. A. Evolution of Local Mutation Rate and Its Determinants. *Mol. Biol. Evol.* **34**, 1100–1109 (2017).
12. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766 (2011).
13. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

14. Ehrlich, M., Norris, K. F., Wang, R. Y., Kuo, K. C. & Gehrke, C. W. DNA cytosine methylation and heat-induced deamination. *Biosci. Rep.* **6**, 387–393 (1986).
15. Rodriguez-Galindo, M., Casillas, S., Weghorn, D. & Barbadilla, A. Germline de novo mutation rates on exons versus introns in humans. *Nat Commun* **11**, 3304 (2020).
16. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48**, 349–355 (2016).
17. Goldmann, J. M. *et al.* Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **50**, 487–492 (2018).
18. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
19. Green, P., Ewing, B., Miller, W., Thomas, P. J. & Green, E. D. Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics* **33**, 514 (2003).
20. Chen, C.-L. *et al.* Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.* **28**, 2327–2337 (2011).
21. Seplyarskiy, V. B. *et al.* Error-prone bypass of DNA lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nature Genetics* **51**, 36 (2019).
22. Bethune, J., Kleppe, A. & Besenbacher, S. A method to build extended sequence context models of point mutations and indels. 2021.12.06.471476 Preprint at https://doi.org/10.1101/2021.12.06.471476 (2021).
23. Jónsson, H. *et al.* Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
24. Wong, W. S. W. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat Commun* **7**, 10486 (2016).
25. Agarwal, I. & Przeworski, M. Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *PNAS* **116**, 17916–17924 (2019).

26. Hodgkinson, A., Ladoukakis, E. & Eyre-Walker, A. Cryptic Variation in the Human Mutation Rate. *PLOS Biology* **7**, e1000027 (2009).

27. Seplyarskiy, V. B., Kharchenko, P., Kondrashov, A. S. & Bazykin, G. A. Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. *Mol. Biol. Evol.* **29**, 1943–1955 (2012).

28. Johnson, P. L. F. & Hellmann, I. Mutation Rate Distribution Inferred from Coincident SNPs and Coincident Substitutions. *Genome Biol Evol* **3**, 842–850 (2011).

29. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* **180**, 568-584.e23 (2020).

30. An, J.-Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).

31. NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).

32. Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H. & Bustamante, C. D. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics* **5**, e1000695 (2009).

33. Crow, J. F. & Kimura, M. *An Introduction to Population Genetics Theory*. (The Blackburn Press, 2009).

34. Weghorn, D. *et al.* Applicability of the Mutation–Selection Balance Model to Population Genetics of Heterozygous Protein-Truncating Variants in Humans. *Molecular Biology and Evolution* **36**, 1701–1710 (2019).

35. Transfer RNA genes experience exceptionally elevated mutation rates | PNAS. https://www.pnas.org/doi/10.1073/pnas.1801240115.

36. Dukler, N., Mughal, M. R., Ramani, R., Huang, Y.-F. & Siepel, A. Extreme purifying selection against point mutations in the human genome. *Nat Commun* **13**, 4312 (2022).

37. Zhang, X.-O., Gingeras, T. R. & Weng, Z. Genome-wide analysis of polymerase III–transcribed Alu elements suggests cell-type–specific enhancer function. *Genome Res.* **29**, 1402–1414 (2019).

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

39. Abascal-Palacios, G., Ramsay, E. P., Beuron, F., Morris, E. & Vannini, A. Structural basis of RNA polymerase III transcription initiation. *Nature* **553**, 301–306 (2018).

40. Anderson, C. J. *et al.* Strand-resolved mutagenicity of DNA damage and repair. 2022.06.10.495644 Preprint at https://doi.org/10.1101/2022.06.10.495644 (2022).

41. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).

42. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nature Communications* **9**, 2626 (2018).

43. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016).

44. Reijns, M. A. M. *et al.* Lagging strand replication shapes the mutational landscape of the genome. *Nature* **518**, 502–506 (2015).

45. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

46. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 3439–3444 (2015).

