Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans

Jedidiah Carlson[1], Laura J Scott[2], Adam E Locke[3], Matthew Flickinger[2], Shawn Levy[4], The BRIDGES Consortium[†], Richard M Myers[4], Michael Boehnke[2], Hyun Min Kang[2], Jun Z Li[1,5*], Sebastian Zöllner[2,6*]

[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

[†]*a full list of BRIDGES collaborators is provided in the supplementary material*

**both authors contributed equally*

Abstract

Precise estimates of the single-nucleotide mutation rate and its variability are essential to the study of human genome evolution and genetic diseases. Here we use ~36 million singleton variants observed in 3,716 whole-genome sequences to characterize the heterogeneity of germline mutation rates across the genome. Adjacent-nucleotide context is the strongest predictor of mutability, with mutation rates varying by >650-fold depending on the identity of three bases upstream or downstream of the mutated site. Histone modifications, replication timing, recombination rate, and other local genomic features further modify mutability; magnitude and direction of this modification varies with the sequence context. Compared to estimates based on common variants used in previous approaches, singleton-based estimates provide a more accurate prediction of the mutation patterns seen in an independent dataset of ~46,000 *de novo* mutations; and incorporating the effects of genomic features further improves the prediction. The effects of sequence contexts, genomic features, and their interactions reported here capture the most refined portrait to date of the germline mutation patterns in humans.

Introduction

Germline mutagenesis is a fundamental biological process, and underlies all heritable genetic variation (see Segurel et al.[1] for a review). Mutation rate models are widely used in genomics research to calibrate variant calling algorithms[2], infer demographic history[3], identify recent patterns of genome evolution[4], and interpret clinical sequencing data to prioritize likely pathogenic mutations[5]. While mutation is an inherently stochastic process, the distribution of mutations in the human genome is not uniform and is correlated with genomic and epigenomic features including local sequence context[6,7], recombination rate[8], and replication timing[9]. Hence, there is considerable interest in studying the regional variation and context dependency of mutation rates to understand the basic biology of mutational processes and to accurately model this variability.

The gold standard for studying the germline mutation rate in humans is direct observation of *de novo* mutations from family-based whole-genome sequencing data[9–12]. These studies have produced accurate estimates of the genome-wide average mutation rate ($\sim 1 - 1.5 \times 10^{-8}$ mutations per site per generation), and uncovered associations between mutation rates and certain features of the genomic landscape. However, given the inherently low germline mutation rate, family-based whole-genome sequencing studies have detected relatively few *de novo* mutations (<60,000 in ~1,000 trios reported so far, or 1-2 per 100,000 bases), making it difficult to precisely estimate mutation rates at a fine scale and identify factors that explain how the mutation rate varies across the genome.

Other studies have used between-species substitutions or within-species polymorphisms to study mutation patterns[7,8,13–16]. Because these variants arose hundreds or thousands of generations ago, they have been subject to evolutionary forces (such as natural selection) that nonrandomly alter the frequencies of variants in the population. To avoid the confounding effects of selection, these studies focus on substitutions/polymorphisms that occur in intergenic (non-coding) regions of the genome, which are less often the target of selective pressure. Nevertheless, even putatively neutral loci may be under some degree of selection[17–19], and are susceptible to GC-biased gene conversion (gBGC), in which recombination-induced mismatches are preferentially repaired to G/C base pairs,

resulting in a greater number of A/T-to-G/C polymorphisms than expected under neutral evolution[11,20,21]. Consequently, these processes bias the resulting distribution of variation, making it difficult to determine which patterns of variation are attributable to the initial mutation processes, and which are due to subsequent selection or gBGC.

We introduce an approach that leverages a new data source to characterize the regional variability of germline mutations at high resolution, and to quantify the influence of genomic context while minimizing the confounding effects of selection and gBGC. Our approach exploits a collection of >36 million extremely rare variants (ERVs), at a sample minor allele frequency (MAF) of 1/7432=0.00013 (i.e., singleton variants in our dataset). Compared to between-species substitutions or common variants in humans, these ERVs are extremely young on the evolutionary timescale, making them much less likely to be affected by evolutionary processes other than drift[1,11,20,22]. ERVs thus represent a relatively unbiased sample of recent mutations, and are far more numerous than *de novo* mutations collected in family-based studies.

Our results show that mutation rate heterogeneity is primarily dependent on sequence context of adjacent bases, and this sequence-specific mutability varies with respect to genomic features in wider surrounding regions, including replication timing, recombination rate, and histone modifications. Importantly, the impact of genomic features often depends on the actual sequence context at adjacent bases, demonstrating the importance of jointly analyzing sequence context and genomic features. Our sample of ~36M ERVs allows systematic estimates of these quantitative and often non-additive effects, providing a refined resource for modeling human germline mutations. We evaluate parameters estimated from ERVs in an independent dataset of *de novo* mutations and show that models that combine sequence context and genomic features substantially improve predictions of *de novo* mutations. Moreover, ERV-based predictions are less biased than predictions using parameters estimated from ancestrally older variants.

Results

ERV data source: variant calling and quality control

In the Bipolar Research in Deep Genome and Epigenome Sequencing (BRIDGES) study, we sequenced the genomes of 3,716 unrelated individuals of European ancestry to an average diploid-genome coverage of 9.6X; after variant calling and filtering, we identified 63,566,013 autosomal single-nucleotide variants in the mappable genome, 36,087,319 (57%) of which were singletons (ERVs), present in only a single individual from the BRIDGES sample (**Materials and Methods**). According to a simple demographic model with a growth rate of 4% per generation and effective *recent* population size of 500,000, we estimate that the average age of an ERV in this sample is approximately 46 generations[23]. We estimate fewer than 2.6% of these ERVs are false positives, which is similar to the validated singleton error rates of other large-scale sequencing studies[24–26], and we present evidence that any such errors are predominantly random across the genome, and are unlikely to be biased by motif-specific genotyping error (**Supplementary Table 1** and **Supplementary Note 1**).

Context-dependent variability in mutation rates

Prior trio sequencing studies have found that the bases immediately adjacent to a mutated site (i.e., a 3-mer sequence context) are an important predictor of variability in mutation rates across the genome[11,27]. To better understand these patterns of context-dependent variability, we examined the surrounding sequence of the 36,087,319 ERVs. We first assigned each ERV into one of six basic mutation types: A>C, A>G, A>T, C>T, C>G, and C>A. Note that each of these types also captures variants that occur in their reverse complement (e.g., the A>C type contains both A>C and T>G singletons, the C>T type contains both C>T and G>A singletons, etc.). We then identified the bases in the reference genome up to 3 positions upstream and downstream from each ERV. Considering all possible combinations of nucleotides in these positions, the 6 basic mutation types can be subdivided into $4^2 \times 6 = 96$ 3-mer subtypes, $4^4 \times 6 = 1{,}536$ 5-mer subtypes, or $4^6 \times 6 = 24{,}576$ 7-mer subtypes. For each K-mer subtype, we divide the number of ERVs observed at the central position of the K-mer by the

number of times the K-mer is seen in the mappable autosomal regions of the reference genome; we term this proportion the *estimated relative mutation rate*. For example, we observed 28,615 A>C or T>G singletons occurring in an AAAAAAA or TTTTTTT 7-mer motif (the underlined base indicates the variant site) and there are 11,823,995 such motifs in the reference genome where this subtype of mutation could be observed, yielding a relative mutation rate estimate of 28,615/11,823,995=0.0024 for this 7-mer subtype. In the absence of selection, gBGC, or sequencing artifacts, the relative mutation rate is proportional to the underlying per-site, per-generation *absolute* mutation rate. We show estimated relative mutation rates for all possible 1-, 3-, 5-, or 7-mer subtypes in **Figure 1** and **Supplementary Tables 2a-2d**.

