Supplementary Material

Jedidiah Carlson[1], Laura J Scott[2], Adam E Locke[3], Matthew Flickinger[2], Shawn Levy[4], Richard M Myers[4], Michael Boehnke[2], Hyun Min Kang[2], Jun Z Li[1,5*], Sebastian Zöllner[2,6*], Devin Absher[4], Huda Akil[7], Gerome Breen[8], Margit Burmeister[1,5,6,7], Sarah Cohen-Woods[9], William G Iacono[10], James A Knowles[11], Lisa Legrand[10], Qing Lu[12], Matthew McGue[10], Melvin G McInnis[6], Carlos N Pato[13], Michele T Pato[14], Margarita Rivera[8], Janel L Sobell[11], John B Vincent[15], Stanley J Watson[7]

[1]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[2]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[3] McDonnell Genome Institute & Department of Medicine, Washington University, St. Louis, MO, USA

[4]HudsonAlpha Institute for Biotechnology, Huntsville, AL, USA

[5]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

[6]Department of Psychiatry, University of Michigan, Ann Arbor, MI, USA

[7]Molecular & Behavioral Neuroscience Institute, University of Michigan, Ann Arbor, MI, USA

[8]MRC Social Genetic and Developmental Psychiatry Centre, Institute of Psychiatry Psychology and Neuroscience, King's College London, London, UK

[9]School of Psychology, Flinders University, Adelaide, South Australia, AU

[10]Department of Psychology, University of Minnesota, Minneapolis, MN, USA

[11]Department of Psychiatry and the Behavioral Sciences, University of Southern California, Los Angeles, CA, USA

[12]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA

[13]Dean of The College of Medicine, Senior Vice President for Research, SUNY Downstate Medical Center, Brooklyn, NY, USA

[14]Department of Psychiatry, SUNY Downstate Medical Center, Brooklyn, NY, USA

[15]Molecular Neuropsychiatry and Development Laboratory; Campbell Family Mental Health Research Institute, Toronto, ON, CA

**both authors contributed equally*

Table of Contents

Supplementary Note

Supplementary Note 1. Potential Sources of bias

1). Motif-specific error rates

It has been shown that certain sequence motifs may be more susceptible to sequencing error, which could lead to a non-random distribution of false positive singleton calls and subsequently bias our analyses[1,2]. Allhoff et al. (2013)[2] reported context-specific errors for the Illumina HiSeq platform, noting that the most common of these are strand-specific T>X errors at 5'-GGGT-3' motifs (i.e., there is no evidence of an excess of A>X errors at the reverse complement 5'-ACCC-3' motifs). We reason that if the BRIDGES ERVs are enriched for such context-specific errors, we should see significantly more T>X ERVs at the 5'-GGGT-3' motif than A>X ERVs at the 5'-ACCC-3' and motif. Of the 115,531 ERVs that occur at this motif, 57,699 were 5'-[A>X]CCC-3' variants, and 57,832 were 5'-GGG[T>X]-3' variants; this difference was not significant, indicating there is no evidence for an enrichment of T>X ERVs at this error-prone motif (exact binomial test; P=0.70). Allhoff et al. (2013) remark that the variants called at error-prone positions tended to have low base quality scores as well as significant strand bias, both of which are detectable with standard filtering protocols[2]. We therefore assume that most motif-specific errors are efficiently filtered by the default strand-bias and quality filters used in our variant calling pipeline, and any undetected errors have a negligible impact on our calculation of relative mutation rates and downstream analyses.

2). Mapping error

We also considered the possibility that ERVs occurring on poorly mapped reads might bias our analysis of regional variation in mutation rates. We expect the majority of ERVs in our data are mapped with high confidence, as the pre-filtering steps in our variant calling pipeline remove sites with average phred-scaled mapping quality score (MQ) <20 and/or with more than 10% of reads that are ambiguously mapped (MQ0>10). This filtering strategy is similar to the filters employed by other large-scale sequencing projects that have demonstrated well-controlled error rates among singleton calls[3,4]. While a more aggressive mapping quality filter would reduce concerns about region-specific error

biases, doing so would primarily filter out ERVs occurring in repeat-rich pericentromeric regions[5] thus precluding our ability to assess the mutation spectrum in these regions. Prior research has found that centromeric and pericentromeric regions evolve more rapidly than elsewhere in the genome[5–7], which is an intriguing phenomenon that would be entirely undetectable if we omit these regions from our analyses.

Supplementary Note 2. Comparison of 7-mer relative mutation rates with independent estimates

Aggarwala & Voight (2016)[8] estimated "substitution rates" using 7,051,667 intergenic variants observed in N=379 Europeans from the 1000 Genomes Phase I study. These substitution rates are analogous to the relative mutation rates used in our study, but are derived from variants across the entire frequency spectrum, encompassing both singletons and common variants. The exact site frequency spectrum for the European intergenic variants is not reported, but Aggarwala & Voight (2016)[8] specify 26% of variants in the 1000G Phase I African sample are singletons or doubletons. Because the BRIDGES sample is ~10 times larger than the 1000G Phase I European sample, we expect many of the 1000G Phase I European singletons are present in the BRIDGES data in multiple individuals (i.e., non-singletons), and hence ancestrally older. The rates estimated by Aggarwala & Voight (2016)[8] are therefore expected to be more similar to the BRIDGES MAC10+-derived relative mutation rates than they are to the BRIDGES ERV-derived rates. As shown in **Supplementary Fig. 3a**, the BRIDGES MAC10+-derived rates are more strongly correlated with rates estimated by 1000 Genomes intergenic variants (r=0.995) than with BRIDGES ERVs (r=0.991). Type-specific correlations between MAC10+-derived and 1000G-derived rates are also higher for all types except A>G and non-CpG C>T transitions (**Supplementary Fig. 3b**). Only 129 of the 24,576 7-mer subtypes (0.5%) have more than a 2-fold difference between MAC10+-derived and 1000G-derived rates (**Supplementary Fig. 3c**), compared to 741 (3%) of 7-mer subtypes with >2-fold difference between MAC10+-derived and ERV-derived rates. The rates estimated by Aggarwala & Voight (2016)[8] constitute a benchmark by which we compare our models' ability to predict true *de novo* mutations; we show that our analogous model based on the

BRIDGES MAC10+-derived rates performs similarly to these previously published rates, and the models based on BRIDGES ERV-derived rates consistently predict *de novo* mutations with greater accuracy (**Supplementary Fig. 6**).

Supplementary Note 3. Tests for enrichment/depletion of *de novo* mutations in feature-associated subtypes.