47. Narasimhan, V. M. *et al.* Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nature Communications* **8**, 303 (2017).

48. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *eLife* **6**, e24284 (2017).

Figures

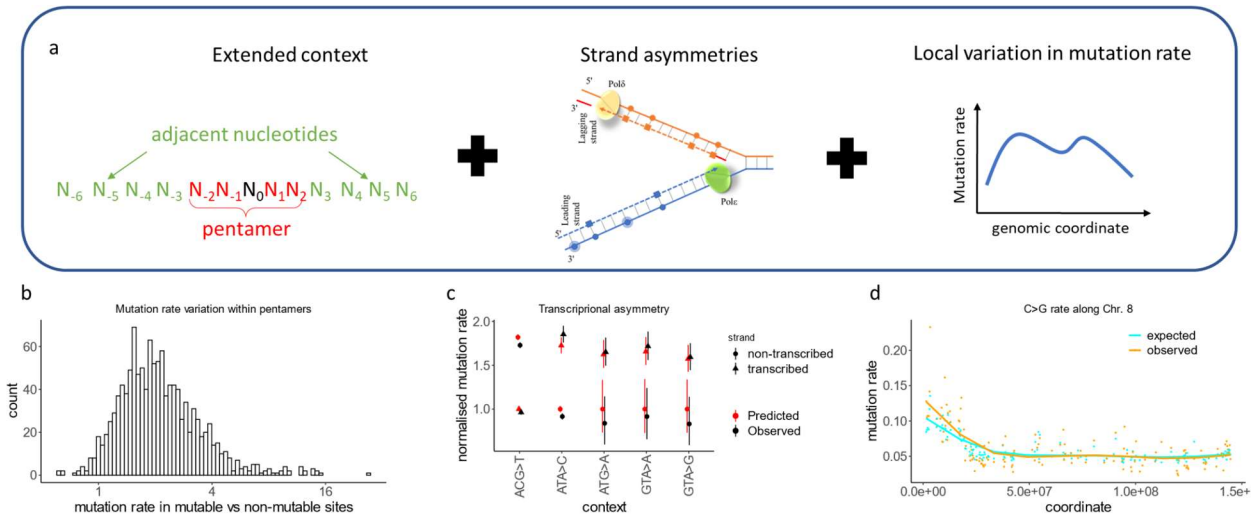

Figure 1. Roulette accounts for extended nucleotide context, strand asymmetries and local variation in mutation rate.

a) Roulette is implemented as logistic regression with pairwise interactions (see Methods). For each pentamer, we model the effect of eight surrounding nucleotides is accounted for (left), strand specific information (middle), and context-specific variation along the genome (right). b) Ratio of observed *de novo* mutation rates between the Roulette predicted most and least mutable deciles for each pentamer shows large variation unexplained by the pentamer context alone. c) Effect of transcriptional asymmetry on the rate of rare synonymous SNVs in the genes with high expression in testis (top quartile). Mutation rate is relative to the least mutable strand. d) Spike of the density of rare synonymous SNVs on the left arm of chromosome 8. This region is known to be affected by increased maternal mutagenesis[4,17,23,24].