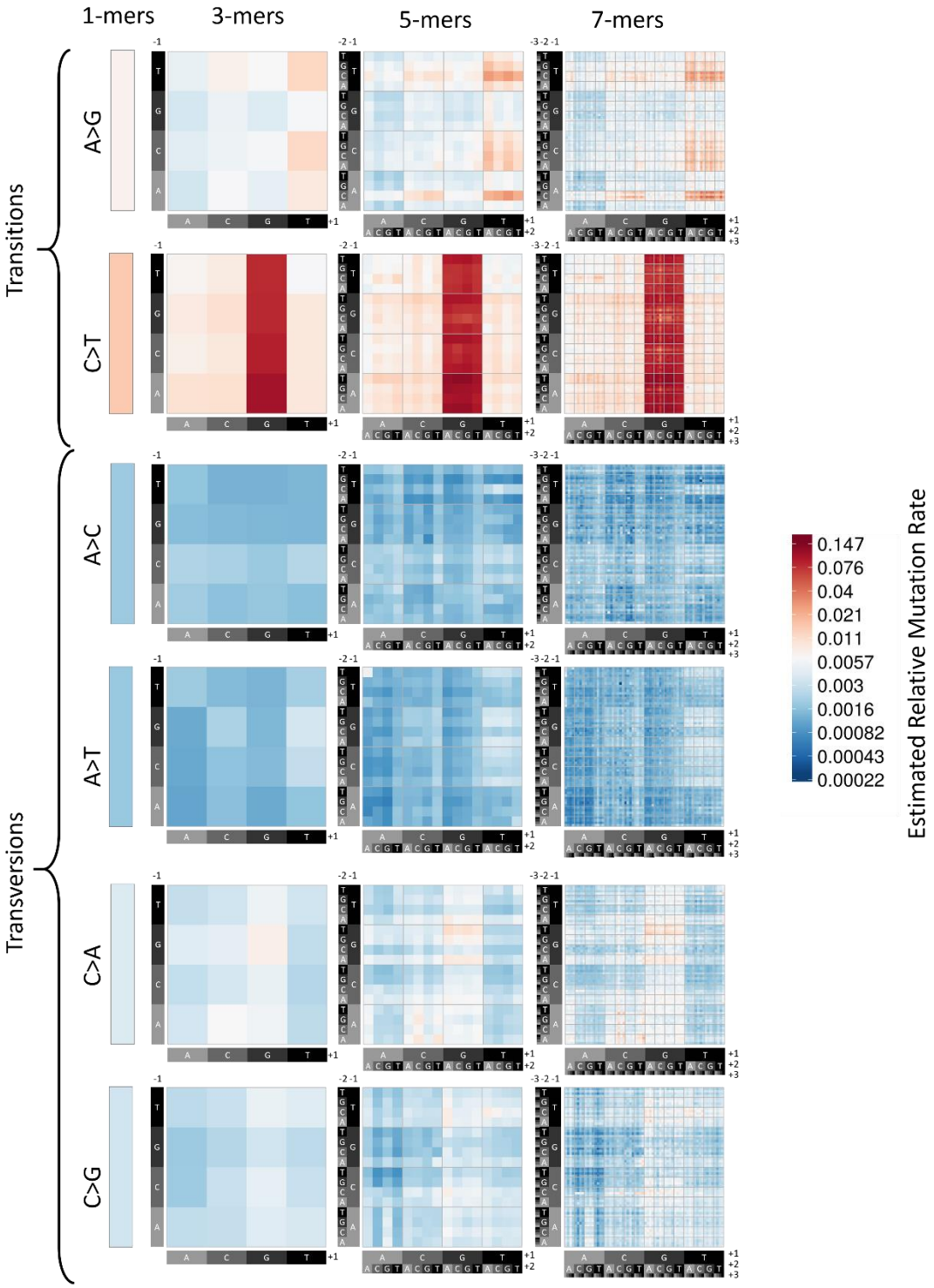

Figure 1 Heatmap of estimated relative mutation rates for all possible subtypes, defined by local sequence context up to a 7-mer resolution. The first column indicates the relative mutation rates for the 6 basic 1-mer types, and the subsequent columns show these rates stratified by increasingly longer sequence context. Each cell delineates a subtype defined by the bases upstream (y-axis) and downstream (x-axis) from the central base. Extreme values in the scale (0.00022 and 0.147) indicate the observed minimum and maximum estimated relative mutation rates among 7-mers.

The 6 basic 1-mer mutation types (**Fig. 1**, first column) distinguish the higher relative mutation rates of transitions (A>G; 0.0062 and C>T; 0.013) compared to transversions (A>C; 0.0016, A>T; 0.0016, C>A; 0.0034, and C>G; 0.0028). When each 1-mer mutation type is split into its 16 constituent 3-mer subtypes (**Fig. 1**, second column), we identify patterns of within-type mutation rate heterogeneity. Hypermutable N[C>T]G subtypes are the most prominent, attributable to increased C>T mutation rates at methylated CpG dinucleotides (relative mutation rate=0.097; 7.4-fold higher than 1-mer C>T relative mutation rate). Expanding to 5-mer subtypes (**Fig. 1**, third column) reveals additional context-dependent heterogeneity, for example, among CA[A>G]TN and CT[A>G]TN subtypes (relative mutation rate=0.024; 3.9-fold higher than 1-mer A>G relative mutation rate). Expansion to 7-mer subtypes (**Fig. 1**, fourth column) shows that relative mutation rates is also affected by the bases 3 positions upstream or downstream from the variant site. A notable example is NTT[A>T]AAA subtypes (relative mutation rate=0.0098; 6.1-fold higher than 1-mer A>T relative mutation rate). The full range of estimated 7-mer relative mutation rates spans nearly three orders of magnitude, with the lowest and highest 7-mer rates differing by 658-fold (lowest: CGT[A>C]TCG; relative mutation rate=0.00022; highest: TTA[C>T]GCA; relative mutation rate=0.146). Assessing the statistical significance of these differences, we found overwhelming evidence for all 96 of the 3-mer subtypes for heterogeneity in the relative mutation rates of their respective 5-mer constituents (all $P < 10^{-259}$). Further, 1462 (95.2%) of the 1536 5-mer subtypes had significantly heterogeneous rates among their respective 7-mer constituents after correcting for multiple testing ($P < 3.3 \times 10^{-5}$) (**Materials and Methods**).

Increasing subtype granularity improves prediction of *de novo* mutations

If the relative mutation rates for successively longer K-mers reflect genuine context-dependent heterogeneity of *de novo* mutation rates, then the probability of observing a true *de novo* mutation should be better predicted by models based on longer K-mers. To test this hypothesis, we fit logistic regression models using each set of ERV-derived relative mutation rates to estimate mutation probabilities of 46,813 autosomal *de novo* mutations observed in 1,074 children from the Genomes of

the Netherlands (GoNL) and Inova newborn screening (NBS) trio-sequencing studies[9,12] (**Materials and Methods**). Successively more variance (measured as Nagelkerke's pseudo-R^2) was explained by models that accounted for longer K-mers, with 8.2%, 13.6%, 14.3%, and 14.5% of the variance in *de novo* mutation rates explained by 1-mer, 3-mer, 5-mer, and 7-mer models, respectively (**Fig. 2a**). Likelihood ratio tests between successive models indicate that these differences are all statistically significant (1-mer to 3-mer: $P < 2.2 \times 10^{-308}$; 3-mer to 5-mer: $P < 2.2 \times 10^{-308}$; 5-mer to 7-mer: $P < 6.7 \times 10^{-136}$; **Supplementary Table 3a**). We also analyzed each mutation type separately to determine for which of the mutation types longer K-mers explained more variance. Here, we subdivided the C>T, C>G, and C>A types into CpG and non-CpG types, and analyzed the resulting 9 basic mutation types. For all 9 types, the 7-mer model provided a significantly better fit than smaller K-mer models, and explained 1.03-fold (A>T) to 1.59-fold (CpG>GpG) more variance than the 5-mer model, and 1.10-fold (A>T) to 4.18-fold (A>C) more variance than the 3-mer model (**Fig. 2b; Supplementary Table 3b**).

