An expanded sequence context model broadly explains variability in polymorphism levels across the human genome

Varun Aggarwala1 & Benjamin F Voight2,3

The rate of single-nucleotide polymorphism varies substantially across the human genome and fundamentally influences evolution and incidence of genetic disease. Previous studies have only considered the immediately flanking nucleotides around a polymorphic site—the site’s trinucleotide sequence context—to study polymorphism levels across the genome. Moreover, the impact of larger sequence contexts has not been fully clarified, even though context substantially influences rates of polymorphism. Using a new statistical framework and data from the 1000 Genomes Project, we demonstrate that a heptanucleotide context explains >81% of variability in substitution probabilities, highlighting new mutation-promoting motifs at ApT dinucleotide, CAAT and TACG sequences. Our approach also identifies previously undocumented variability in C-to-T substitutions at CpG sites, which is not immediately explained by differential methylation intensity. Using our model, we present informative substitution intolerance scores for genes and a new intolerance score for amino acids, and we demonstrate clinical use of the model in neuropsychiatric diseases.

Measured at the level of the chromosome down to individual base, rates of single-nucleotide substitution vary substantially by position across mammalian genomes, including in humans1. An exquisite example of the role for sequence context in contributing variability in substitution rate is provided by CpG dinucleotides, where spontaneous deamination of 5-methylcytosine results in ~14-fold higher C-to-T substitution rates as compared to the genome-wide average1-3. Modeling the variability in nucleotide substitution rates will inform understanding of evolutionary processes, help identify functional noncoding regions4 and mutation-promoting motifs, suggest mechanisms behind spontaneous mutation and aid in prediction of the clinical impact of polymorphisms discovered through resequencing5. Such models will need to determine not only the optimal window of local sequence context but should also integrate knowledge of functional constraint on the genome due to pressure from purifying selection.

Results
Sequence context modeling of substitution probabilities
We hypothesized that local sequence context—the nucleotides that flank a polymorphic site—could explain the observed variability in nucleotide substitution probabilities. To test this hypothesis, we defined a statistical model (Online Methods and Supplementary Fig. 1) whereby the probability that a nucleotide substitution occurs at a genomic site varies according to (i) the identities of the nucleotides that flank the site and (ii) the size of the 5′-to-3′ local sequence context window. To minimize the impact of natural selection, we focused on intergenic noncoding regions of the genome (Online Methods). As the estimated nucleotide substitution probabilities were robust (Supplementary Table 1a), we developed a likelihood-ratio testing procedure to evaluate competing local sequence context models (Online Methods).

First, we calculated the likelihood of the observed data assuming a ‘1-mer’ model, which allowed different substitution classes (for example, A to G, C to T, etc.) to occur at different rates but ignored the effects of sequence context on substitution probabilities. We compared the 1-mer model to the trinucleotide (‘3-mer’) sequence context model where single 5′ and 3′ nucleotides flanking the polymorphic middle position influence the rate of substitution. As expected, the 3-mer model significantly improved fit to the data (log-likelihood ratio (LLR) = 6,070,948, $P < 1 \times 10^{-600}$; Supplementary Table 1a).

Studies of complex human disease have incorporated a simple trinucleotide sequence context6-7 into models to quantify the probability of de novo mutational events8-10, to clarify the distribution of somatic mutational events segregating in different cancers11 and to model the purifying selective pressure on gene sequences12. As their focus was clinical, these reports did not determine whether this context model best captured the extent to which flanking nucleotides influence the variability in genome-wide nucleotide substitution rates. Here we report a statistical framework that compares the extent to which different local sequence lengths influence the probability of nucleotide substitution, tested using data from the 1000 Genomes Project13; apply our models to the coding genome; and demonstrate use of the model to interpret de novo mutations identified in studies of neuropsychiatric disorders. We define the probability of nucleotide substitution as the chance that a nucleotide in the human genome reference is polymorphic, that is, the nucleotide position segregates alternative nucleotides within the population. This probability depends on population history, selection, sample ascertainment and local context features that influence the rate of mutation.

1 Genomics and Computational Biology Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. 2 Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. 3 Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence should be addressed to B.F.V. (bvoight@upenn.edu).

Received 14 April 2015; accepted 22 January 2016; published online 15 February 2016; doi:10.1038/ng.3511
Nature America, Inc. All rights reserved.

24.4
−1.0
30.0
3,072
3
7.5
48.7
b
4,096
366
−0.8
a
−1.0
15.2
95
44.4
7.1
4,096
1,024
1,024
R
197
266
28.2
81.0
86.8
58.3
−0.6
19.0
17.1
−1.2
58.7
1,024
1.0
−1.8
−1.4
−1.0
−0.8
−0.6
−1.8
−1.6
−1.4
−1.2
−1.0
−0.8
−0.6

Next, we evaluated whether the inclusion of additional local nucleotides could further improve fit to the observed data. We demonstrate that, when compared to the 3-mer model or the pentanucleotide (‘5-mer’) model (with two flanking nucleotides on each side), the larger, heptanucleotide (‘7-mer’) model (with three flanking nucleotides on each side) fit the data better (both LLR >494,212, P < < 1 x 10−100; Supplementary Table 1b). To further validate the models, we estimated substitution probabilities using 1,659,929 HapMap14 variants found in our noncoding regions (Online Methods) and observed that 7-mer context probabilities strongly correlated with probabilities estimated from 1000 Genomes Project data (Supplementary Fig. 2 and Supplementary Table 2) and provided the best fit to the observed polymorphisms (Supplementary Table 3). Our model recapitulates expected shifts in probabilities consistent with population histories15 (Supplementary Fig. 3) and the downward shift in the average substitution probability for the X chromosome16 relative to the autosomes (Supplementary Table 4) due to the smaller effective population size at the X chromosome. Taken collectively, our analyses demonstrate for the first time, to our knowledge, that a 7-mer sequence context model explains the observed distribution of polymorphisms found in human populations.