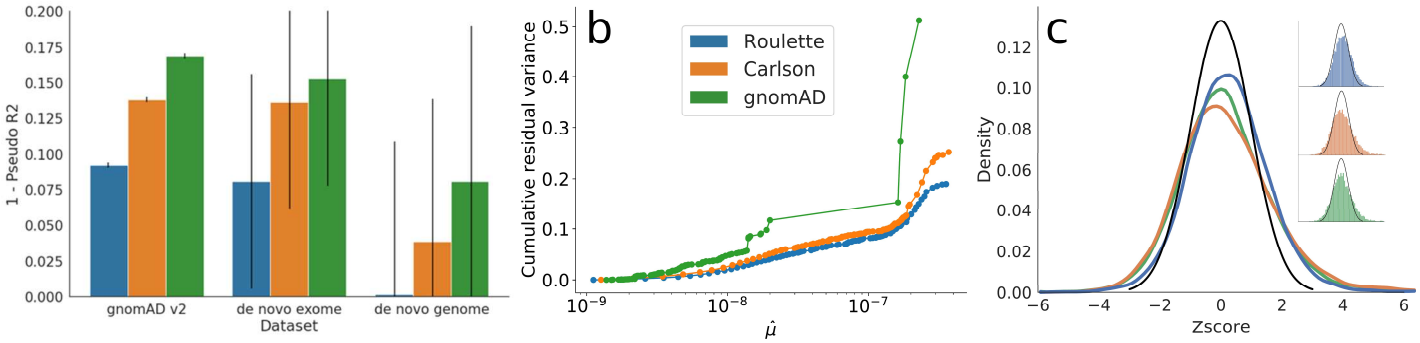

Figure 2 Roulette outperforms existing mutational models, under both per-gene and per-site metrics

a) 1 – Pseudo R^2 of the three mutational models on synonymous variants observed in population sequencing data (gnomAD v2) and two *de novo* mutation datasets[18,27,28]. A pseudo R^2 of 0 is equivalent to using genome-wide mean mutation rate for every site. A pseudo R^2 of 1 is the best per-site mutation rate estimates we can achieve, under the constraint that the mutation rates of synonymous sites follow the predicted genome-wide distribution. b) The estimated cumulative residual variance for the Carlson, gnomAD and Roulette models after binning mutation rate estimates. Within-bin variance is scaled by the total variance estimated for Roulette. c) Error distributions on the Z-scale for predicted counts of synonymous mutations within genes in gnomAD v2. The standard normal density is shown in black to provide a reference for the expected error distribution if mutation rates were known without error.