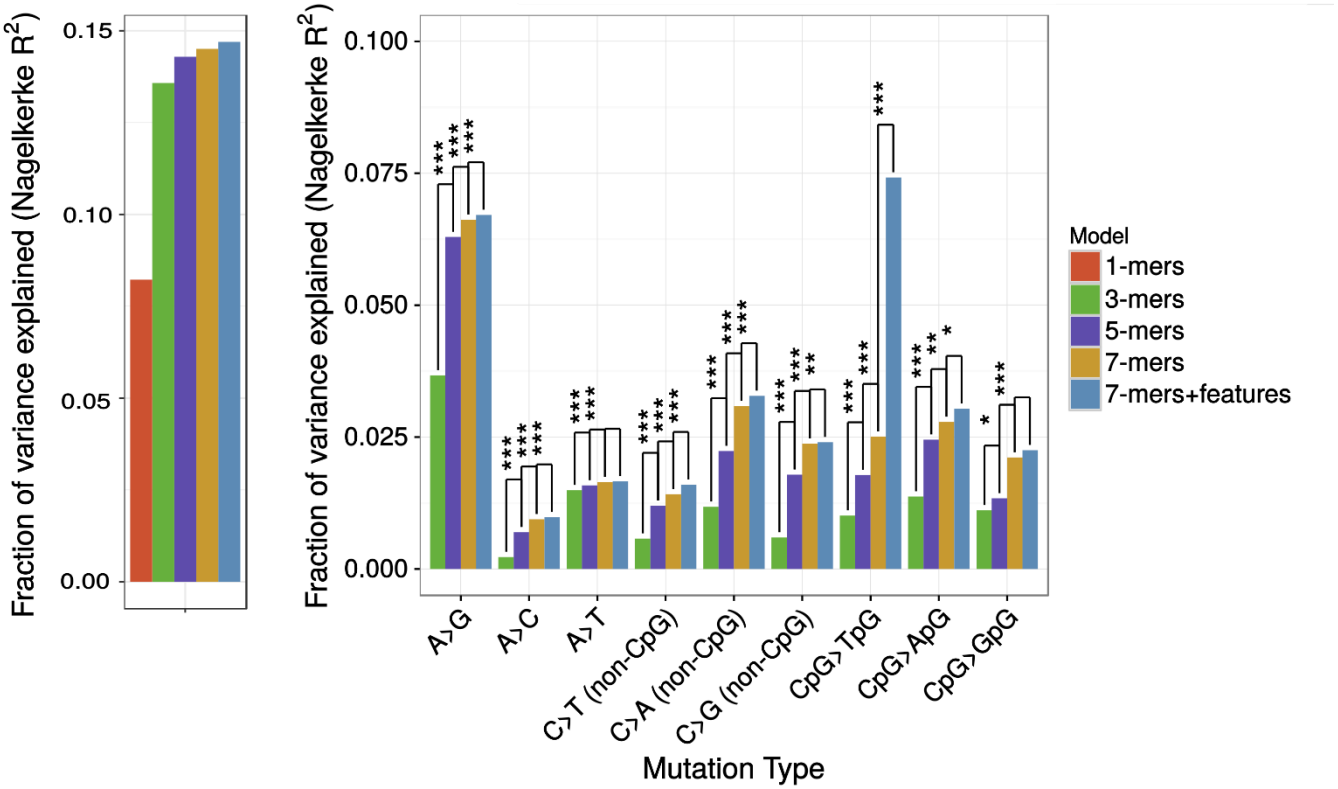

Figure 2 Comparison of model fit for different ERV-based mutation models. **(a)** Nagelkerke's R^2 for models using 1-mers (6 types), 1-mers with CpGs and non-CpGs considered separately (9 types), 3-mers (96 subtypes), 5-mers (1,536 subtypes), 7-mers (24,576 subtypes), and 7-mers+genomic features across all 9 mutation types combined. (The effect of genomic features is described in a section below.) **(b)** Nagelkerke's R^2 for 3-mer, 5-mer, 7-mer, and 7-mer+genomic features models run separately for each mutation type. The p-value from the likelihood ratio test (with 1 degree of freedom) comparing each model with the next model containing an additional parameter is indicated above each pair of bars (***P < 0.001; **P < 0.01; *P < 0.05) Exact p-values are reported in **Supplementary Table 3**.

Discordant mutation signatures of ERVs and common polymorphisms

Our finding that longer K-mers improve prediction of *de novo* mutations is not unexpected; previous studies using polymorphism data have demonstrated that local sequence context is an important predictor of polymorphism rate heterogeneity[7,8,13]. However, estimates derived from common polymorphisms may not accurately represent the basal mutation processes due to the confounding effects of selection and gBGC[1]. We compared our ERV-derived 7-mer relative mutation rates to 7-mer relative mutation rates calculated among 12,088,037 variants with a minor allele count ≥10 in the BRIDGES data, which represent ancestrally older variants (hereafter referred to as the "MAC10+" set of variants). To control for the effects of sample size, we downsampled the 36,087,319 singletons to 12,088,037 and recalculated the 7-mer relative mutation rates on this subset. Across all 24,576 7-mer mutation types, the ERV-derived and MAC10+-derived relative mutation rates were highly correlated (Spearman's r=0.99; **Fig. 3a**). When stratified by mutation type, however, we observed much weaker correlations between the ERV-derived and MAC10+-derived relative mutation rate estimates (r=0.45 to 0.95; **Fig. 3b**). Under the null hypothesis that neither ERV-derived nor MAC10+-derived rates are biased, the respective distributions of these rates should be similar. We used Kolmogorov-Smirnov tests to compare the two distributions of relative mutation rates for each type, and found strong evidence for each of the 9 types that the two distributions were not equivalent ($P < 2.3 \times 10^{-5}$; **Supplementary Fig. 1**). Moreover, for every type except A>T transversions, we found that the discordance between ERV-derived and MAC10+-derived rates was not random with respect to sequence context, but varied between low-GC motifs (3 or fewer GC bases in flanking region) and high-GC motifs (4-6 G/C bases in flanking region) (**Fig. 3c**; **Supplementary Table 4**). This distinction is particularly strong for A>G transition subtypes ($P < 2.4 \times 10^{-161}$). A>G polymorphisms are known to occur at a frequency higher than expected under neutral evolution, because gBGC during meiotic recombination results in preferential transmission of G/C alleles[11,20]. Moreover, gBGC tends to occur most often in GC-rich DNA[21,28], consistent with our observation of an excess of GC-rich motifs with >2-fold higher MAC10+ rates compared to ERV rates (**Supplementary Fig. 2**).

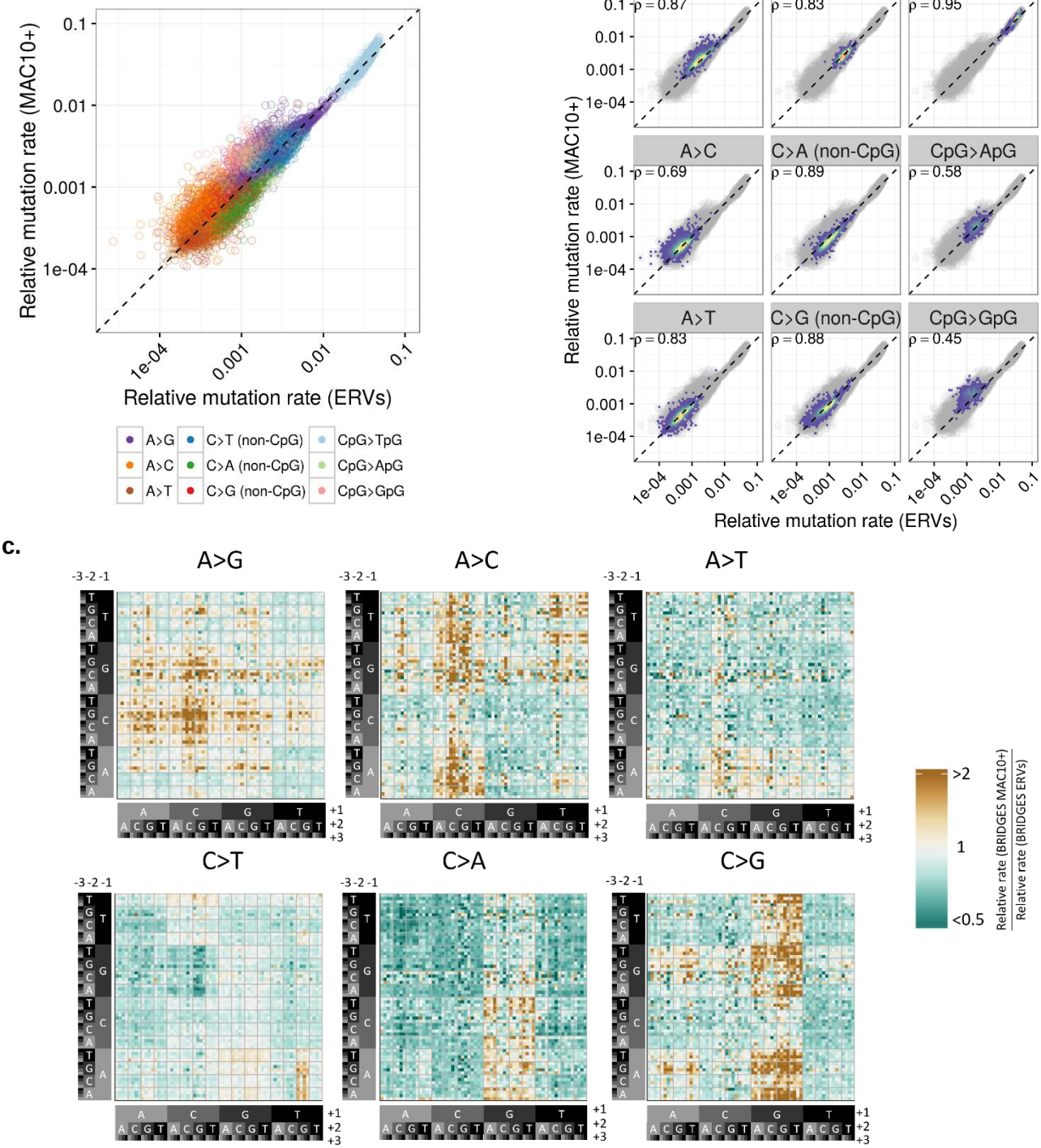

Figure 3 (a) Relationship between 7-mer relative mutation rates estimated using ERVs (x-axis) and variants with a minor allele count >= 10 (MAC10+; y-axis) on a log-log scale. We note that the strength of this correlation is driven by hypermutable CpG>TpG transitions. **(b)** Type-specific 2D-density 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.