To incorporate prior information, we developed a Bayesian formulation using objective conjugate priors for analysis of the noncoding genome (Online Methods). Consistent with our previous analysis, the 7-mer context model proved superior in comparison to all other models (approximate Bayes factor (ABF) >> 1,000; Supplementary Table 1c). In subsequent analyses, we use these posteriors for the nucleotide substitution probabilities.

7-mer context predicts noncoding substitution rates
To quantify the variance in the posterior probabilities that a 7-mer sequence context model could explain, we considered each substitution class separately, as well as CpG site contexts (nine classes in total). We employed forward regression (Online Methods) to select features from a 7-mer context window to predict substitution probabilities and considered up to four-way interactions at positions within the window. When compared to single-base and position models without interactions, incorporating higher-order interactions substantially improved the fit to data (Supplementary Table 5). Specifically, we found that our selected models in a separately held test data set explained a median of 81% of the variability (as compared to 30% explained by the 3-mer context) in probabilities across all substitution classes, covering 84% of all mutational events and fitting well the probability of C-to-T substitution at CpGs (Fig. 1a and Table 1). Although we identified a common set of interactions across classes (Supplementary Table 6), many common features did not always influence substitution probabilities in the same way and others had class-specific effects. These observations indicate that core and class-specific features based on sequence context are predictive of the potential for nucleotide substitution.

Methylation cannot fully explain patterns at CpG sites
The spontaneous deamination of 5-methylcytosine at CpG sites results in ~14-fold higher rates of C-to-T substitutions generally3,17. Although a previous report indicated that divergence at CpG sites varies as a function of local context, the focus was on introns and did not consider population-level polymorphisms in humans18. Thus, we hypothesized that the surrounding sequence context further influences the probability of nucleotide substitution at CpGs and examined the C-to-T substitution class within the subset of contexts that contained a CpG at positions 4 and 5 in the 7-mer. Simulations using a model that ignored additional genomic context, or considered the 3-mer context (Supplementary Fig. 4), using a fixed CpG substitution probability generated significantly less variability in 7-mer CpG substitution probabilities than was empirically observed (empirical P < < 1 x 10−100; Fig. 1a). These data indicate that (i) not all CpG sites accrue substitutions at the same rate and (ii) the sequence context surrounding CpG sites correlates with biological features or mechanisms that influence this rate.

To explore the possibility that the excess variability depends on variation in methylation intensity across sequence contexts, we reanalyzed whole-genome bisulfite sequencing data obtained from germline and other tissues of healthy individuals19,20. Comparing the CpG sites that are consistently methylated with those that are consistently unmethylated across subjects, we observed as expected that methylation correlates with an increase in the probability of C-to-T substitution (P < < 1 x 10−100; Supplementary Fig. 5). Unexpectedly, when we compared the methylation intensity in sperm at 7-mer CpG contexts with the probability of substitutions, we found a positive but imperfect correlation (R² = 0.33, P < 1 x 10−04; Fig. 1b), with similar results in other tissues (Supplementary Fig. 6), noting instances of methylation status decoupled from substitution probabilities. For example, nearly every genomic instance of the sequence contexts

Table 1 Summary and performance of forward regression model for feature selection using the 7-mer context in the intergenic noncoding genome

| Substitution class | Contexts (n) | Substitutions (%) | Parameters (n) | Model R² (7-mer) | Model R² (3-mer) |
|--------------------|-------------|-------------------|---------------|-----------------|-----------------|
| Outside CpG dinucleotide context | A to C | 4,096 | 7.3 | 266 | 56.5 | 11.2 |
| | A to G | 4,096 | 28.2 | 366 | 91.5 | 40.9 |
| | A to T | 4,096 | 7.1 | 197 | 58.7 | 37.4 |
| | C to A | 3,072 | 8.5 | 282 | 83.5 | 30.0 |
| | C to G | 3,072 | 7.5 | 268 | 81.0 | 17.1 |
| | C to T | 3,072 | 24.4 | 254 | 86.8 | 37.6 |
| Within CpG dinucleotide context | C to A | 1,024 | 1.0 | 26 | 58.3 | 19.0 |
| | C to G | 1,024 | 0.8 | 95 | 48.7 | 9.5 |
| | C to T | 1,024 | 15.2 | 96 | 93.1 | 44.4 |

*The percentage of substitutions for that class observed in the genome. *2The number of features selected in the best 7-mer model. *3Prediction accuracy in the test data set alone (not used for model training) with the best model using the 7-mer sequence context features. *4Prediction accuracy with only the 3-mer sequence context features.
GTACGCA and GATCGCGA showed consistent methylation signals (both methylated in >94% of occurrences in sperm); the probability of C-to-T transition was more than twofold different for these two contexts (0.148 versus 0.07, respectively). These data are consistent with the hypothesis that local context features beyond DNA methylation influence probabilities of C-to-T transitions at CpG sites, although we cannot exclude the possibility that sub-tissue methylation differences could explain these patterns.