Our multivariate models identify specific 7-mer subtypes found to be enriched (or depleted) for ERVs when occurring in the presence of a genomic feature. While our model validation results demonstrate that accounting for these features in aggregate improves prediction of *de novo* mutations, it does not show that, for a given single feature, these subtype-specific effects could also be detected among actual *de novo* mutations. Because the available catalogs of *de novo* mutations are relatively sparse, validating each individual feature-associated 7-mer subtype is not feasible. Instead, we looked across all 7-mer subtypes associated with a given feature in the same direction, and tested if the *de novo* mutations of those subtypes were higher (or lower) than expected under the null assumption that a feature has no effect on those subtypes' mutability.

Hence, for each feature, we identified regions of the genome covered by that feature, and calculated the expected number of ERVs in those regions based on the 7-mer relative mutation rates (i.e., assuming the feature has no effect on mutability) of subtypes significantly associated in the same direction with the feature. Assuming no systematic bias, this number is proportional to the expected number of *de novo* mutations (e.g., if we expect 36,000 BRIDGES ERVs in those regions [0.1% of all ERVs], we would expect ~47 [0.1%] of the GoNL[9]/Inova[10] *de novo* mutations occur in the same regions). We compared the expected number to the observed number of *de novo* mutations using one-sided Pearson's Chi-squared tests, each with 1 degree of freedom (prop.test() function in R). A significant result indicates that observed counts of *de novo* mutations in the feature vary as predicted (higher or lower) from the expected count. Ten of the 15 tests showed a significant enrichment or

depletion of observed *de novo* mutations (**Supplementary Table 6a**). These results are not solely a result of feature-associated DNA methylation, as the associations remained significant when subtypes with CpG dinucleotides were excluded (**Supplementary Table 6b**). Note that four of the non-significant tests described in **Supplementary Table 6a** where we predicted an increase in *de novo* mutations had fewer observed *de novo* mutations than expected (CpG islands, GC content, H3K27me3, and lamin-associated domains). This may indicate false positive sin our model, but is also consistent withThis may a limited ability to confidently call *de novo* mutations in the GoNL/Inova datasets due to low coverage in GC-rich regions[9,10]. We conclude that most of the mutagenic effects of genomic features inferred by our model are likely operative in the germline and play a role in shaping mutation rate heterogeneity across the genome.

Supplementary Note 4. Potential mechanisms for TTAAAA hypermutability

Our finding of a 3-fold depletion of TTAAAA AT>TA motifs in DNase hypersensitive sites provides an excellent example of how our results can be leveraged to better understand the origins of certain mutation patterns. We identify two possible mechanisms that might explain the context-dependent mutation probabilities of AT>TA mutations at TTAAAA hexamers. As described in the main text, L1 EN nicking activity has been shown to vary according to the nucleosomal context of its target motifs, usually occurring at a higher rate in nucleosome-free DNA, but in some cases actually decreasing in nucleosome-free DNA[11]. Therefore, under the L1 EN model, it is possible to see either a positive or negative association between TTAAAA mutability and DHS.

Slipped-strand mispairing, also known as replication slippage, is another plausible hypothesis for the hypermutability of this motif[8]. Because the nucleosomal architecture is disrupted ahead of the replication fork[12], and reassembled almost immediately thereafter[13], nascent DNA containing unresolved lesions that is packaged in nucleosomes could be inaccessible to mismatch repair machinery, thus preserving any errors caused by slippage. In this case, it is also possible to see a negative association between TTAAAA mutability and DHS.

This slippage mechanism, however, appears to be unlikely for the following reasons. First, replication slippage inherently results in short insertions or deletions rather than point mutations. Mapping error could potentially cause an insertion/deletion to be falsely identified as a single-nucleotide variant, but such errors would need to be extremely prevalent in our data (and also context-dependent) in order to observe a 3-fold depletion of these singletons in DHS. Given the quality metrics we report for the BRIDGES singletons, it seems unlikely that these results are purely a technical artifact. Furthermore, if slippage were the primary mechanism, we would expect other motifs ending in poly-A 4-mers to also show an inverse association with DHS. Among the 13 NNNAAAA subtypes whose mutability is significantly associated with DHS, only five are inversely associated, three of which are NNTAAAA motifs (i.e., conforming closely to the canonical target for L1 EN nicking activity). The other eight subtypes all show *higher* mutation rates in DHS, which conflicts with the proposed slippage+chromatinization mechanism.

Supplementary Note 5. Derivation of false discovery rate by Ts/Tv statistics

(1) Let $TS_o = TS_{tp} + TS_{fp}$ be the number of observed transitions, consisting of both true positives (TS_{tp}), and false positives (TS_{fp})

(2) Let $TV_o = TV_{tp} + TV_{fp}$ be the number of observed transversions.

(3) Based on findings from other large-scale sequencing studies, the true positive Ts/Tv ratio, $TSTV_T = \frac{TS_{tp}}{TV_{tp}}$ is between 2.0 and 2.1[14].

(4) Because there are 8 possible transversions and 4 possible transitions, if errors are occurring at random, the Ts/Tv ratio for random false positive errors ($TSTV_\epsilon$) should be 0.5, that is, $\frac{TS_{fp}}{TV_{fp}} = 0.5$.

Solving this system of four equations, it follows that $TV_{fp} = \frac{TSTV_T \times TV_o - TS_o}{TSTV_T - 0.5}$ and $TS_{fp} = 0.5 \times TV_{fp}$, so the false discovery rate is estimated as:

$$\frac{TS_{fp} + TV_{fp}}{TS_o + TV_o} = \frac{0.5\left(\frac{TSTV_T \times TV_o - TS_o}{TSTV_T - 0.5}\right) + \frac{TSTV_T \times TV_o - TS_o}{TSTV_T - 0.5}}{TS_o + TV_o}$$

Assuming a true $TSTV_T$ between 2.0 and 2.1, we estimate a false discovery rate of 0.6-2.6% among the BRIDGES ERVs.

Acknowledgements

The BRIDGES study was supported by R01 MH094145 to Michael Boehnke and Richard M. Myers and U01 MH105653 to Michael Boehnke. The collection and storage of cases and controls from the Centre for Addiction and Mental Health (CAMH) in Toronto and from the Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London in London, U.K. was supported by funding from GlaxoSmithKline, from the Canadian Institutes of Health Research to John B. Vincent, MOP-172013 (CAMH), and funding from the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London (IoPPN). The views expressed are those of the author(s) and not necessarily those of the UK NHS, the NIHR or the UK Department of Health. Case and control collection was supported by Heinz C. Prechter Bipolar Research Fund at the University of Michigan Depression Center to Melvin G. McInnis (Prechter). Data and biomaterials were collected for the Systematic Treatment Enhancement Program for Bipolar Disorder (STEP-BD), a multi-center, longitudinal project selected from responses to RFP #NIMH-98-DS-0001, "Treatment for Bipolar Disorder" which was led by Gary Sachs and coordinated by Massachusetts General Hospital in Boston, MA with support from 2N01 MH080001-001. The Genomic Psychiatric Cohort wishes to acknowledge all of the research participants in this cohort; the study was supported by U01 MH105641, R01 MH085548, R01MH104964. The MCTFR study was supported through grants from the National Institutes of Health DA037904, DA024417, DA036216, DA05147, AA09367, DA024417, HG007022, and HL117626.