These patterns remained after excluding subtypes with fewer than 50 ERVs or MAC10+ variants, thus are unlikely to be an artifact of low-frequency subtypes (**Supplementary Fig. 3**). Likewise, the type-specific MAC10+-derived relative mutation rates were strongly correlated with rates independently estimated in a previous study based on 7,051,667 intergenic variants observed in 379 European individuals sequenced by the 1000 Genomes Project[7], indicating the distribution of ancestrally older BRIDGES variants is consistent with what we would expect in a similar population (**Supplementary Fig. 4**; **Supplementary Note 2**). Collectively, these results point to systematic differences in the estimated context-specific effects between common and rare variants. The two most plausible sources of such differences are either that 1) even after our careful data cleaning and filtering, the ERV-derived rates are affected by context-specific sequencing errors, or 2) the MAC10+-derived estimates are under the influence of non-mutational evolutionary processes that have altered the relative frequencies among subtypes.

ERVs predict occurrence of de novo mutations more accurately than MAC10+ variants

To distinguish the origin of these systematic differences, we compare the ability of both sets of rates to predict *de novo* mutations by fitting logistic regression models to predict the 46,813 autosomal *de novo* mutations described above, using either the ERV-derived or MAC10+-derived 7-mer relative mutation rates as predictors (**Materials and Methods**). We reasoned that if ERV-derived rates are biased by sequencing artifacts, the ERV model will explain less variance in the independent testing data than the MAC10+ model, whereas if MAC10+-derived rates are biased by evolutionary processes, the ERV model will explain more variance than the MAC10+ model. Results of this comparison revealed that the ERV model explains more variance (i.e., higher Nagelkerke's R^2) than the MAC10+ model overall (11.9% vs. 11.4%; **Fig. 4**), and with lower AIC (314,723 vs 316,400; **Supplementary Table 3a**). When stratifying the *de novo* testing data by the 9 types, the ERV model also explains more within-type variance than the MAC10+ model (**Supplementary Table 3b**).

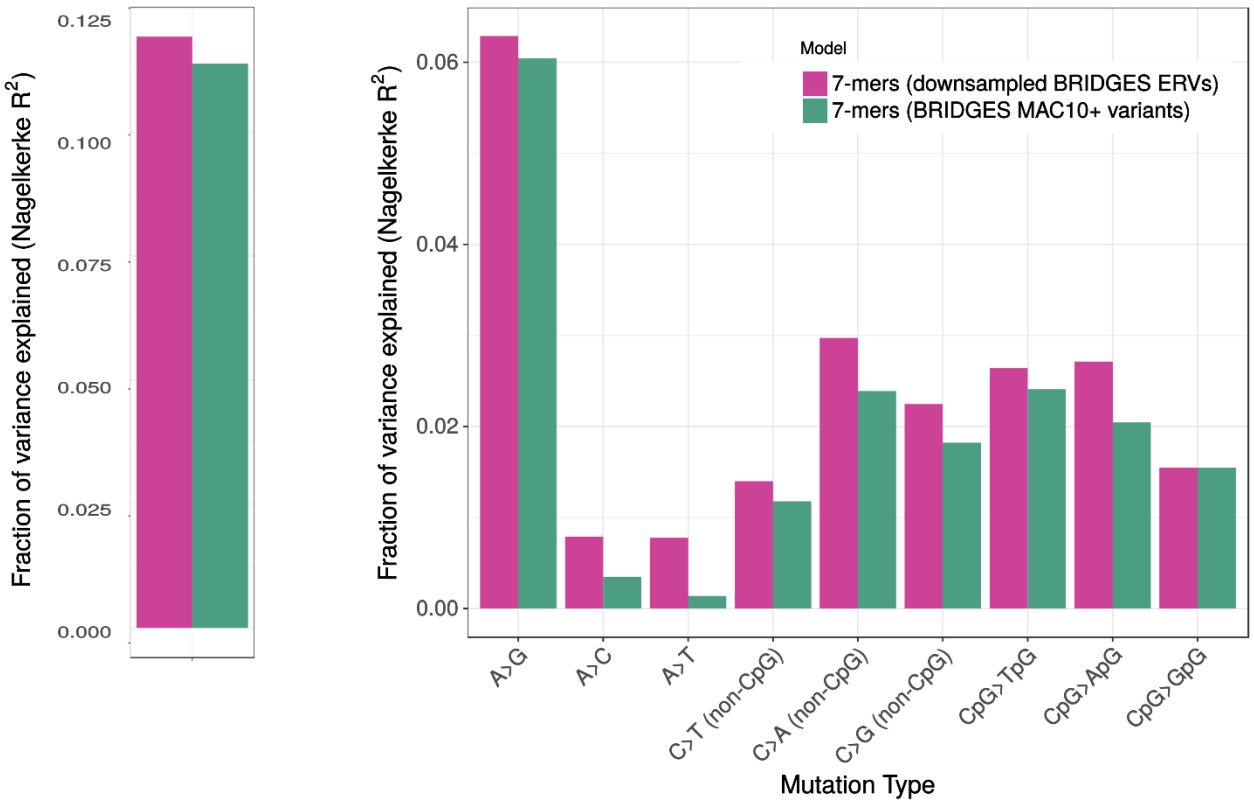

Figure 4 Variance in *de novo* mutation rates explained by ERV-derived vs. MAC10+-derived 7-mer relative mutation rate models, measured as Nagelkerke's R^2 for each model across **(a)** all mutation types combined, and **(b)** stratified by each basic mutation type. Exact values for Nagelkerke's R^2 and AIC of these models are reported in **Supplementary Table 3**.

Effects of genomic features vary by mutation type and sequence context

Family-based sequencing studies have demonstrated that germline mutation rates are associated with genomic features such as chromatin structure, replication timing, and recombination rate[9,11,27], but none of these studies have accounted for the joint influence of multiple features and local sequence context. To more precisely model how various features influence mutability, we used the BRIDGES ERVs to jointly estimated the effects of 14 genomic features (**Supplementary Table 5**) on relative mutation rates across the genome at a single-base resolution (**Materials and Methods**). For each of the 24,489 7-mers with at least 10 ERVs, we estimated the effects of 14 features, resulting in 342,846 parameter estimates. The distributions of effects for each feature by mutation type shown in **Supplementary Fig. 5.** To identify significant effects among the many associations tested, we applied a false discovery rate (FDR) cutoff of 0.05 to the p-values for each feature across all subtype-specific models. Of the 24,489 7-mer subtypes analyzed, 3,940 had at least one of the 14 genomic features significantly associated with mutability, with 6,514 significant associations among 342,846 tests. The distributions of these significant associations, separated by feature and mutation type, are shown in **Fig. 5**. Eleven of the 14 features had a significant association with relative mutation rate for at least one 7-mer; no significant associations were detected for exons, H3K27ac peaks, or H3K9ac peaks.

While seven of the 11 features had consistent directions of effect across all significantly associated 7-mer subtypes (top two rows, **Fig. 5**), four other features (H3K36me3 peaks, DNase hypersensitive sites [DHS], GC content, and CpG islands) exhibited bidirectional effects, both within and between mutation types (bottom row, **Fig. 5**). For a notable example, 99% (1021/1024) of CpG>TpG 7-mer subtypes are negatively influenced by CpG islands (likely because CpG islands are generally unmethylated[29], and thus less likely to mutate). Within five other mutation types, in a small fraction of subtypes (1.4%), CpG islands were associated with higher relative mutation rates and among non-CpG C>T subtypes associated with CpG islands, we observe both positive and negative associations.