**Identification of new mutation-promoting motifs**

We next investigated the substitution probabilities for 7-mer contexts partitioned by the substitution class (Fig. 2 and Supplementary Table 7). First, we noted that several classes (C-to-A and C-to-G in addition to C-to-T changes) appeared to segregate as mixtures of two distributions, explainable by CpG effects. These observations are consistent with studies demonstrating elevated rates of substitution at CpGs in humans, although this early work was not powered to measure context dependencies surrounding CpGs as we are here. As the methylation transition state intermediate 5-formylcytosine can induce spontaneous C-to-A or C-to-G substitutions, one possibility is that methylation also elevates these rates in this context. We next determined whether local sequence context motifs—analogs to but beyond CpG dinucleotides—correlate with variable substitution probabilities across classes (Online Methods). We noted that poly(CG) sequences in the lower tail of C-to-T substitutions for the CpG context were enriched ($P < 1 \times 10^{-16}$; Table 2). This observation is consistent with previous reports, as this context is found proximal to genes (Supplementary Fig. 7) and is associated with lower methylation intensities (Supplementary Fig. 8). In the upper tail of the A-to-T substitution class, we observed a poly(T) + poly(A) motif in the outlier sequences ($P < 1 \times 10^{-5}$; Table 2). We also observed a similar A$_3$ motif in the lower tail of the A-to-G class ($P < 1 \times 10^{-10}$). One possible mechanism that might contribute is the ‘slippage’ of protein machinery during DNA replication. Our analysis also identified motifs without an obvious contributing mechanism. First, in the upper tail of CpG rates, we observed enrichment of a TACGG motif ($P < 1 \times 10^{-10}$; Table 2) that was strongly methylated (Supplementary Fig. 8), but, curiously, a similar motif shifted by one position was enriched in the lower tail of the A-to-C class ($P < 1 \times 10^{-4}$). Second, the ApT dinucleotide was found to correlate with the substitution probabilities (Fig. 2) for the A-to-G ($P < 1 \times 10^{-25}$) and A-to-T ($P < 1 \times 10^{-17}$) classes, although not statistically significantly so for the A-to-C class. Finally, we observed a CAAT motif also enriched in the upper tail of the A-to-G substitution class ($P < 1 \times 10^{-53}$), reported in an earlier study of dbSNP variants. These latter cases indicate potentially new mechanisms contributing to elevated nucleotide substitutability, not documented by the commonly used trinucleotide context model. As a final analysis of robustness, keeping in mind limitations due to variant ascertainment, we estimated the substitution probabilities using HapMap variants and found similar mutation-promoting motifs across substitution classes (Supplementary Table 8).

**Experiments to validate the noncoding rate model**

If the estimated noncoding substitution probabilities reflect properties of mutation, one would expect that these rates should (i) be influenced by rates of recombination, (ii) strongly correlate with rates of species divergence, (iii) be consistent for both rare and common genetic variants, and (iv) also be reflected in de novo mutational events. We explored each of these predictions in turn. First, we estimated the 7-mer substitution rates from all intergenic noncoding variants separately for regions with high and low recombination rates and found a strong correlation between the two ($R^2 = 0.97, P < 1 \times 10^{-100}$; Online Methods and Supplementary Fig. 9), indicating that substitution probabilities estimated from the noncoding genome are correlated across high and low rates of recombination. Next, using human-chimpanzee and human-macaque alignments over intergenic noncoding sequences, we found a strong correlation between divergence and substitution probabilities for our 7-mer contexts (both $R^2 = 0.96, P < 1 \times 10^{-100}$; Online Methods, Supplementary Fig. 10 and Supplementary Table 9).

We then estimated 7-mer probabilities from all intergenic noncoding rare variants (singletons and doubletons) separately from low- and high-frequency (>1%) variants and found a strong correlation ($R^2 = 0.98, P < 1 \times 10^{-100}$; Online Methods and Supplementary Fig. 11), as well as a superior 7-mer context fit to data across variant frequencies (Supplementary Table 10). Finally, we obtained 4,748 de novo mutational events from a high-quality pedigree sequencing data set on 78 parent-offspring trios. We tested for the presence of the motifs we identified in Table 2 around de novo events and observed a significant enrichment (Online Methods and Supplementary Table 11). Taken collectively, these findings provide additional validation for the hypothesis that our substitution probabilities capture features of germline mutation.

**7-mer context also predicts exonic substitution rates**

Assuming that the processes that generate spontaneous mutations apply uniformly across the genome, we hypothesized that sequence context could explain variability in substitution probabilities in the coding genome. We therefore extended our initial framework (Online Methods and Supplementary Fig. 1) to the coding genome.
Table 2 Enrichment of motifs identified in posterior nucleotide substitution probabilities for the 7-mer sequence context models inferred from the intergenic noncoding genome

| Motif                  | Substitution class | Effect on substitution probability | Enrichment P value | OR (95% CI) | Fold change in substitution rate
|------------------------|--------------------|-----------------------------------|--------------------|-------------|-------------------------------|
| NNNCGNN                | C to T             | Higher                            | 2 × 10⁻²⁶          | 134.4 (18.4–977.4) | 13.9             |
| A to C                 | Higher             | 1.9                               | 12.8 (5.9–27.7)    | 2.4         |
| C to A                 | Higher             | 1.7                               | 9 × 10⁻¹²          | 60.8 (14.6–252.1) | 2.7               |
| C to T (CpG+)          | Lower              | 1.6                               | 9 × 10⁻¹⁰          | 366.3 (45.6–2,939.5) | 1.5               |
| Poly(T) and poly(A) combination (AAAATT, TTTAAA) | A to T | Higher | 4 × 10⁻³⁰ | 304.2 (31.0–2,987.6) | 12.7           |
| Aₜ (AAAANN, NAAANN, NNAAN, NNNAA) | A to T | Higher | 5 × 10⁻¹⁰ | 10.2 (7.3–14.1) | 1.9              |
| NTACG[C/A/G]A/C/G       | C to T (CpG+)      | Higher                            | 3 × 10⁻¹⁰          | 102.5 (27.4–383.2) | 1.7               |
| NNTACGN                | A to G             | Lower                             | 3 × 10⁻¹³          | 9.4 (3.6–24.8)    | 1.5               |
| NNNATNN                | A to T             | Higher                            | 2 × 10⁻¹⁰          | 22.3 (8.7–57.1)   | 1.6               |
| [C/T]CAAT[C/G/T]N      | A to G             | Higher                            | 1 × 10⁻⁵           | 131.2 (18.0–954.2) | 2.0               |
|                        | A to G             | Higher                            | 8 × 10⁻⁵⁵          | 5,966 (2,091–17,021) | 5.1              |