References

1. Minoche, A., Dohm, J. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* **12,** R112 (2011).

2. Allhoff, M. *et al.* Discovering motifs that induce sequencing errors. *BMC Bioinformatics* **14 Suppl 5,** S1 (2013).

3. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

4. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).

5. Horvath, J. E. *et al.* Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. *Hum. Mol. Genet.* **9,** 113–23 (2000).

6. Dover, G. A. Evolution of genetic redundancy for advanced players. *Curr. Opin. Genet. Dev.* **3,** 902–910 (1993).

7. Bensasson, D. Evidence for a high mutation rate at rapidly evolving yeast centromeres. *BMC Evol. Biol.* **11,** 211 (2011).

8. Aggarwala, V. & Voight, B. F. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nat. Genet.* **48,** 349–55 (2016).

9. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).

10. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48,** 935–939 (2016).

11. Cost, G. J., Golding, A., Schlissel, M. S. & Boeke, J. D. Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res.* **29,** 573–577 (2001).

12. Groth, A., Rocha, W., Verreault, A. & Almouzni, G. Chromatin Challenges during DNA Replication and Repair. *Cell* **128,** 721–733 (2007).

13. Leman, A. R. & Noguchi, E. The replication fork: understanding the eukaryotic replication machinery and the challenges to genome duplication. *Genes (Basel).* **4,** 1–32 (2013).

14. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

15. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

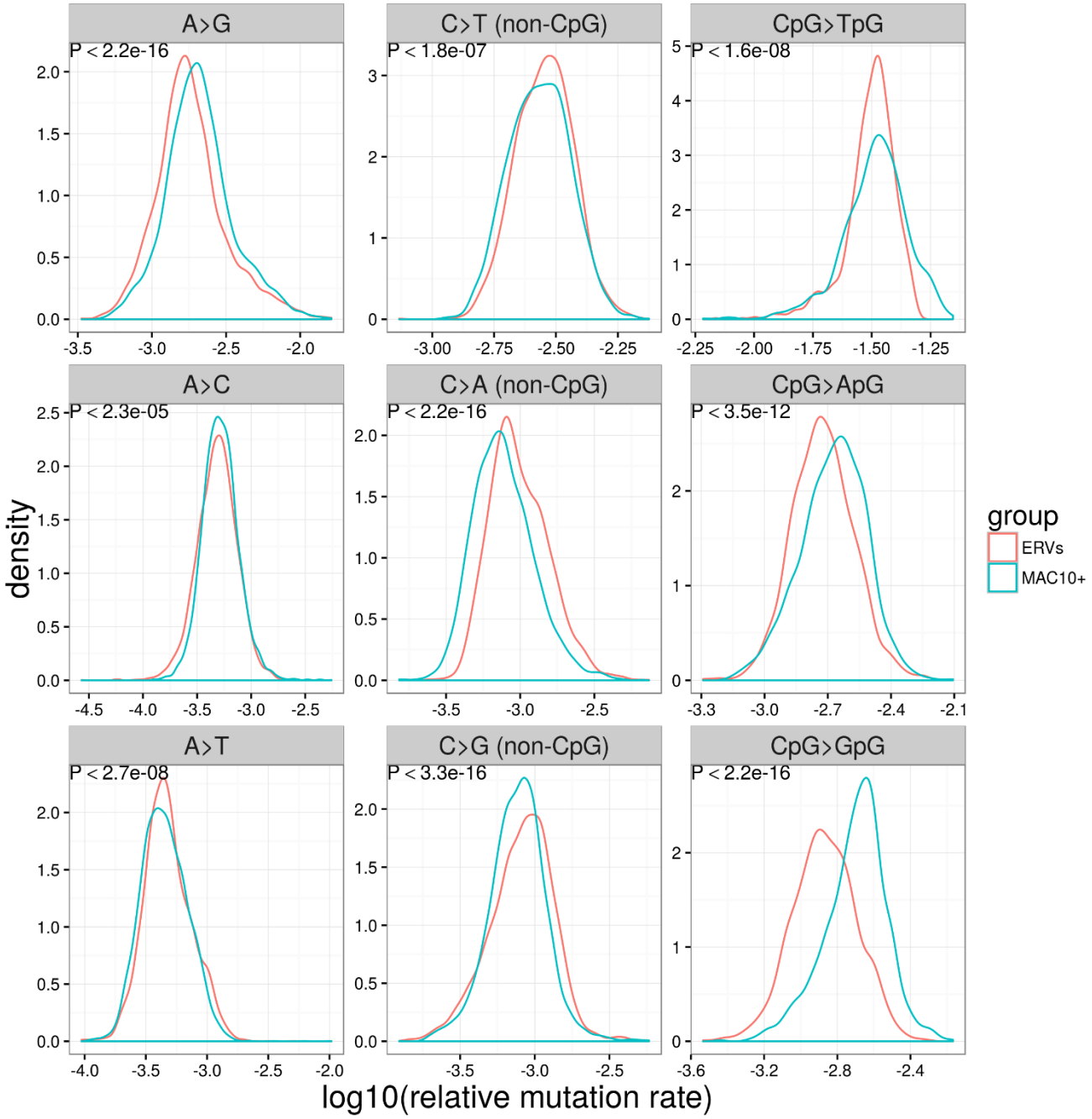
16. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91,** 1033–1040 (2012).

17. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103 (2010).

18. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–51 (2008).