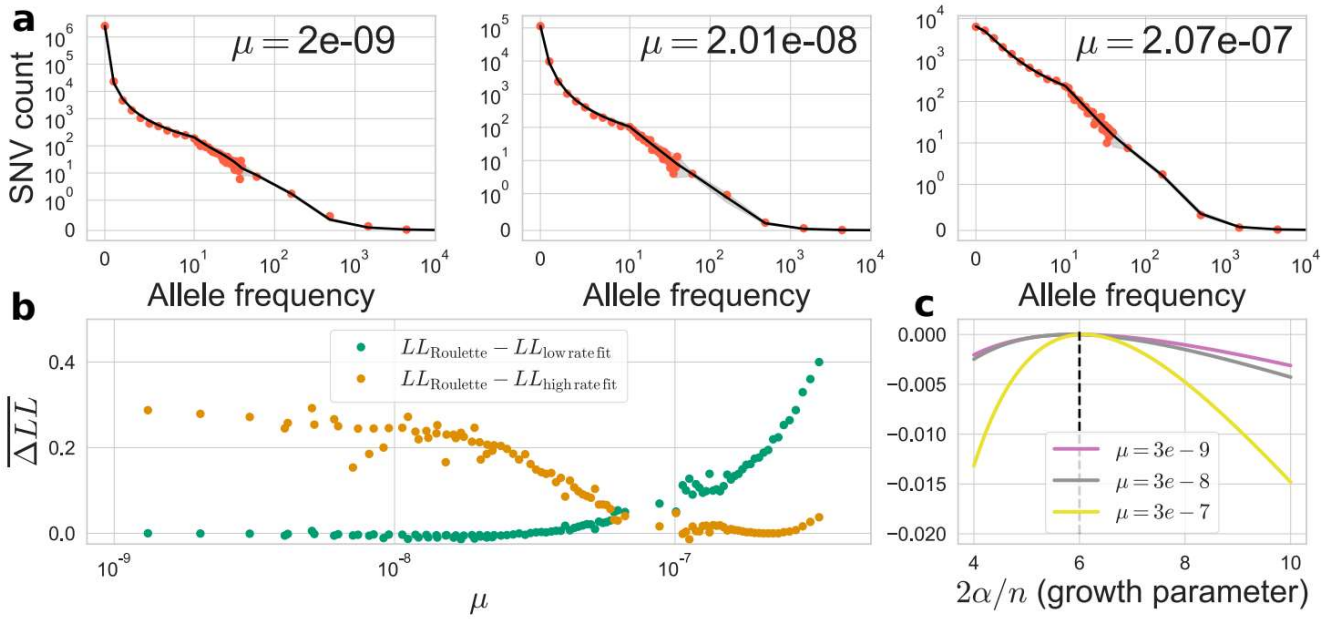

Figure 3 Accurate per-site mutation rate estimates improve population genetics inference
a) Estimated demographic history fits the SFS with mutation rate bins at different orders of magnitude. Red dots show the observed counts at synonymous sites in gnomAD and black lines show expectations of the demographic model with shaded areas giving 95% binomial confidence intervals. Entries 0-40 in the SFS are used as is and binned logarithmically with base 3 above that. b) Roulette bins improve fits to the shape of the SFS compared to demographic model predictions scaled to either low (1e-09 – 3.3e-09) or high rate (1e-07 – 3e-07) bins. Average log-likelihoods (per-SNV) are higher for Roulette after subtracting one to account for the additional parameter used to refit the mutation rate within each bin. Roulette improves over the model trained on sites with low mutation rate (mostly non-recurrent sites) because recurrent mutations change the shape of the SFS It also and improves over the high-rate model as one moves away from the mean mutation rate within the high-rate bin. c) High mutation rate SNVs are more informative about growth parameters. The expected per-SNV log-likelihood relative to the maximum is shown using rare SNVs (1-40 allele counts). The compound growth-rate / sample size parameter was chosen to approximate the observed synonymous SFS in gnomAD v2.

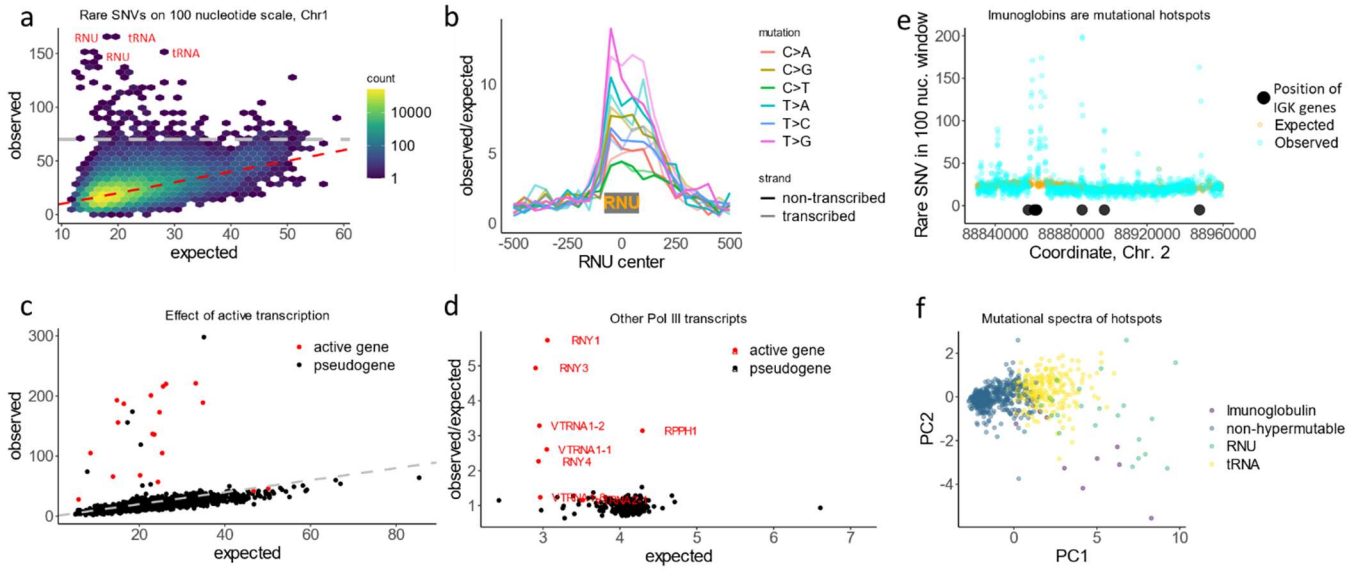

Figure 4. Polymerase III transcripts and transcription binding sites are mutational hotspots