In the motifs, the polymorphic nucleotide is underlined.

aThe distribution of sequence contexts that include a CpG site (where the polymorphic site at position 4 is C and position 5 is fixed as G). bBased on the enrichment of the motif in the 1% tail of the given substitution class. “Higher” implies enrichment in the upper 1% tail of the sequence context probability distribution, and “Lower” implies enrichment in the lower 1% tail of this distribution. cThe odds ratio (and 95% confidence interval) of enrichment of the motif in the upper or lower 1% tail of the sequence context probability distribution. dThe fold increase or decrease in substitution rates for the motif relative to its substitution class.

by (i) using information obtained from our model on the noncoding genome as the prior and (ii) allowing for context dependence of codons and local sequence context in our estimates of substitution probabilities to accommodate purifying selective pressure. Our new model substantially improved the fit to the data as compared to the 3-mer sequence context models with or without codon context (ABF > 1,000; Supplementary Table 12). To further validate, we tested our model on a different large-scale exome sequencing data set from ~4,300 individuals and noted that our 7-mer model fit patterns of exonic polymorphisms better than competing models (ABF > 1,000; Online Methods and Supplementary Table 12). These results demonstrate, for the first time to our knowledge, that a broader sequence context—beyond simple codon or trinucleotide context—captures the forces that shape variability in nucleotide substitutions in the coding genome.

We then examined the posterior distribution of substitution probabilities for all contexts stratified by the type of amino acid substitution (Supplementary Fig. 12 and Supplementary Table 13) and found excess variability in each class over that expected under simulation (Online Methods and Supplementary Table 14). Next, we estimated the substitution probability profiles for each amino acid change and found certain nonsense and missense substitution probabilities to be higher than synonymous levels (Supplementary Fig. 13), partially explained by CpG contexts. These observations caution against the practice—invoked in rare variant association tests—of ignoring codon and sequence context when testing for the burden of functional substitutions. Our results here demonstrate that functional substitutions may not be equally likely or tolerated with respect to purifying selection.

**7-mer context improves power to detect pathogenic variants**

We now turn to applications of our model to improve the interpretation of variation discovered by clinical resequencing. Efforts to prioritize variants from such studies often rely on classifying variants that are deleterious with respect to population genetic fitness, hypothesizing that such variants are more likely to be pathogenic. As our coding substitution probabilities are influenced by forces both of mutation (estimated from the noncoding genome) and selection, we hypothesized that the ratio of these probabilities quantifies the action of selective pressure and could be used to prioritize pathogenic variants. To test this hypothesis, we calculated the log ratio of intergenic noncoding and coding substitution probabilities, defined as the sequence constraint score, for missense (n = 48,450) and nonsense (n = 12,054) variants present in the Human Gene Mutation Database (HGMD; Online Methods). We observed that the distribution of sequence constraint scores for HGMD variants was shifted toward larger values (intolerance) as compared to 1000 Genomes Project variants (P < 1 × 10⁻¹⁰⁰; Fig. 3a), compatible with the ‘intolerant variant, pathogenic variant’ hypothesis. Moreover, the distribution of scores based on our 7-mer model was further shifted toward intolerance with a thicker tail, as compared to a 3-mer model (P < 1 × 10⁻¹⁰⁰, Supplementary Fig. 14). These data demonstrate that a coding model...
that includes the codon and 7-mer context improves the identification of variants that are potentially pathogenic.

Describing genic intolerance to mutation via 7-mer context

Several groups have argued that the power to identify causal disease genes from clinical resequencing data could be enhanced by incorporating estimates of selective constraint on genes. The underlying hypothesis behind this concept is that genes that are under selective constraint are more likely to show functional consequences when altered and are therefore most likely to be pathogenic and have fewer functional variants (‘intolerant gene, pathogenic gene’). The community has successfully applied this concept to neurodevelopmental and psychiatric disorders; however, the existing approaches have not incorporated the 7-mer sequence or codon context in their models.

Therefore, we applied our 7-mer coding substitution probabilities to develop an intolerance score (Online Methods and Supplementary Table 15) quantifying the difference between the expected and observed numbers of functional variants at a gene, with higher scores consistent with functional constraint. To further validate, we found gene scores on a separate, larger exome sequencing data set and observed a strong correlation between the two (Supplementary Fig. 15). We found that genes belonging to putatively essential or ubiquitously expressed categories scored strongly for genic intolerance (P < 1 × 10^-4). In contrast, gene sets representing keratin and olfactory categories were found to be highly tolerant of functional changes (Fig. 3b). Next, we applied this to Online Mendelian Inheritance in Man (OMIM) genes or the known genes behind several neuropsychiatric disorders such as autism, epilepsy, developmental disorder and intellectual disability and found these genes to have significantly higher intolerance scores (P < 1 × 10^-4). We then compared our gene scores to previously reported scores (Online Methods and Supplementary Fig. 16) and found that our approach improved classification or performed comparably to other approaches for genes in each set, including the disease categories (Supplementary Table 16). These results demonstrate that the most accurate scoring of genic tolerance to functional substitution can be achieved by modeling 7-mer sequence and coding context.