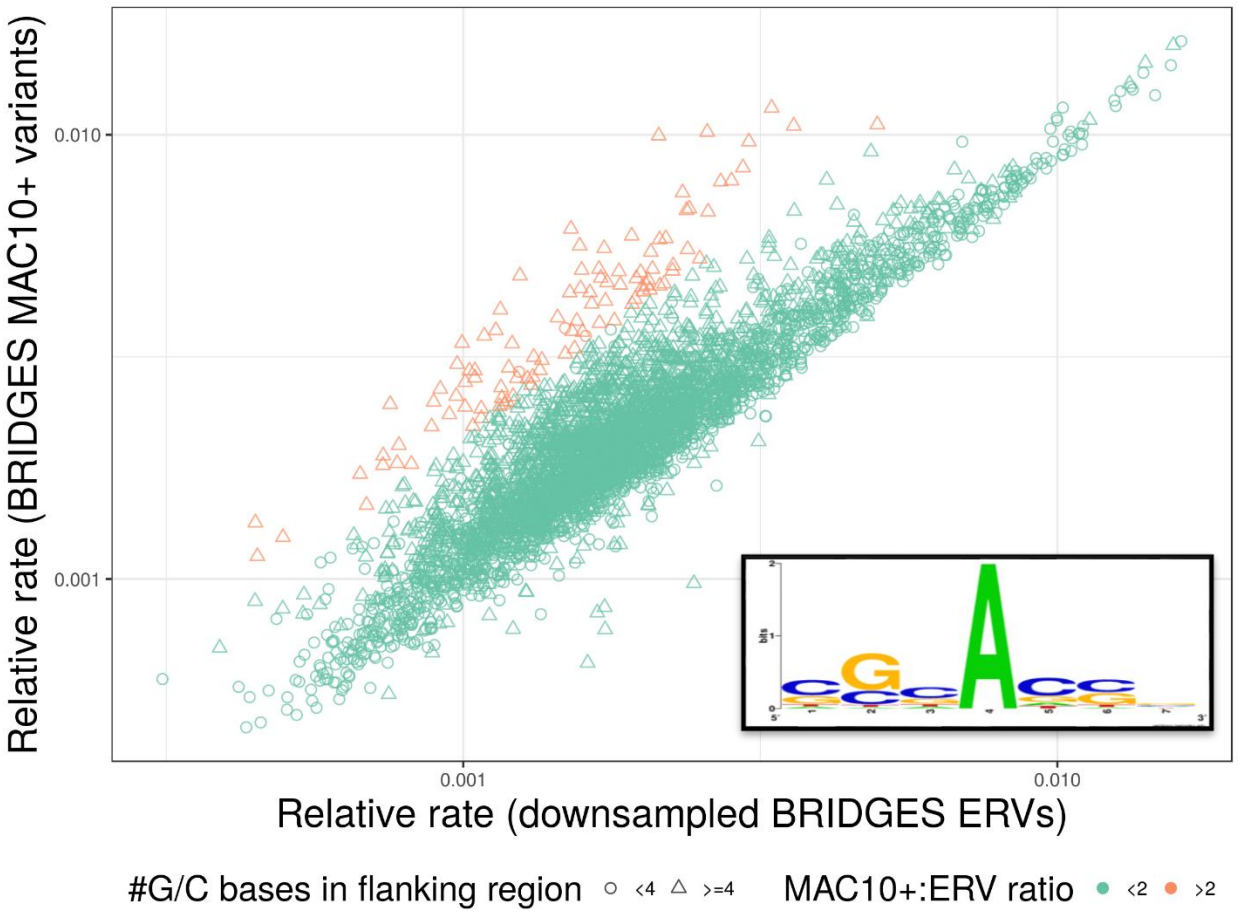
19. Wu, H., Caffo, B., Jaffee, H. A., Irizarry, R. A. & Feinberg, A. P. Redefining CpG islands using hidden Markov models. *Biostatistics* **11,** 499–514 (2010).

Supplementary Figures

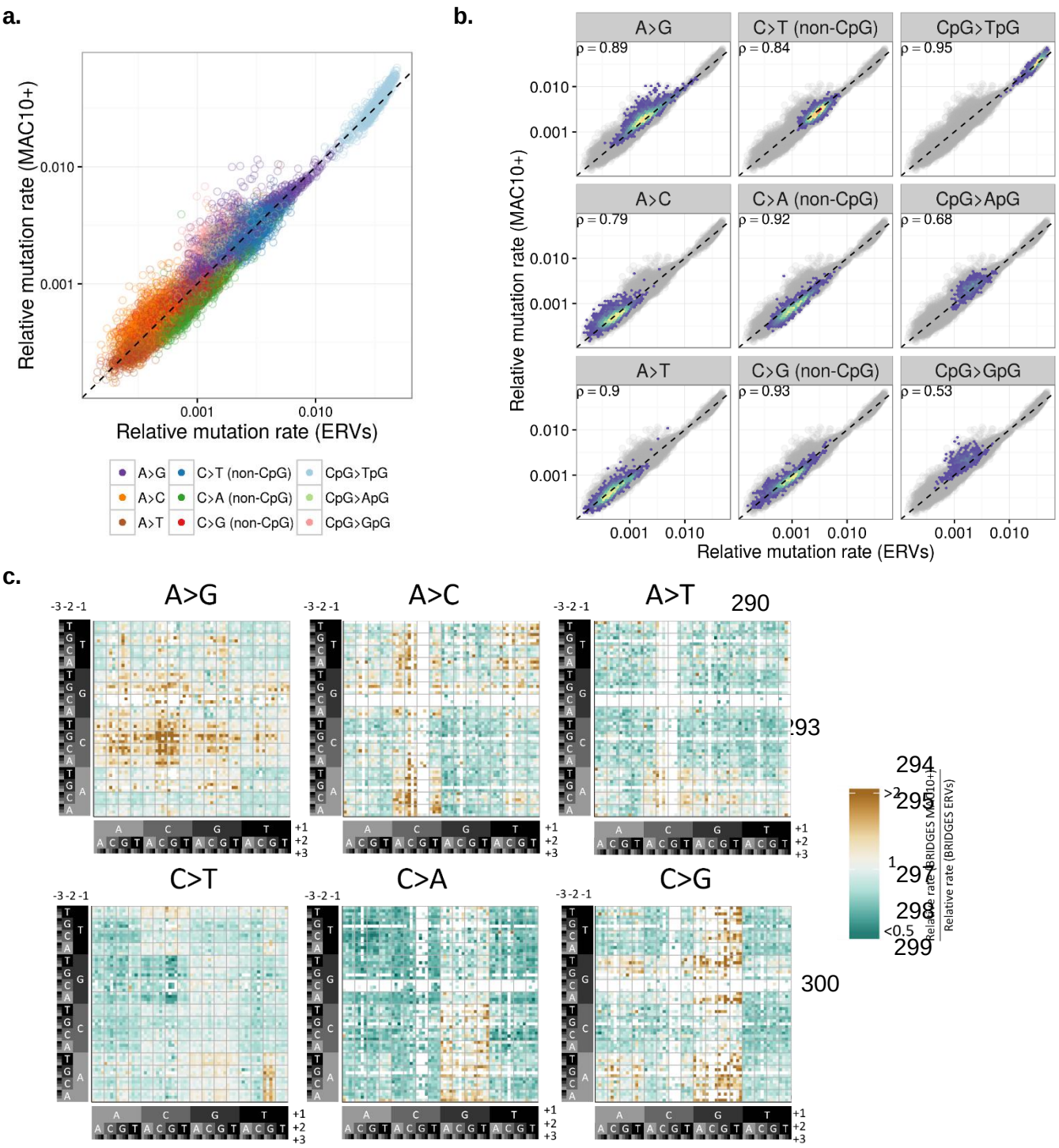


Supplementary Figure 1

Densities of $\log_{10}$-scaled 7-mer relative mutation rates, estimated using the downsampled BRIDGES ERVs (red) and BRIDGES MAC10+ variants (blue). P-values from the Kolmogorov-Smirnov test for distributional equivalence are shown in the upper left corner of each panel.