a) Number of rare SNVs in 100 nucleotide non-overlapping windows. Expectation is calculated with Roulette. While mutation counts in most regions show minor deviations from the prediction, a few loci have much higher mutation rates (>70 SNVs, above the grey line). These loci are heavily enriched with Polymerase III transcripts. b) Mutation rate at and around small nuclear RNAs (RNU); median size of RNU depicted as a gray rectangle. c) Number of rare SNVs at active RNUs and pseudogenes. d) Rate of rare SNVs in other classes of Polymerase III transcripts. e) Segment of chromosome 2 harboring immunoglobulin kappa genes. f) Spectra of mutations in regions annotated as immunoglobulins, RNU or tRNA. Some of the annotated regions are not hypermutable (less than 70 SNVs per 100 nucleotide), likely because these regions are tRNA pseudogenes.

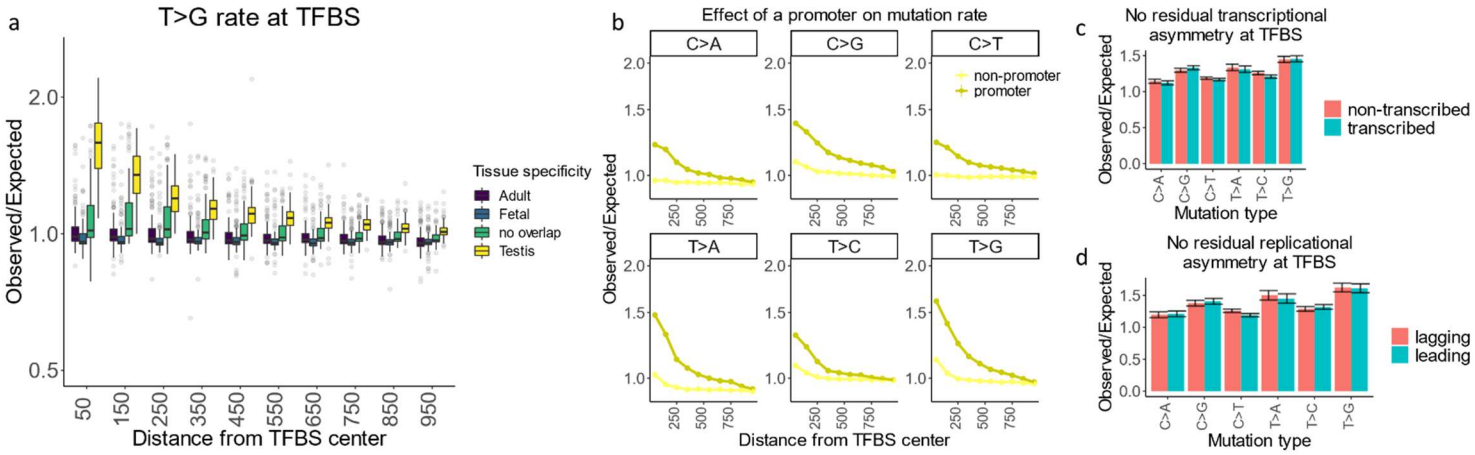

Figure 5. TFBS are prone to high mutation rate.

a) Box plot for the observed to expected rate of rare T>G mutations across different transcription factors. Positions occupied with TF were annotated with chip-seq data. Tissues where TFBSs are active were determined through overlap with tissue specific DHS peaks. b) mutagenic effect of TFBS active in testis overlapping promoter (- 2 kb upstream of transcription start site, dark yellow) or not (light yellow). c) and d) strand resolved observed to expected mutation rates at 100 nucleotide windows around TFBS centers in promoters.