An amino acid score for pathogenic variant prioritization

Beyond the average rate of amino acid replacement that a gene might tolerate, genes could be further intolerant to specific types of amino acid substitutions, signifying added localized selective constraint or importance for gene functionality. Therefore, we developed a score measuring the intolerance at the level of amino acid replacement for a gene (Online Methods and Supplementary Table 17), after quantifying the difference between the expected and observed numbers of functional variants for a specific amino acid at a gene. Across all genes represented in HGMD with a large number of putatively pathogenic amino acid changes for a specific substitution, we found that these genes segregate larger intolerance scores for that amino acid (empirical P < 1 × 10^-10). Moreover, a gene might score ‘tolerant’ for functional substitution but ‘intolerant’ for specific amino acid changes. For example, VWF (encoding von Willebrand factor), a blood glycoprotein involved in hemostasis, is tolerant to substitution overall (within the top 8% of gene tolerance) but intolerant to cysteine substitution (within the top 3.5% of cysteine intolerance). These data are consistent with a causal mechanism for von Willebrand disease: protein misfolding when cysteine residues are substituted. We note that 5,652 genes segregate a profile similar to that of VWF, showing average genic tolerance but amino acid intolerance.

Interpretation of de novo mutations discovered in autism

Autism spectrum disorder is a disease with complex etiology, and recent efforts have aimed to identify de novo mutational events that may contribute to disease. To highlight the usefulness of gene and amino acid scores, we applied them to interpret the de novo mutations collected from 2,508 autism spectrum disorder cases and 1,911 control family trios. First, we found that the most intolerant genes based on our gene score segregated a significant burden of de novo mutations in cases as opposed to controls (odds ratio (OR) = 1.66, P < 0.0001; Fig. 4a and Online Methods), even after removing known autism-associated genes (OR = 1.54, P < 0.001), and similar although slightly attenuated burden using other scores (Fig. 4a). Next, we found that the average amino acid score for de novo mutations at autism-associated genes in cases was higher (more intolerant) than that found in controls or at other genes in cases (P = 0.002; Fig. 4b and Online Methods). We further observed higher (intolerant) average amino acid scores for variants in genes with a positive variant burden in cases relative to controls (+2 or +3 allele count excess in cases; both P < 0.01; Fig. 4b). Finally, several genes from the excess allele count set stood out with amino acid–specific intolerance (all within the top four percentiles of intolerance): MYO9B, WDFY3, NAV2, STIL and SCUBE2. Aside from WDFY3, these genes are generally tolerant, on
the basis of their gene score, indicating the usefulness of subgene-wise measurement of functional intolerance. While MYO9B has been implicated in autism and Wdfy3 deletions in a mouse model have been shown to cause autism-like symptoms, our analysis points to the remaining candidates for future follow-up.

**DISCUSSION**

We report a sequence context model that explains patterns of nucleotide substitution observed in the human genome. Our motivation was based on the need to statistically evaluate competing models for sequence context. We demonstrate that the commonly used context that includes one nucleotide flanking each side of a polymorphic site does not fully capture the complete spectrum of where, what type and how frequently nucleotides are expected to change. Furthermore, by using population-level data rather than data on de novo or somatic events, we were able to improve the resolution of substitution models and identify new mutation-promoting motifs. Our approach also characterized average selective pressures operating in the coding genome at a finer level of detail. Our model indicates substantial variability across all amino acid replacement classes and, in some cases, synonymous substitutions that were less prone to change than missense or even nonsense substitutions. We suggest that inference of the presence and strength of selection on genes might further benefit from incorporating information at this resolution.

One question in the field has been how much sequence context can explain patterns of nucleotide substitution in genomes. Our results suggest that a substantial fraction of variability can be robustly predicted by sequence context alone, although specific substitution classes may require more than sequence context for their prediction. In evolutionary genetics studies, the set of substitutions that occurs at nearly constant rates proportional to the lineage (the most ‘clock-like’) is important for accurate dating of divergence events. Although we did not apply our model to other species, the strong correlation with divergence suggests that our features are potentially conserved across primates.

We acknowledge that a number of features remain to be formally evaluated in the genome, for example, recombination in the coding genome or replication timing. Our framework has the flexibility to model the complexity found in any sequences that contain features hypothesized to be important. We also acknowledge that context models including more than three flanking nucleotides on each side of a polymorphism were not considered. The regression approach we have presented does suggest that the 7-mer models could be refined, perhaps allowing a broader context to be considered.

With an appropriate background model for nucleotide substitution, new statistics for clinical resequencing studies can be envisioned, based on the occurrence of discovered variation. Such approaches may complement statistics that assay allele frequency differences between cases and controls at one or more polymorphic sites. Moreover, comparative genomics applications to identify non-neutrally evolving regions, genome alignments or tree reconstruction would benefit from accurate models of nucleotide substitution. Although the underlying mechanisms that determine how nucleotide sequences change over time remain to be addressed, we posit that the features identified from our model provide important clues in elucidating these fundamental principles.

**URLs** Exome Variant Server (EVS), http://evs.gs.washington.edu/EVS/

**METHODS**

Methods and any associated references are available in the online version of the paper.

**Note:** Any Supplementary Information and Source Data files are available in the online version of the paper.

**ACKNOWLEDGMENTS**

We thank C. Brown, M. Bucan, P. Babb, K. Siewert, K. Johnson, S. Bumgarner and two anonymous reviewers for helpful comments on the manuscript. B.F.V. is grateful for support of the work from the Alfred P. Sloan Foundation (BR2012-087), the American Heart Association (13SDG1433006), the W.W. Smith Charitable Trust (H1201) and the US National Institutes of Health/National Institute of Diabetes and Digestive and Kidney Disorders (R01DK101478).

**AUTHOR CONTRIBUTIONS**

V.A. and B.F.V. conceived and designed the experiments, developed the model, performed the statistical analysis, developed and contributed analysis tools, and wrote the manuscript. B.F.V. supervised the research.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. Nat. Rev. Genet. 12, 756–766 (2011).