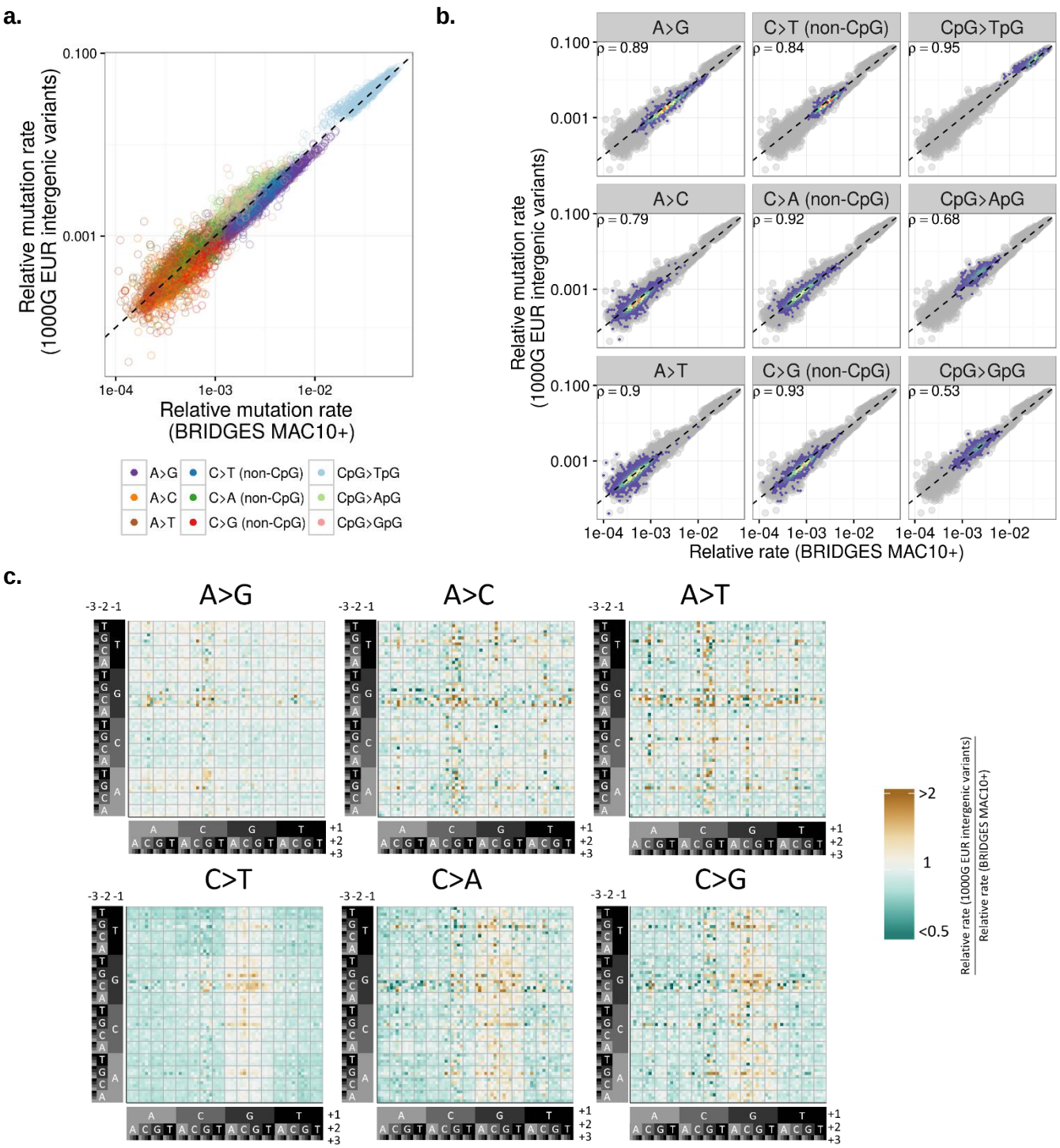


Supplementary Figure 2 Detailed comparison between ERV-derived and MAC10+-derived A>G transition rates. Points are colored by the ratio between the two rates for that subtype (orange: MAC10+:ERV>2; green: MAC10+:ERV<2). The shape of each point indicates the number of G or C bases in the +/-3 nucleotides flanking the variant site. Among the 103 7-mer motifs with a MAC10+:ERV ratio >2, 100 have 4 or more G/C bases in the flanking region. **(inset)** Sequence logo for these 103 7-mer subtypes with MAC10+:ERV ratio >2 shows flanking regions are enriched for G/C bases.