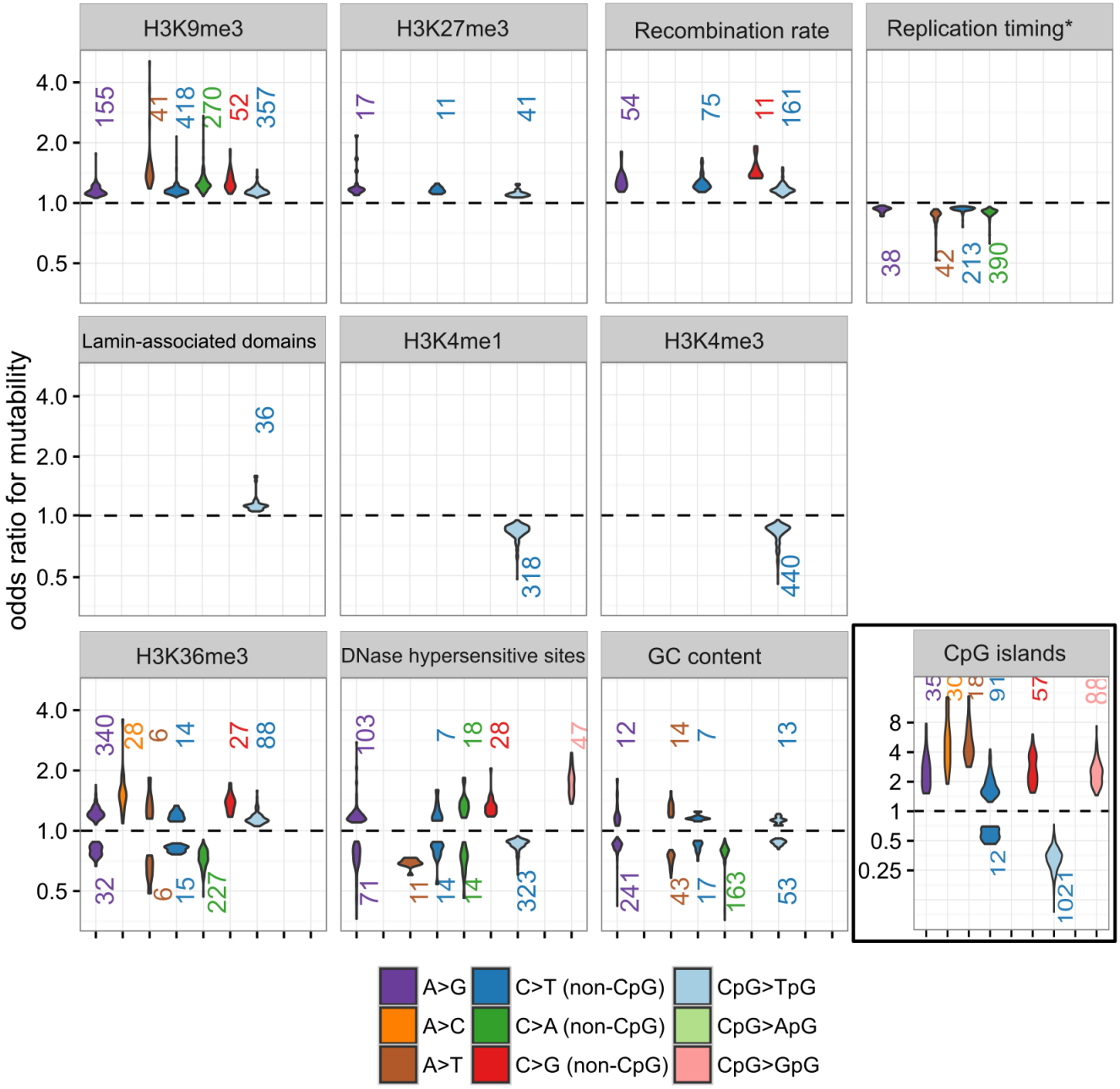

Figure 5 Distributions of significant mutagenic effects for 11 genomic features. For each feature, we identified all statistically significant (FDR<0.05) effects among the 24,576 7-mer subtypes, and plotted the empirical distributions of these subtype-specific odds ratios for each basic mutation type, separated by positive associations (odds > 1; above dashed line) and negative associations (odds < 1; below dashed line). The number of 7-mer subtypes within each type for which that feature is statistically significant in a positive or negative direction is shown above each or below distribution. Distributions are only shown for types with 5 or more 7-mer subtypes associated in the same direction. *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. Odds ratios for the 3 continuous features (recombination rate, replication timing, and GC content) indicate the change in mutability per 10% increase in the value of that feature.

291 ***De novo* mutations correlate with local genomic features**

292 To evaluate the importance of accounting for genomic features in mutation rate models, we used the
293 relative mutation rates estimated from 7-mers and genomic features to predict the 46,813 *de novo*
294 mutations from the GoNL[9] and Inova NBS[12] studies. We compared model fit statistics (likelihood ratio
295 test, AIC, and Nagelkerke's R^2) between this 7-mer+features model and the 7-mer-only model
296 (**Materials and Methods**). The 7-mer+features model explained significantly more variation in the *de*
297 *novo* mutations than the 7-mer-only model, both overall ($P = 3.1 \times 10^{-147}$; **Fig. 2a; Supplementary**
298 **Table 3a**) and within 6 of the 9 basic mutation types ($P < 0.05$), except for A>C transversions ($P =$
299 0.38), CpG>ApG transversions ($P = 0.18$), and CpG>GpG transversions ($P = 0.08$) (**Fig. 2b;**
300 **Supplementary Table 3b**). Including genomic features had the largest effect on the prediction of
301 CpG>TpG transitions, and nearly tripled the variance explained by the 7-mer-only model (7-mers-only:
302 2.7%; 7-mer+features: 7.5%; $P < 7.23 \times 10^{-108}$).

303 These models evaluate the aggregate effects of genomic features, but do not confirm the
304 specific effects of individual features described in **Fig. 5**. To address this, we examined each of the 15
305 features across all 7-mer subtypes that showed significant increase or decrease in mutation rate when
306 associated with that feature in the ERV data, and tested for a corresponding enrichment or depletion of
307 true *de novo* mutations in genomic regions where that feature is present (**Supplementary Note 3**). Of
308 the 15 tests performed, 10 showed significant enrichment or depletion of *de novo* mutations within
309 genomic regions predicted to affect mutability (**Supplementary Table 6**).

310

311 **Context-dependency reveals potential mechanisms of hypermutability**

312 **T**hese patterns of fine scale variability can indicate previously unknown context-dependent mutation
313 mechanisms in the germline. Here we describe two examples. Recent studies of various cancers
314 revealed an elevated somatic mutation rate in transcription factor binding sites within DNase
315 hypersensitive sites (DHS), likely caused by inhibition of nucleotide excision repair machinery[30–32]. One
316 of the most common binding targets in the genome is the CCAAT motif, which is targeted by a family of

transcription factors known as CCAAT/Enhancer Binding Proteins (CEBPs)[33]. Because CEBP binding sites were found to be significantly enriched for somatic mutations in multiple cancer types[30], we hypothesized that a similar mechanism may be operative in the germline. Adjusting for other genomic features, our model indeed shows DHS are significantly enriched for A>G (but not A>C or A>T) singletons at five of the 16 possible CCAATNN motifs (1.1 to 1.3-fold enrichment; $P < 9.3 \times 10^{-4}$), indicating that transcription factor binding may also associate with increased mutability in the germline.

A second example are 5'-TTAAAA-3' hexamer motifs, which harbor A>T mutations at a rate ~6.1-fold higher than the background (1-mer) A>T rate (**Supplementary Table 2d**). In ATTAAAA or TTTAAAA motifs occurring in DNase hypersensitive sites, however, the mutation rate is reduced by over 3-fold ($P < 2.8 \times 10^{-22}$). This TTAAAA hexamer is a common insertion target for LINE-1 retrotransposons and *Alu* elements[34], and is known to be nicked by L1 endonuclease (L1 EN) at the A/T dinucleotide, even when no retrotransposition takes place[35]. Moreover, the rate of L1 EN-induced damage has been shown to vary according to the nucleosomal context of target motifs[36], consistent with the DHS association detected by our model. Overall, this pattern of sequence- and feature-dependent mutability suggests that L1 EN nicks are mutagenic, resulting in A>T transversions. A more detailed analysis of the potential sources behind this mutation signature is presented in **Supplementary Note 4**.

Discussion

Our results highlight the relationship between a site's genomic context and likelihood of mutation, and demonstrate the importance of accounting for adjacent nucleotides, genomic features, and their joint effects in mutation models. Though prior family-based studies have shown that di- and tri-nucleotide motifs are predictive of mutation[9,11,27], the density of singletons in our dataset enables an even more granular description of sequence specificity, with a combination of bases up to 3 positions upstream and/or downstream from a given site influencing that site's mutability. Collectively, these results suggest that the nucleotides up to 3 positions away from a mutation site are a nontrivial source of

mutation rate heterogeneity, and even more distal nucleotides may have an appreciable effect on mutability.