2. Ehrlich, M. & Wang, R.Y. 5-methylcytosine in eukaryotic DNA. Science 212, 1350–1357 (1981).

3. Riedelt, W.M. III, Coetzeet, G.A., Olumi, A.F. & Jones, P.A. 5-methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. Science 249, 1288–1290 (1990).

4. Arbiza, L. et al. Genome-wide inference of natural selection on human transcription factor binding sites. Nat. Genet. 45, 723–729 (2013).

5. Yang, Y. et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N. Engl. J. Med. 369, 1502–1511 (2013).

6. Hwang, D.G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA 101, 13994–14001 (2004).

7. Blake, R.D., Hess, S.T. & Nicholson-Tuell, J. The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. J. Mol. Evol. 34, 189–200 (1992).

8. Neale, B.M. et al. Patterns and rates of de novo mutations in autism spectrum disorders. Nature 485, 242–245 (2012).

9. Michaelson, J.J. et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151, 1431–1442 (2012).

10. Fromer, M. et al. De novo mutations in schizophrenia impact synaptic networks. Nature 506, 179–184 (2014).

11. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–218 (2013).

12. Samocha, K.E. et al. A framework for the interpretation of de novo mutation in human disease. Nat. Genet. 46, 944–950 (2014).

13. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 66–69 (2012).

14. International HapMap Consortium. A haplotype map of the human genome. Nature 437, 1299–1320 (2005).

15. Campbell, M.C. & Tishkoff, S.A. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu. Rev. Genomics Hum. Genet. 9, 403–433 (2008).

16. Schaffner, S.F. The X chromosome in population genetics. Nat. Rev. Genet. 5, 43–51 (2004).

17. Nachman, M.W. & Crowell, S.L. Estimate of the mutation rate per nucleotide in humans. Genetics 156, 297–304 (2000).

18. Mugal, C.F. & Ellegren, H. Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. Genome Biol. 12, R58 (2011).

19. Okae, H. et al. Genome-wide analysis of DNA methylation dynamics during early human development. PLoS Genet. 10, e1004868 (2014).

20. Hovestedt, V. et al. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature 510, 537–541 (2014).

21. Wang, L.C. & Furihata, A.Y. The mutational spectrum of non-CpG DNA varies with CpG content. Genome Res. 20, 875–882 (2010).

22. Kamiya, H. et al. Mutagenicity of 5-formylcytosine, an oxidation product of 5-methylcytosine, in DNA in mammalian cells. J. Biochem. 132, 551–555 (2002).

23. Deaton, A.M. & Bird, A. CpG islands and the regulation of transcription. Genes Dev. 25, 1010–1022 (2011).

24. Levinson, G. & Gutman, G.A. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203–221 (1987).

25. Panchin, A.Y., Mitrofanov, S.I., Alexeevski, A.V., Spirin, S.A. & Panchin, Y.V. New words in human mutagenesis. BMC Bioinformatics 12, 268 (2011).

26. Lanfear, R., Welch, J.J. & Bromham, L. Watching the clock: studying variation in rates of molecular evolution between species. Trends Ecol. Evol. 25, 495–503 (2010).

27. Kong, A. et al. Rate of de novo mutations and the importance of father’s age to disease risk. Nature 488, 471–475 (2012).
28. Bustamante, C.D. et al. Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).
29. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
30. Cooper, G.M. & Shendure, J. Needle in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
31. Stenson, P.D. et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9 (2014).
32. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
33. Georgi, B., Voight, B.F. & Bucan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* **9**, e1003484 (2013).
34. Uddin, M. et al. Brain-expressed exons under purifying selection are enriched for de novo mutations in autism spectrum disorder. *Nature* **515**, 216–221 (2014).
35. De Rubeis, S. et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **519**, 209–215 (2015).
36. Epi4K Consortium & Epilepsy Phenome/Genome Project. De novo mutations in epileptic encephalopathies. *Nature* **501**, 217–221 (2013).
37. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
38. Hamdan, F.F. et al. De novo mutations in moderate or severe intellectual disability. *PLoS Genet.* **10**, e1004772 (2014).
39. Rauch, A. et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380**, 1674–1682 (2012).
40. de Ligt, J. et al. Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367**, 1921–1929 (2012).
41. Ginsburg, D. & Bowie, E.J. Molecular genetics of von Willebrand disease. *Blood* **79**, 2507–2519 (1992).
42. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
43. Orosco, L.A. et al. Loss of Wdfy3 in mice alters cerebral cortical neurogenesis reflecting aspects of the autism pathology. *Nat. Commun.* **5**, 4692 (2014).
44. Eyre-Walker, A. & Eyre-Walker, Y.C. How much of the variation in the mutation rate along the human genome can be explained? *G3 (Bethesda)* **4**, 1667–1670 (2014).
45. Kimura, M. & Ohta, T. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA* **71**, 2848–2852 (1974).
46. 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).
47. Hussin, J.G. et al. Recombination affects accumulation of damaging and disease-associated mutations in human populations. *Nat. Genet.* **47**, 400–404 (2015).
48. Koren, A. et al. Genetic variation in human DNA replication timing. *Cell* **159**, 1019–1026 (2014).
49. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468–488 (2004).
ONLINE METHODS

Sourcing population samples. Samples were obtained from phase 1 of the 1000 Genomes Project. We considered only the variants from individuals of African, European and East Asian ancestry.

Selection of intergenic noncoding sequences. Intergenic sequences were defined as the full set of genomic sequences that are not annotated in Ensembl Biomart (version 75) and RefSeq Genes. We then removed centromeres, telomeres, repetitive regions and sequences not present in the accessibility mask (version 20120824) filter of the 1000 Genomes Project. Within these intergenic regions, we identified variants for the three populations for use in downstream analysis. More details are provided in the Supplementary Note.