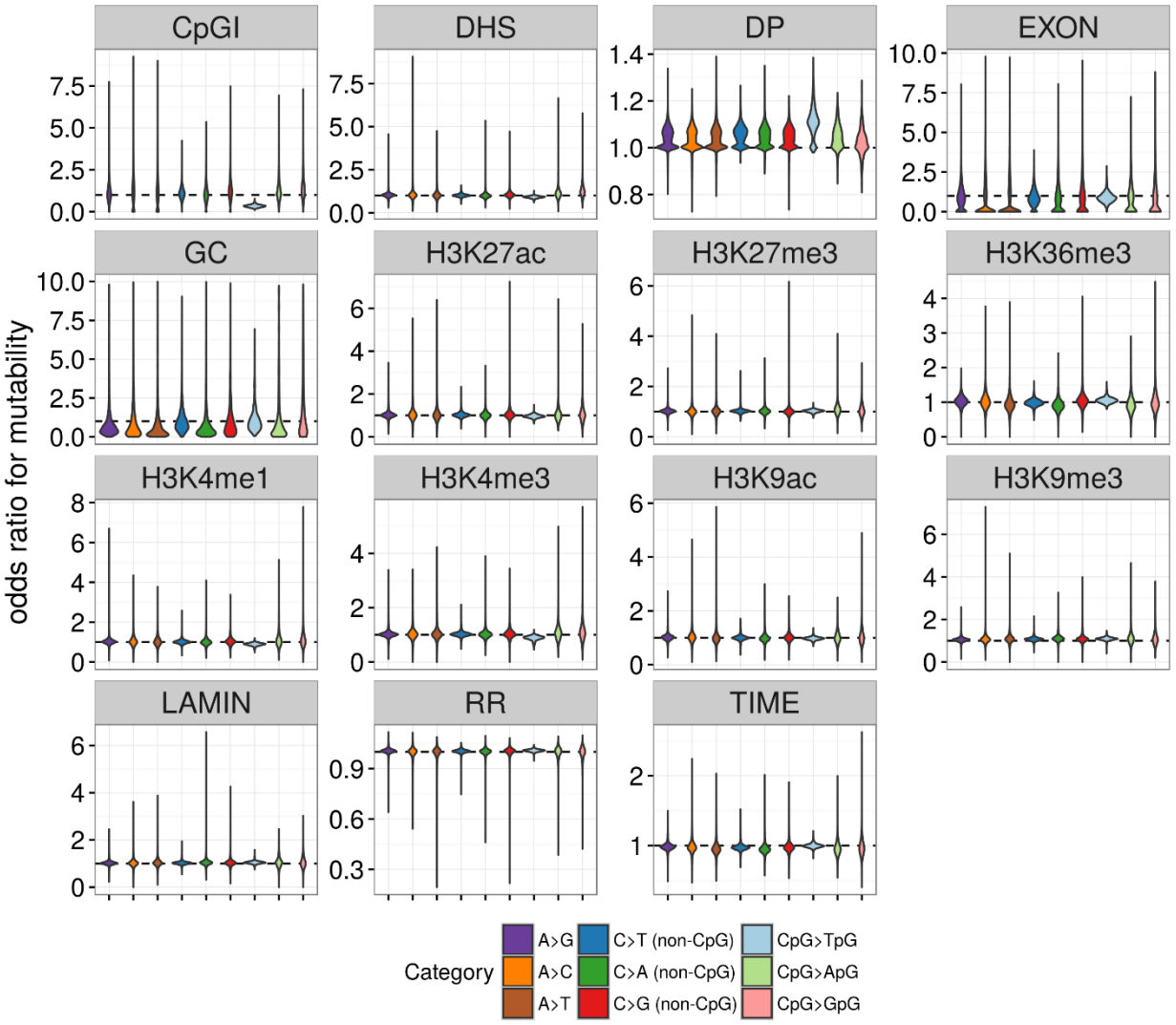
Supplementary Figure 3

(a) Relationship between 7-mer relative mutation rates estimated using down-sampled BRIDGES ERVs (x-axis) and variants with a minor allele count >= 10 (MAC10+; y-axis), excluding subtypes with <50 variants in either dataset. **(b)** Type-specific 2D-density plots, as situated in the scatterplot of **a**. The dashed line indicates an expected least-squares regression line if there is no bias present. **(c)** Heatmap shows ratio between relative mutation rates calculated on MAC10+ variants and ERVs for each 7-mer mutation subtype. Subtypes with higher MAC10+-derived rates relative to ERV-derived rates are shaded gold, and subtypes with lower MAC10+-derived rates relative to ERV-derived rates are shaded green.



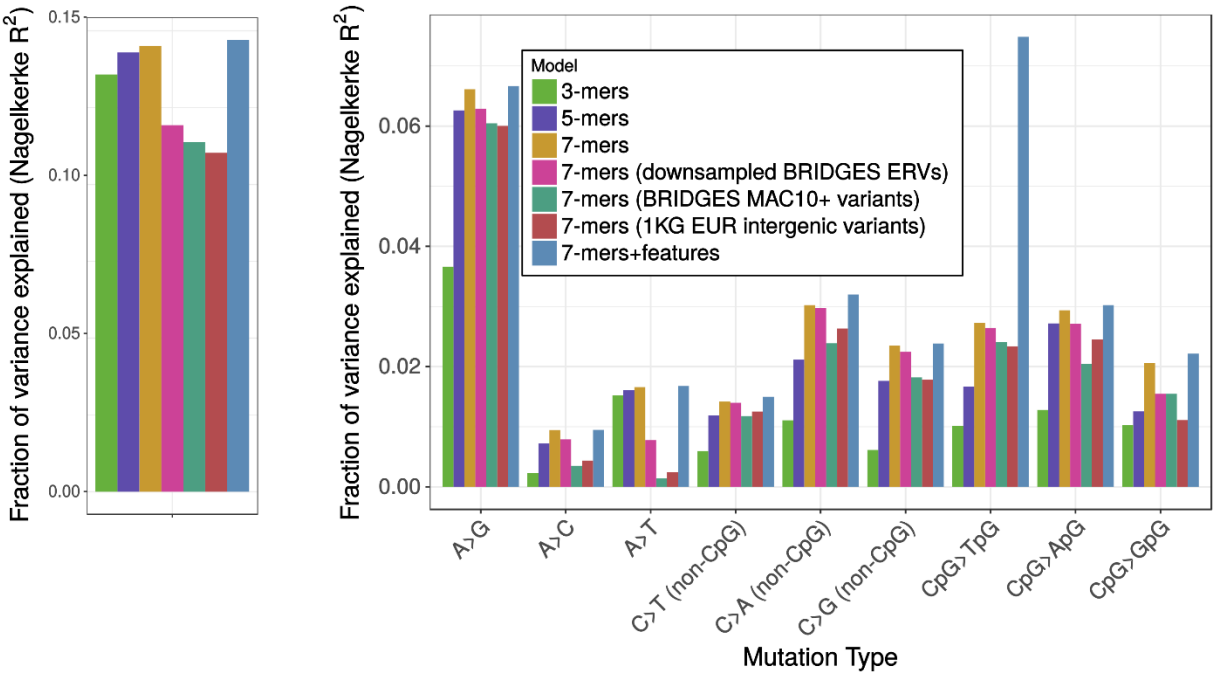
Supplementary Figure 4

(a) Relationship between 7-mer relative mutation rates estimated using BRIDGES variants with a minor allele count >= 10 (MAC10+; x-axis), and 7-mer rates calculated from intergenic variants in the European 1000G phase I sample (y-axis) **(b)** Type-specific 2D-density plots, as situated in the scatterplot of **a**. The dashed line indicates an expected least-squares regression line if there is no bias present. **(c)** Heatmap shows ratio between relative mutation rates calculated on MAC10+ variants and 1000G variants for each 7-mer mutation subtype. Subtypes with higher 1000G-derived rates relative to MAC10+-derived rates are shaded gold, and subtypes with lower 1000G-derived rates relative to MAC10+-derived rates are shaded green. 1000G-derived rates shown here are scaled relative to the MAC10+-derived rates.



Supplementary Figure 5 Distributions of effect sizes (including non-significant effects) on mutability for the 14 genomic features considered in the logistic regression model. For each feature, we plotted the empirical distributions of these subtype-specific odds ratios for each basic mutation type. *Replication timing is coded with negative values indicating later replicating regions, so an OR<1 means mutation rate increases in late-replicating regions. Note that effects in CpG islands are shown on a wider scale than other features.

a. **b.**



Supplementary Figure 6 Comparison of variance explained by all models for **(a)** all mutation types combined, and **(b)** stratified by mutation type.