Mutation probabilities of heptameric motifs have been catalogued by others using variants across the entire frequency spectrum[7], but we show that exclusively using the rarest variants results in more accurate inference of context-dependent *de novo* mutation patterns. This result has two important implications. First, we conclude that variants that occur at higher frequency are likely subject to biased evolutionary processes. Second, any cryptic biases in sequencing that affect detection of ERVs are relatively weak in comparison to evolutionary biases. This is most clearly exemplified in our observation of increased A>G polymorphism rates: this pattern is consistent with known effects of gBGC, and supports the conclusion that mutation models based on older variants are affected by evolutionary biases.

Building upon previous attempts to holistically model the relationship between sequence context, genomic features, and mutation rate[9,27], we extend our 7-mer model to account for the potential mutagenicity of multiple genomic features. Our model both confirms the presence of known feature-associated effects (such as a higher mutation rate in late-replicating regions[9]), and identifies additional associations (such as the bidirectional effects found in DHS and CpG islands). Moreover, we describe how such effects vary between and within mutation types according to local sequence context, providing new insight into how the genomic landscape shapes the mutation rate. These results suggest that a single genomic feature may act to both suppress and promote mutability of different mutation types and, in some cases, the direction of the effect can vary between different sequence motifs of the same basic mutation type. These findings may be used to generate additional testable hypotheses for how and where specific mutation mechanisms act in the genome.

We note that power to detect feature-associated mutability was in some cases limited by the number of ERVs of a given 7-mer subtype; 93% of the 6,514 significant associations were detected in 7-mer subtypes with more than 741 ERVs (the median number of ERVs per 7-mer subtype). Although we show genomic features to associate with observed patterns of mutation, the features used in our

model are a generic approximation of the genomic landscape in germ cells, and cannot account for changes to the genomic features through different stages of gametogenesis[11]. Additionally, because ERVs in our model reflect mutations that have occurred in both males and females over dozens of generations, we are unable to model sex-specific mutation rate heterogeneity or parental age effects[9,12,37]. Prediction of *de novo* mutations will almost certainly be improved by accounting for these temporal and cellular sources of variability.

Despite these limitations, our model describes patterns of mutability at an unprecedented level of detail, and with greater accuracy than previous models. The catalog of predicted single-base mutation rates we have generated can be integrated easily into existing analysis and simulation pipelines. To better incorporate this mutation rate map, we have developed a utility to annotate genomic variants with the estimated mutation rate, and generated a genome browser track to visualize predicted mutation rates alongside other genomic data. We envision a variety of applications for this data resource: for example, single-base mutation rates can be aggregated over an entire gene or regulatory locus to accurately estimate the number of mutations in "typical" individual. This is a crucial step in identifying which functional elements might be disrupted by *de novo* mutations in affected cases[38]. Pathogenicity scoring algorithms such as CADD[39] that rely on comparisons between observed and simulated variants can potentially benefit by adopting a context-dependent simulation model. Moreover, we expect that the analytical framework presented here will serve as a baseline for understanding mutation rate heterogeneity and its consequences, providing insight into both the history and future of human evolution and genetic diseases.

Acknowledgements

Funding for this research was provided by US National Institutes of Health (NIH) US National Institutes of Health grant R01HG005855 (S.Z. and J.L.). J.C. was supported by the NIH/National Human Genome Research Institute Genome Science Training Program (T32HG00040). The BRIDGES study was supported by NIH grants R01MH094145 (M.B., R.M.M.) and U01MH105653 (M.B.). Additional acknowledgements from collaborating members of the BRIDGES consortium are detailed in the **Supplementary Information**.

Author contributions

J.C., J.L., and S.Z. designed the models and wrote the manuscript. J.C. performed the analyses and created the online annotation utility and interactive heatmap. L.S., M. B., and H.M.K. provided critical feedback and evaluation of the manuscript. A.L., M.F., and H.M.K. performed variant calling and filtering of the BRIDGES samples and curated the raw data. Sequencing was led by S.L. and R.M.

Methods

Sample description. The BRIDGES sample contains 3716 European American bipolar disorder (BD) cases and controls. In all studies, DNA was extracted from blood-based samples. The BD cases and controls from the Centre for Addiction and Mental Health (CAMH) in Toronto (n=801) and from the Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London in London, U.K (n=817) were collected as previously described[40]. The BD cases and controls were collected as previously described for the Genomic Psychiatry Cohort (GPC) (n=1,079)[41] and the Prechter Repository (n=296)[42]. The STEP-BD BD cases (n=302), obtained from the NIMH repository, were collected as previously described[43]. Control samples were obtained from the Minnesota Center for Twin and Family Research (MCTFR) study (n=421)[44]. All human research was approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. All participants provided written informed consent.

Sample quality control and library preparation. The concentration of each DNA sample was measured by fluorometric means (PicoGreen, Thermo Fisher, Woburn, MA, USA) followed by agarose gel electrophoresis to verify the integrity of DNA. Six-hundred nanograms of DNA was sheared with acoustic shearing (Covaris, Woburn, MA, USA) to an average size of 400nt. Following shearing, the samples are transformed to a sequencing library using standard protocols to create a paired-end library. Briefly, sheared DNA was end-repaired, A-tailed and ligated with Illumina adaptors (New England Biolabs, Ipswitch, MA, USA). Following ligation, indexed primers were used to amplify the final libraries for each sample. Each sample received two indexes: 96 i7 indexes were used to identify each sample in each 96-well reaction plate while a single i5 index was used for each plate. This combination of indexes uniquely coded all samples in the project when both the i7 and i5 indexes were read at during sequencing. Following six cycles of PCR (Kapa Biosystems, Wilmington, MA, USA), libraries were purified and quality controlled by assaying the final library size using the Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and quantitating the final library via real-time PCR

(Kappa Biosciences). A single peak between 300-400bp indicates a properly constructed and amplified library ready for sequencing. PCR cycles for amplification are kept to a minimum to minimize PCR duplication rate and maximize library complexity.

Sequencing. Sequencing was performed per Illumina protocol, essentially as described by Bentley et al.[45]. Libraries were pooled in sets of 12 samples and each pool sequenced on a single lane of a HiSeq 2500 flowcell using version 3 Illumina chemistry at paired-end 100nt read lengths. Each library pool was loaded at 13pM to generate 160-180M paired reads per lane. Multiple flowcells of the library pools were performed to generate a final data set corresponding to 5-12x genome coverage per sample (Average coverage 9.6x). Reads were mapped to Build 37 of the human reference genome, with decoy sequence[24]. Alignment and variant calling were performed using the GotCloud pipeline[46].

Identification of singleton variants. We identified 36,087,319 singleton variants, including 314 that occurred at 157 multiallelic sites (i.e., two singletons occurring at the same site but representing two different alternative alleles. Notably, these all occurred at tetralleic sites where the third non-reference allele was a non-singleton variant). All downstream analyses included these multiallelic singletons as independent observations. To assess the quality of singleton variants, we compared quality metrics between this full set of singletons with a more stringently filtered subset of singletons, including only singletons with a variant quality score ≥30, average mapping quality score >56, and falling in the 1000 Genomes Strict Accessibility Mask generated by the 1000 Genomes Project, which restricts variant calls to the most uniquely mappable regions of the genome, comprising 72% of non-N bases[47].

We compared data quality measures (Ts/Tv ratio and % in dbSNP, build 142) both before and after applying these filters (**Supplementary Table 1**). The transition:transversion ratio (Ts/Tv) is a commonly used diagnostic to assess the overall quality of variant calls from sequencing data. Prior studies have suggested a Ts/Tv between 2.00 and 2.10 is to be expected for true variants in genome-wide datasets[48], whereas this ratio is close to 0.5 for random errors. Without downstream filters, the full

454 set of singletons have Ts/Tv ratio 2.02, compared to 2.00 for singletons within the strict filter set. If we
455 assume the Ts/Tv of true positive singletons is between 2.00 and 2.10, our observed Ts/Tv of 2.02
456 corresponds to a false positive rate of 0.6% to 2.6% (details described in **Supplementary Note 5**). We
457 may also estimate a false negative rate by analyzing the site frequency spectrum. Multiple sequencing
458 studies have reported an abundance of rare variants, particularly singletons, in European populations,
459 consistent with a model of recent faster-than-exponential population growth; in these studies, singletons
460 comprise 50-60% of all observed single-nucleotide variants[49,50]. In the BRIDGES data, singletons
461 comprise 56.8% of all variants (36,087,319 of 63,566,013). If we assume 2.6% of these singletons are
462 false positives, and further assume that non-singletons are all true positives, 55.3% of variants would
463 be true singletons, consistent with the proportion of variants reported to be singletons in the ExAC
464 exome sequencing study (54%)[51], suggesting a low false negative rate for singletons in the BRIDGES
465 data.