Statistical framework to model substitution probabilities for intergenic noncoding regions. We initially describe a simple model that does not take into account local sequence context and then build upon this by incorporating additional local sequence contexts.

Suppose that we observe $n_C$ occurrences of nucleotide C in the reference genome. A subset of these $n_C$ sites will be polymorphic within the population of individuals. Let $n_{CA}$ represent the number of sites where a C-to-A nucleotide change has occurred. Similarly, $n_{CG}$ is the number of sites where a C-to-G change has occurred and $n_{CT}$ is the number of sites where a C-to-T change has occurred. Then, the probability of nucleotide substitution or polymorphism within the population can be described using a multinomial distribution

\[
\frac{n_C!}{(n_C - n_{CA} - n_{CG} - n_{CT})!n_{CA}!n_{CG}!n_{CT}!} \alpha_{CA}^{n_{CA}}\alpha_{CG}^{n_{CG}}\alpha_{CT}^{n_{CT}}(1 - \alpha_{CA} - \alpha_{CG} - \alpha_{CT})^{(n_C - n_{CA} - n_{CG} - n_{CT})}
\]

where the probabilities of observing a C-to-A, C-to-G or C-to-T substitution are expressed as $\alpha_{CA}$, $\alpha_{CG}$ and $\alpha_{CT}$, respectively. After iterating over all possible substitutions (A to C, A to G, A to T, C to A, C to G, C to T, T to A, T to G, T to C, G to A, G to C and G to T), we merge the reverse-complementary pairs (for example, A to C was merged with T to G, etc.) to yield six ‘substitution classes’ as parameters for the simple model, which we refer to as the 1-mer model. We then use maximum-likelihood estimation (MLE) to find the substitution probability estimates for all possible substitutions.

This model can be naturally extended to consider the effects of local sequence context by replacing the count of $n_C$ occurrences of nucleotide X with the count of the occurrences of a particular nucleotide sequence context. For example, if we want to consider the local sequence context ACA, then we count the number of sites where a C-to-A change has occurred at the middle position C, thereby estimating a total of 1,536 parameters for the 5-mer model and the 7-mer model by considering additional fixed nucleotides (two and three, respectively) on either side of the polymorphic site, thereby estimating a total of 1,536 parameters for the 5-mer model and 24,576 parameters for the 7-mer model. More details are provided in the Supplementary Note.

Incorporating prior information into the statistical framework. Because the likelihood of our framework is based on a multinomial distribution, we use its conjugate prior, that is, the dirichlet distribution, for different sequence context models. For inference in the intergenic noncoding genome, we select the objective version of the prior for our analysis, with all concentration parameters of the dirichlet prior set as 1. We then use the maximum a posteriori (MAP) estimates to find the substitution probability estimates for all possible substitutions. More details are provided in the Supplementary Note.

Bayes factor analysis for model comparison. We calculated the approximate posterior likelihood, using Chib's method, on the overall data using the MAP estimates of the substitution probabilities for a specific sequence context model found before. We then calculate the ABF as

\[
\frac{\text{Posterior likelihood under model}_2}{\text{Posterior likelihood under model}_1} = \frac{\text{Prob}(\text{data | context}_2) \times \text{Prob}(\text{context}_2)}{\text{Prob}(\text{data | context}_1) \times \text{Prob}(\text{context}_1)}
\]

where $S_1$ and $S_2$ represent parameters estimated from two competing sequence context models. We use the Jefferey's scale for interpreting the ABFs, where the ratio if greater than 100 is considered to be decisive evidence against model1. More details are provided in the Supplementary Note.

Regression modeling and feature selection. We considered each substitution class separately and created an additional substitution class for each of the three possible changes within a CpG context, resulting in nine substitution classes. For each substitution class, we consider the initial regression model

\[
\Pr[X_1 \rightarrow X_2 | S] = \alpha + \beta_1 P_1^C + \beta_2 P_1^G + \beta_3 P_1^T + \ldots + \beta_n P_n^T + \epsilon
\]

where the probability that a nucleotide changes from $X_1$ to $X_2$ is modeled using a position-based variable $p$, a set of bases (for example, $C$, $G$ or $T$) where A is the reference base) denoted by the superscript for $p$, each position ($= 1, 2, 3, 5, 6$ or 7) denoted by the subscript for $p$ within sequence context $S$, intercept $\alpha$ and error term $\epsilon$. We assigned A as the reference nucleotide at each position and encoded the single nucleotide present at each position as the combination of three thermometer variables (for example, 0,0,0 = A; 0,0,1 = C; 0,1,0 = G; 1,0,0 = T). Next, we examined non-additivity (interactions) between nucleotides at sequence context positions. Rather than including all possible interaction terms, we employed feature selection (model training and testing to select the most informative features) and incorporated these terms into the

\[
-2\ln\left(L_1[\text{data | context } S_1]\right) + 2\ln\left(L_1[\text{data | context } S_2]\right)
\]

where $S_1$ and $S_2$ represent parameters estimated from two competing sequence context models. The test is $\chi^2$ distributed, with the degrees of freedom equal to the difference in the number of parameters between the two models (for example, comparing the 3-mer model with the 1-mer model requires 90 degrees of freedom; comparing the 7-mer model with the 3-mer model requires 24,480 degrees of freedom).

In Supplemental Note 8, we provide additional details on the statistical framework to model substitution probabilities for intergenic noncoding regions. We also provide details on the selection of intergenic noncoding sequences and the Sourcing population samples. Finally, we describe the Statistical framework to model substitution probabilities for intergenic noncoding regions in more detail.
We obtained human-chimpanzee and human-pan data.

We defined rare variants as those occurring fewer than two times in the population and low- and high-frequency variants as those with minor allele frequency (MAF) >1%. We only considered variants with a frequency greater than 0.03%. More details are provided in the Supplementary Note.