Supplementary Tables

Supplementary Table 1 Quality comparison between filtered partitions of BRIDGES singletons

Partition	# Singletons	Ts/Tv ratio	%dbSNP (b142)	% of Full Set
Full Set	36,087,319	2.02	17.4	100
Filter 1 (QUAL>=30)*	20,796,900	2.03	17.8	58
Filter 2 (MQ>56)	33,550,098	2.01	17.3	93
Filter 3 (passed 1000G strict mask)	28,958,837	1.94	17.5	80
All Filters (MQ>56, QUAL>=30, 1000G strict mask)	16,535,856	2.00	17.6	46

**Quality score cutoff uses raw base quality scores obtained prior to recalibration.*

Supplementary Tables 2a-2d Relative mutation rate estimates for 1-mers, 3-mers, 5-mers, and 7-mers

[see separate spreadsheet, table_S2_K-mer_relative_rates.xlsx]

Each table contains data used to calculate relative mutation rates for K-mers of a given length. Each row in the table contains the following columns: 1) basic mutation type; 2) K-mer motif corresponding to a reference base A or C at the central mutated position (the reverse complement of each motif, corresponding to reference base T or G is given in parentheses); 3) number of singletons observed in the BRIDGES data of the K-mer subtype defined by columns 1 and 2; 4) total number of times the motif in column 2 is observed in the reference genome; 5) relative mutation rate of singletons in that subtype (column 3 divided by column 4). For 7-mer subtypes (Supplementary Table 2d), we include four additional columns: 6) number of singletons in that subtype, after downsampling to 12M; 7) relative mutation rate of downsampled singletons in that subtype (column 6 divided by column 4); 8) number of MAC10+ variants observed in the BRIDGES data of that subtype; 9) relative mutation rate of polymorphisms of that subtype (column 8 divided by column 4).

Supplementary Table 3a Summary of overall model fit statistics for *de novo* testing data

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
1-mers	0.082 (0.082)	326076 (326076)	--
3-mers	0.136 (0.111)	309619 (317089)	<2.2e-308
5-mers	0.143 (0.117)	307405 (315331)	<2.2e-308
7-mers	0.145 (0.119)	306738 (314705)	1.56e-148
7-mers+features	*0.147* (0.119)	*306146* (314943)	3.08e-147
7-mers (downsampled BRIDGES ERVs)	0.119	314723	--
7-mers (BRIDGES MAC10+ variants)	0.114	316400	--
7-mers (intergenic 1000G polymorphisms)[8]	0.110	317490	--

Due to the nested structure of the first 5 models in this table (described in **Materials and Methods**), Nagelkerke's R^2 is slightly biased upwards for models with more parameters. For a more direct comparison with the other 3 models, we repeated each these models with only 1 composite predictor (as was done for the downsampled ERV, MAC10+, and 1KG polymorphism models), and we include Nagelkerke's R^2 and AIC values for these models in parentheses. Note that the relative differences in Nagelkerke's R^2 between non-nested K-mer and (K+2)-mer models are nearly identical to what we observe in the nested modeling framework. Also, because all models are applied to the same testing data, AIC is a valid means of comparison between all models, regardless of number of predictors; the nested 7-mer+features model achieves the lowest AIC, indicating this model provides the best overall fit. The last column of P-values come from likelihood ratio test between each nested model and the corresponding model in the preceding row, where such nested models exist.

Supplementary Table 3b Summary of type-specific model fit statistics for *de novo* testing data. Each type is shown in a sub-table, with the number of *de novo* mutations and non-mutated sites used in the partitioned testing data indicated in the subheading.

A>C (2920 *de novo* mutations; 198481 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.0023	30447	--
5-mers	0.0072	30309	4.20e-32
7-mers	0.0094	30248	1.94e-15
7-mers+features	0.0095	30249	0.385
7-mers (downsampled BRIDGES ERVs)	0.0079	30288	--
7-mers (BRIDGES MAC10+ variants)	0.0035	30413	--
7-mers (intergenic 1000G polymorphisms)[8]	0.0043	30388	--

A>G (11400 *de novo* mutations; 198793 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.037	85999	--
5-mers	0.063	84087	<2.2e-308
7-mers	0.066	83829	2.22e-58
7-mers+features	0.067	83792	3.68e-10
7-mers (downsampled BRIDGES ERVs)	0.063	84065	--
7-mers (BRIDGES MAC10+ variants)	0.060	84244	--
7-mers (intergenic 1000G polymorphisms)[8]	0.060	84278	--

A>T (2455 *de novo* mutations; 198320 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.015	26123	--
5-mers	0.016	26103	3.27e-06
7-mers	0.017	26092	3.16e-04
7-mers+features	0.017	26090	0.038
7-mers (downsampled BRIDGES ERVs)	0.008	26307	--
7-mers (BRIDGES MAC10+ variants)	0.001	26466	--
7-mers (intergenic 1000G polymorphisms)[8]	0.002	26440	--

non-CpG C>A (3620 *de novo* mutations; 128765 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.011	32901	--
5-mers	0.021	32606	9.73e-67
7-mers	0.030	32340	3.75e-60
7-mers+features	0.032	32290	4.60e-13
7-mers (downsampled BRIDGES ERVs)	0.030	32351	--
7-mers (BRIDGES MAC10+ variants)	0.024	32523	--
7-mers (intergenic 1000G polymorphisms)[8]	0.026	32451	--

non-CpG C>G (3561 *de novo* mutations; 128746 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.006	32603	--
5-mers	0.018	32271	1.25e-75
7-mers	0.023	32102	4.79e-39
7-mers+features	0.024	32093	1.10e-03
7-mers (downsampled BRIDGES ERVs)	0.022	32127	--
7-mers (BRIDGES MAC10+ variants)	0.018	32251	--
7-mers (intergenic 1000G polymorphisms)[8]	0.018	32263	--

non-CpG C>T (10321 *de novo* mutations; 128774 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.006	73240	--
5-mers	0.012	72902	5.40e-76
7-mers	0.014	72771	9.49e-31
7-mers+features	0.015	72728	1.92e-11
7-mers (downsampled BRIDGES ERVs)	0.014	72779	--
7-mers (BRIDGES MAC10+ variants)	0.012	72905	--
7-mers (intergenic 1000G polymorphisms)[8]	0.013	72863	--

CpG>ApG (304 *de novo* mutations; 6108 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.013	2424	--
5-mers	0.027	2397	5.82e-08
7-mers	0.029	2394	3.43e-02
7-mers+features	0.030	2395	0.18
7-mers (downsampled BRIDGES ERVs)	0.027	2395	--
7-mers (BRIDGES MAC10+ variants)	0.020	2408	--
7-mers (intergenic 1000G polymorphisms)[8]	0.024	2400	--

CpG>GpG (270 *de novo* mutations; 6292 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.010	2218	--
5-mers	0.013	2216	0.037
7-mers	0.021	2203	9.90e-05
7-mers+features	0.022	2202	0.083
7-mers (downsampled BRIDGES ERVs)	0.015	2208	--
7-mers (BRIDGES MAC10+ variants)	0.015	2208	--
7-mers (intergenic 1000G polymorphisms)[8]	0.011	2217	--