466 Although these estimates suggest that most singletons are high-quality calls, specific sources of
467 bias may occur, such as 1) error-prone sequence motifs may have an excess of false positives, leading
468 to overestimated mutation rates for the corresponding mutation subtypes; or 2) true variants that are
469 incorrectly mapped may confound our analysis of the relationship between genomic features and
470 mutation. We discuss in **Supplementary Note 1** why these concerns are unlikely to have made a
471 strong contribution to the main findings. Hence, we decided to use the full set of singletons (rather than
472 the stringently filtered subset) in our analyses.
473
474 **Mutation subtypes and subtype-specific relative mutation rates.** We assigned each of the
475 36,087,319 singletons into one of 6 basic mutation types: A>C, A>G, A>T, C>T, C>G, and C>A. Where
476 applicable, we separated C>T, C>G, and C>A mutations into CpG and non-CpG types (for a total of 9
477 basic mutation types), to account for the known hypermutability of cytosine at CpG dinculeotides[52,53].
478 The notation of A>C includes both A-to-C mutations and complementary T-to-G mutations. For each
479 mutation type, we further define a set of mutation subtypes by the bases flanking the variant site. Since

there are 4 possible bases at both the +1 position and the -1 position, there are 4x4=16 possible 3-mers containing each basic mutation type at the central position, producing 6x16=96 3-mer subtypes. Likewise, there are $6x4^4$=1,536 5-mer subtypes, and $6x4^6$=24,576 7-mer subtypes.

For every subtype, we calculate a genome-wide relative mutation rate as the observed number of singletons of that subtype divided by the number of K-mers in the accessible reference genome that could generate that subtype. K-mers in the reference genome were counted by a 1-bp sliding window, so that every possible K-mer was accounted for (e.g., a run of 4 As is counted as two AAA 3-mers shifted by one base). For simplicity, we denote a subtype by the sequence motif containing either an A or a C as the reference base at the central position (e.g., either CGT[A>X]TCG or CGT[C>X]TCG). As each K-mer can be split into 16 possible (K+2)-mers that share the same internal motif but differ in their terminal bases, the relative mutation rate for each K-mer subtype is the weighted mean of the rates found among its 16 possible (K+2)-mer constituent subtypes. To assess the heterogeneity of relative mutation rates among each set of 16 (K+2)-bp constituent subtypes that share the same K-bp motif, we performed a chi-squared test for uniformity of these rates, with each test having 15 degrees of freedom. We applied Bonferroni correction to account for the multiple testing burden of this procedure, and considered a test to be statistically significant if $P < \frac{0.05}{N}$, where N is the number of tests (i.e., N=96 when testing for heterogeneity of each 3-mer subtype's constituents, and N=1,536 when testing for heterogeneity of each 5-mer subtype's constituents).

Mutation prediction model and validation. We used these ERV-derived relative mutation rate estimates as predictors in a series of logistic regression models to determine if longer sequence context more accurately described the observed distribution of true *de novo* mutations. The dependent variable in the logistic regression models was whether or not a site was one of 46,813 *de novo* mutations identified by two independent studies: 11,020 *de novo* mutations detected in 258 Dutch families by the Genomes of the Netherlands (GoNL) project[9], and 35,793 *de novo* mutations from 816 families sequenced by the Inova Newborn Screening Cohort[12]. These observed mutations were combined with

1 million randomly selected sites from the mappable autosomal regions of the reference genome to serve as a non-mutated background. Because each non-mutated site can be ambiguously considered as the background for 3 different mutation types, we divided the 1 million non-mutated sites into 3 non-overlapping sets. We designated A/T and C/G reference bases in the first set (consisting of 333,334 unique sites) as non-mutated A>G and C>T types, respectively, and so on for the second set (A>C or C>G types), and the third set (A>T or C>A types), each of which contained 333,333 unique sites. So, we consider a total of 1,046,813 testing sites (1,000,000 unmutated sites and 46,813 de novo mutations), each with one possible mutation event, in our prediction models.

Now let $\boldsymbol{i} = \{\mathbf{1}, \ldots, \mathbf{1,046,813}\}$ be an index for the 1,046,813 testing sites. We coded $\boldsymbol{D_i} = \mathbf{1}$ if site $\boldsymbol{i}$ is a de novo mutation and $\boldsymbol{D_i} = \mathbf{0}$ otherwise. We then annotated each of the testing sites with four quantities, according to the site's sequence context: 1) R_1=the 1-mer relative mutation rate 2) R_3=the difference between the 3-mer relative mutation rate and 1-mer relative mutation rate, 3) R_5=the difference between 5-mer relative mutation rate and 3-mer relative mutation rate, and 4) R_7=the difference between 7-mer relative mutation rate and 5-mer relative mutation rate, so the overall estimated 7-mer relative mutation rate for site $\boldsymbol{i}$ can be decomposed into $\boldsymbol{R_{i,1} + R_{i,3} + R_{i,5} + R_{i,7}}$. Then we can assess how well the estimated relative mutation rates for the 7-mer model predict the observed mutation rates by fitting the logistic model:

$$\boldsymbol{ln\left(\frac{P(D_i = 1)}{P(D_i = 0)}\right) = \alpha_0 + \alpha_1 R_{i,1} + \alpha_3 R_{i,3} + \alpha_5 R_{i,5} + \alpha_7 R_{i,7}} \qquad (\mathbf{1})$$

and calculating the maximized likelihood. We can similarly assess the 5-mer, 3-mer, and 1-mer models by removing all terms of a higher order than the specified K-mer from the logistic model (1). We structured models in this way so that each K-mer model is nested within the next (K+2)-mer model, enabling us to statistically evaluate model fit using a likelihood ratio test (each successive test between a K-mer model and (K+2)-mer model having 1 degree of freedom). We also compared model fit using Nagelkerke's R^2 as a measure of variance in *de novo* mutation rates explained.

Because these models evaluate trends jointly across the basic mutation types, they do not provide information about which types benefit most strongly from using expanded sequence motifs. For example, it is possible that any improvement to the overall model fit is elicited by context-dependent heterogeneity of a single mutation type, whereas other types might not be significantly affected by using longer sequence motifs, and do not contribute to the improved model fit. To identify these type-specific trends, we stratified our testing data by each of the 9 basic types, and for each type, repeated the 3-mer, 5-mer, and 7-mer models on only the sites of that type. Within each set of type-specific models, we again used likelihood ratio tests to evaluate if relative mutation rates estimated in longer K-mers result in significantly better prediction of *de novo* mutations. We also calculated Nagelkerke's R^2 to quantify the amount of within-type variability explained by a given model.

Comparison of ERV and polymorphism models. To compare the ERV-derived, MAC10+-derived, and 1000 Genomes-derived[7] 7-mer mutation rates, we applied a logistic modeling framework similar to equation (1). Because the 1000 Genomes-derived rates are only available for 7-mers[7], we used only the respective 7-mer relative mutation rate as a single predictor in each model, without decomposing into four separate components. Each of these models was applied to predict the probability of a *de novo* mutation at each of the 1,046,813 testing sites. These 7-mer models are based on different sets of predictors and are not nested, so their relative performance cannot be assessed with a likelihood ratio test; instead, we compared model fit with Akaike Information Content (AIC) and Nagelkerke's R^2.