CpG>TpG (6960 *de novo* mutations; 6289 non-mutated sites)

Model	Nagelkerke's R^2	AIC	P-value (likelihood ratio test)
3-mers	0.010	18067	--
5-mers	0.017	18005	8.36e-16
7-mers	0.027	17900	7.45e-25
7-mers+features	0.075	17415	7.23e-108
7-mers (downsampled BRIDGES ERVs)	0.026	17905	--
7-mers (BRIDGES MAC10+ variants)	0.024	17928	--
7-mers (intergenic 1000G polymorphisms)[8]	0.024	17935	--

Supplementary Table 4 **t-tests for differences in mean MAC10+/ERV ratio of GC-poor vs. GC-rich 7-mer motifs**

Type	Mean MAC10+/ERV ratio (≤3 C/G bases)	Mean MAC10+/ERV ratio (≥4 C/G bases)	P-value
A>C	0.97	1.12	8.00e-30
A>G	1.00	1.28	2.37e-161
A>T	0.89	0.89	0.81
C>A (non-CpG)	0.76	0.72	2.61e-09
C>G (non-CpG)	0.89	0.93	2.98e-04
C>T (non-CpG)	0.93	0.85	1.75e-39
CpG>ApG	1.15	0.96	4.97e-22
CpG>GpG	1.46	1.33	2.80e-04
CpG>TpG	1.02	0.98	1.01e-09

For each mutation subtype, we calculated the ratio between MAC10+-derived and ERV-derived relative mutation rates. Then, for each of the 9 basic types, we grouped 7-mer subtypes into low C/G subtypes (≤3 C/G bases in the +/-3 flanking positions) and high C/G subtypes (≥4 C/G bases in the +/-3 flanking positions) and performed t-tests for differences in the mean MAC10+/ERV ratios of these two groups.

Supplementary Table 5 Genomic features used in mutation models

Feature	Source	Cell Type	Resolution
H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3	Roadmap Epigenomics Project[15]	Peripheral Blood Mononuclear Primary Cells	1bp (inside vs. outside of broad peak)
Replication timing	Koren et al., 2012[16]	Lymphoblastoid	1kb window
Recombination rate	Kong et al., 2010[17] (deCODE sex-averaged recombination rate map)	--	10kb window
Lamin B1 domains	Guelen et al., 2008[18]	Tig3ET normal human embryonic lung fibroblasts	1bp (inside vs. outside of LAD)
DNase hypersensitivity sites	ENCODE	multiple	1bp (inside vs. outside of DHS region)
Exonic site	RefSeq gene database	--	1bp (inside vs. outside of exon)
CpG island	Wu et al., 2010[19]	--	1bp (inside vs. outside of CpG island)
% GC content	Calculated from reference genome	--	10kb

A script to download the exact external data files used in this paper is available at https://github.com/carjed/smaug-genetics

Supplementary Table 6a Univariate tests for enrichment or depletion of *de novo* mutations occurring in feature-associated subtypes identified by logistic regression models.

Feature	Expected direction of effect	# mutations in feature-associated subtypes[a]	# mutations in feature		p-value	Consistent direction?
			expected[b]	observed[c]		
CpG Islands	Increased	648	17	14	0.65	No
	Decreased	7072	331	84	$\mathbf{7.5\times10^{-35}}$	**Yes**
GC content	Increased	256	22	8	0.99	No
	Decreased	2350	56	23	$\mathbf{1.1\times10^{-4}}$	**Yes**
H3K36me3	Increased	3731	589	682	$\mathbf{2.4\times10^{-3}}$	**Yes**
	Decreased	898	162	98	$\mathbf{1.3\times10^{-5}}$	**Yes**
H3K9me3	Increased	7361	1165	1905	$\mathbf{3.8\times10^{-51}}$	**Yes**
H3K27me3	Increased	896	274	252	0.86	No
H3K4me1	Decreased	2839	557	463	$\mathbf{1.2\times10^{-3}}$	**Yes**
H3K4me3	Decreased	3406	566	487	$\mathbf{8.1\times10^{-3}}$	**Yes**
DHS	Increased	1091	177	184	0.38	Yes
	Decreased	2898	701	645	**0.04**	**Yes**
Lamin-associated domains	Increased	485	187	171	0.85	No
Recombination rate	Increased	2190	306	377	$\mathbf{1.9\times10^{-3}}$	**Yes**
Replication timing	Increased	2359	278	321	**0.03**	**Yes**

For each feature for a given effect direction, we included all 7-mer subtypes with significant association of the feature with relative mutation rate (**Fig. 5**). We counted: [a]the total number of *de novo* mutations of those subtypes, [b]the number of these mutations expected to occur in the feature under the null expectation that the feature has no impact on mutability (i.e., assuming only an effect of sequence context), and [c]the number of these mutations that were observed to occur in the feature. Significant associations are indicated by a one-sided p-value (observed numbers are consistent with model predictions) in bold.

Supplementary Table 6b Univariate tests for enrichment or depletion of *de novo* mutations occurring in feature-associated subtypes (excluding CpG subtypes) identified by logistic regression models.

Feature	Expected direction of effect	# mutations in feature-associated subtypes[a]	# mutations in feature		p-value	Consistent direction?
			expected[b]	observed[c]		
H3K36me3	Increased	2844	356	474	$\mathbf{1.1\times10^{-5}}$	**Yes**
	Decreased	887	160	96	$\mathbf{1.8\times10^{-5}}$	**Yes**
H3K9me3	Increased	4050	728	1101	$\mathbf{4.9\times10^{-23}}$	**Yes**
H3K27me3	Increased	238	72	60	0.25	No
CpGI	Increased	610	12	9	0.64	No
DHS	Increased	1061	167	173	0.78	Yes
	Decreased	378	101	56	$\mathbf{8.4\times10^{-5}}$	**Yes**

For each feature, we identified all 7-mer subtypes where our model estimated a significant association (Fig. 3), separated into either an increased or decreased direction of effect, and counted: [a]the total number of non-CpG *de novo* mutations of those subtypes, [b]the number of these mutations that would occur in regions of the genome where that feature was present, under the null expectation that the feature has no impact on mutability (i.e., assuming only an effect of sequence context), and [c]the number of these mutations that were observed in the presence of that feature. Significant associations are indicated by a one-sided p-value in bold. Note that only 5 of the 15 groups described in **Supplementary Table 6a** contained sufficient numbers of non-CpG *de novo* mutations to perform these tests.

Supplementary Table 7 **Parameter estimates for genomic features model**

[see separate spreadsheet, table_S7_feature_parameter_estimates.xlsx]

This table contains effect size estimates and standard errors of 16 parameters (14 features, plus intercept and read depth) for each of the 24,489 7-mer subtypes with at least 10 singletons in the BRIDGES data.