Estimating the effect of local genomic features. We estimated the effect of 14 genomic features (data sources for these features are described in **Supplementary Table 5**) on the relative mutation rate of each 7-mer subtype using the modeling framework described below. Let $\boldsymbol{K}$ be the index across all 7-mer subtypes with more than 10 observed singletons ($\boldsymbol{K} \in \{\mathbf{1}, \ldots, \mathbf{24}, \mathbf{489}\}$). Let $\boldsymbol{j_K}$ be the index across all sites that are centered at the 7-mer motif that could produce a mutation of subtype $\boldsymbol{K}$, and let $\boldsymbol{Z_{j_K}} = \mathbf{1}$ if the site carries a singleton of subtype $\boldsymbol{K}$ and $\boldsymbol{Z_{j_K}} = \mathbf{0}$ otherwise. We annotated each site of the

considered subtype for 14 genomic features, generating predictors $\boldsymbol{F_{j_K,1}, \ldots, F_{j_K,14}}$. For 11 binary features (broad histone peaks [H3K4me1, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3], lamin-associated domains, CpG islands, DNase hypersensitive sites, exons), we set the predictor $\boldsymbol{F_{j_K,g} = 1, g \in \{1, \ldots, 11\}}$ if the central site of the motif was inside the specified regions and $\boldsymbol{F_{j_K,g} = 0}$ otherwise. For the 3 continuous features (recombination rate, replication timing, surrounding GC content), we set the predictor $\boldsymbol{F_{j_K,g}, g \in \{12, 13, 14\}}$ to the mean value of that feature in a 10kb window centered at the site. To avoid confounding effects of genomic features like GC content with the effect of read depth, we included read depth at the central site of the 7-mer as covariate $\boldsymbol{F_{j_K,DP}}$. We then evaluated the effect of the genomic predictors on the relative mutation rate of 7-mer subtype $\boldsymbol{K}$ using the following logistic regression equation:

$$\boldsymbol{ln\left(\frac{P(Z_{j_K} = 1)}{P(Z_{j_K} = 0)}\right) = \beta_0^K + \beta_1^K F_{j_K,1} + \cdots + \beta_{14}^K F_{j_K,14} + \beta_{DP}^K F_{j_K,DP}}$$

Where $\boldsymbol{(\beta_1^K, \ldots, \beta_{14}^K)}$ are effects of the 14 considered genomic features on the mutation rate of subtype $\boldsymbol{K}$, and $\boldsymbol{\beta_{DP}^K}$ is the effect of the local sequencing depth. The intercept of this model, $\boldsymbol{\beta_0^K}$, represents the feature-adjusted relative mutation rate for the considered 7-mer subtype. We estimated the beta parameters of each model in R v3.2.3 by fitting a generalized linear model with the speedglm() function from the *speedglm* package, which directly solves the system of regression equations using an iteratively reweighted least squares algorithm. We performed this procedure for each of the $\boldsymbol{K \in \{1, \ldots, 24, 489\}}$ 7-mer subtypes; the resulting beta values and standard errors for 16 x 24,489 estimated parameters are provided in **Supplementary Table 7**.

To generate a map of mutation rates across the genome, we used the estimated regression coefficients $\boldsymbol{(\hat{\beta}_0^K, \hat{\beta}_1^K, \ldots, \hat{\beta}_{14}^K, \hat{\beta}_{DP}^K)}$ to predict the relative mutation rate (i.e., probability of observing a singleton) at each site $\boldsymbol{j}$ where a mutation of a given 7-mer subtype could occur. Because there are three possible mutations at every base, we predict 3 independent mutation probabilities (one for each possible allele). For example, for a site centered at ACGATTG motif, we predict probabilities for A>C, A>G, and A>T alleles, using the parameters estimated from those models. This prediction uses all

estimated effects, not just the effects determined to be statistically significant. We also note that we did not generate predictions for sites within 5 megabases of the start/end of a chromosome, since recombination rate data were not available for these regions[54].

To assess if inclusion of these genomic features improved the prediction of *de novo* mutations over the 7-mers-only model, we again tested this model's ability to predict the known *de novo* mutations from the GoNL[9] and Inova NBS[12] studies. We annotated each of the $\boldsymbol{i} = \{\mathbf{1}, \ldots, \mathbf{1,046,813}\}$ testing sites with the quantity $\boldsymbol{R_{i,L}}$=the difference between 7-mers+features relative mutation rate and 7-mer relative mutation rate at each site, and added this parameter to the 7-mer model. We compared these models to the 7-mer-only models again using a likelihood ratio test (with 1 degree of freedom), AIC, and Nagelkerke's R^2.

Data availability. We are in the process of submitting the BRIDGES sequence-based genotypes to dbGaP. K-mer-based relative mutation rate estimates are provided in **Supplementary Table 2**. Predicted mutation rates based on sequence context and genomic features at each site are available in a compressed wiggle (bigwig) format for use with the UCSC Genome Browser, and can be accessed at http://mutation.sph.umich.edu.

Code availability. The GotCloud mapping and variant calling pipeline can be accessed at http://genome.sph.umich.edu/wiki/GotCloud. All custom scripts used in downstream data processing and analyses are available at https://github.com/carjed/smaug-genetics. A web-based utility and command-line code for annotating genetic variants with estimated 7-mer mutation rates can be accessed at http://www.jedidiahcarlson.com/mr-eel/.

References

1. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15,** 47–70 (2014).
2. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).
3. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
4. Nielsen, R. *et al.* Genomic scans for selective sweeps using SNP data. *Genome Res.* **15,** 1566–1575 (2005).
5. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).
6. Zhang, W., Bouffard, G. G., Wallace, S. S. & Bond, J. P. Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. *J. Mol. Evol.* **65,** 207–214 (2007).
7. 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).
8. Lercher, M. J. & Hurst, L. D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18,** 337–340 (2002).
9. Francioli, L. C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat. Genet.* **47,** 822–826 (2015).
10. 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).
11. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48,** 1–11 (2015).
12. Goldmann, J. M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.* **48,** 935–939 (2016).
13. Panchin, A. Y. *et al.* New words in human mutagenesis. *BMC Bioinformatics* **12,** 268 (2011).
14. Nachman, M. W. & Crowell, S. L. Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304 (2000).
15. Jiang, C. & Zhao, Z. Mutational spectrum in the recent human genome inferred by single nucleotide polymorphisms. *Genomics* **88,** 527–534 (2006).
16. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
17. Williamson, S. H. *et al.* Localizing recent adaptive evolution in the human genome. *PLoS Genet.* **3,** 0901–0915 (2007).
18. Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. & Clark, A. G. Recent and ongoing selection in the human genome. *Nat. Rev. Genet.* **8,** 857–868 (2007).
19. Cai, J. J., Macpherson, J. M., Sella, G. & Petrov, D. A. Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* **5,** e1000336 (2009).
20. Schaibley, V. M. *et al.* The influence of genomic context on mutation patterns in the human genome inferred from rare variants. *Genome Res.* **23,** 1974–1984 (2013).
21. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).
22. Messer, P. W. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics* **182,** 1219–1232 (2009).
23. Slatkin, M. Allele age and a test for selection on rare alleles. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **355,** 1663–8 (2000).
24. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
25. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).
26. Nelson, M. R. *et al.* An Abundance of Rare Functional Variants in 202 Drug Target Genes

Sequenced in 14,002 People. *Science (80-.).* **337,** 100–104 (2012).
27. Michaelson, J. J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–1442 (2012).
28. Meunier, J. & Duret, L. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21,** 984–990 (2004).
29. Deaton, A. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
30. Melton, C., Reuter, J. a, Spacek, D. V & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47,** 710–716 (2015).
31. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532,** 264–267 (2016).
32. Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532,** 259–263 (2016).
33. Mantovani, R. A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res.* **26,** 1135–1143 (1998).
34. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci.* **94,** 1872–1877 (1997).
35. Gasior, S. L., Wakeman, T. P., Xu, B. & Deininger, P. L. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J. Mol. Biol.* **357,** 1383–1393 (2006).
36. 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).
37. Wong, W. S. W. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7,** 10486 (2016).
38. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).
39. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (2014).
40. Scott, L. J. *et al.* Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 7501–6 (2009).
41. Pato, M. T. *et al.* The genomic psychiatry cohort: Partners in discovery. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **162,** 306–312 (2013).
42. Langenecker, S. A., Saunders, E. F. H., Kade, A. M., Ransom, M. T. & McInnis, M. G. Intermediate: Cognitive phenotypes in bipolar disorder. *J. Affect. Disord.* **122,** 285–293 (2010).
43. Sklar, P. *et al.* Whole-genome association study of bipolar disorder. *Mol. Psychiatry* **13,** 558–569 (2008).
44. Miller, M. B. *et al.* The Minnesota Center for Twin and Family Research genome-wide association study. *Twin Res. Hum. Genet.* **15,** 767–74 (2012).
45. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–9 (2008).
46. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25,** 918–925 (2015).
47. McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
48. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).
49. Reppell, M., Boehnke, M. & Zöllner, S. The impact of accelerating faster than exponential population growth on genetic variation. *Genetics* **196,** 819–828 (2014).
50. Gao, F. & Keinan, A. Inference of super-exponential human population growth via efficient computation of the site frequency spectrum for generalized models. *Genetics* **202,** 235–245 (2016).

51. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
52. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274,** 775–780 (1978).
53. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287,** 560–561 (1980).
54. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467,** 1099–1103 (2010).