We selected the exonic coordinates of the 1000 Genomes Project and EVS data, we manually annotated the type of codon change caused by each variant specific to the transcript.

We obtained CpG methylation data for our intergenic regions of interest from whole-genome bisulfite sequencing studies performed on germline (sperm or oocyte), blastocyst, blood and brain tissues. We performed our analysis on the 7,059,740 intergenic CpG sites that were methylated and the 651,479 intergenic CpG sites that were unmethylated across all samples for a tissue by calculating the mean intensity.

Sequence motif identification. We examined the top and bottom ten sequences for each substitution class and manually identified a total of six sequence-motif classes. More details are provided in the Supplementary Note.

Scaling the substitution probability estimates for a larger sample. To calibrate our model (built using the 1000 Genomes Project data set) for use with the larger EVS data set, we rescaled the substitution probabilities estimated using the 1000 Genomes Project data to make them proportional to the EVS data set. We used a constant scaling factor defined as

\[ \text{Overall substitution probability in the new data set} = \frac{\text{Overall substitution probability in the 1000 Genomes data set}}{\text{Overall substitution probability in the 1000 Genomes data set on all substitution probabilities in the new data set}} \] (7)

Overall substitution probability in the 1000 Genomes data set

Simulating variability in substitution probabilities within amino acid replacement classes. Variant across the frequency spectrum. We defined rare variants as those occurring fewer than two times in the population and low- and high-frequency variants as those with minor allele frequency (MAF) >1%. We only considered the intergenic noncoding variants present in the 1000 Genomes Project.

De novo mutations. We only considered the de novo mutations occurring in the accessible regions of the 1000 Genomes Project. For each motif class, we found the expected number of mutations under a normalized 1-mer sequence context model. More details are provided in the Supplementary Note.

Selection of coding sequences. We selected the exonic coordinates of the longest transcript for each gene annotated in Ensembl Biomart (version 75). We only considered transcripts where (i) the total exonic region length was a multiple of three and (ii) 90% or more of this length was present in the combined accessibility mask (version 20120824) filter of the 1000 Genomes Project. This yielded 16,386 autosomal transcripts and 679 transcripts from the X chromosome.

To test our model in a different data set, SNP sites for ~4,300 individuals of European ancestry were obtained from the Exome Variant Server (EVS; downloaded on 26 August 2013). For EVS data, to obtain a representative spectrum of allele frequencies (and the impact of background selection) observed from the smaller set of individuals found in the 1000 Genomes Project data, we only considered variants with a frequency greater than 0.03%. More details are provided in the Supplementary Note.

Annotation of SNP variants in the autosomal coding genome. For both 1000 Genomes Project and EVS data, we manually annotated the type of codon change caused by each variant specific to the transcript.

Recombination and substitution rates. We obtained a recombination rate map of the YRI (Yoruba) population from the phase 1 release of the 1000 Genomes Project and segregated our intergenic noncoding regions of interest into ones with high (>3 cM/Mb) and low (<0.05 cM/Mb) recombination rates. More details are provided in the Supplementary Note.

De novo mutations. We only considered the de novo mutations occurring in the accessible regions of the 1000 Genomes Project. For each motif class, we found the expected number of mutations under a normalized 1-mer sequence context model. More details are provided in the Supplementary Note.
Calculating tolerance scores for genes. We find the expected distribution of polymorphism levels for each gene by performing simulations from the standard multinomial distribution using our coding substitution probability estimates. We then normalize the difference between the observed levels of polymorphism and those generated from simulations, to obtain the gene tolerance score defined as

\[
\frac{\left( \mu_{NS} - n_{NS} \right)}{\sigma_{NS}}
\]

(8)

where \(\mu_{NS}\) and \(\sigma_{NS}\) represent the mean and standard deviation of nonsynonymous polymorphisms generated from simulations based on our model and \(n_{NS}\) is the empirical number of nonsynonymous polymorphism observed in the data. A positive gene score indicates that the number of observed substitutions is fewer than expected and identifies genes experiencing stronger than average purifying selection.

Categorizing genes on the basis of tolerance scores. We subdivided genes into various categories, that is, essential genes (where knockout of the mouse homolog is lethal), ubiquitously expressed genes, genes with known phenotypes described in OMIM, immune-related genes, keratin genes and olfactory genes. The data set from ref. 33 was used to find the first two categories, and ref. 32 was used to classify genes in OMIM. OMIM subcategorizes genes according to mutational models, including de novo, dominant, haploinsufficient or recessive. In our analysis, we merged OMIM’s de novo, dominant and haploinsufficient categories, treating them as a single category. We used the DAVID ontology database (version 6.7) to classify immune-related, keratin and olfactory genes. The data set from ref. 38–40 was used to find the first two categories, and ref. 32 was used to classify genes in OMIM. OMIM subcategorizes genes according to mutational models, including de novo, dominant, haploinsufficient or recessive. In our analysis, we merged OMIM’s de novo, dominant and haploinsufficient categories, treating them as a single category. We considered the gene lists published in the latest de novo sequencing analysis reports on autism\(^{35}\), epilepsy\(^{36}\), intellectual disability\(^{38–40}\) and developmental disorder\(^{37}\) as the gene sets belonging to these diseases. We merged the gene lists of the aforementioned diseases, treating them as a single category belonging to ‘all neuropsychiatric disease’.

AUC comparison between competing gene scores on different gene sets. We used the receiver operating characteristic (ROC) curve to compare the performance of our gene scores against previously annotated scores for classifying genes into the gene sets we described above. We fitted a linear classifier using the three different gene scores on each gene set and found the area under the curve (AUC) for each. More details are provided in the Supplementary Note.

Calculating tolerance scores for amino acids. We find the expected distribution of polymorphism levels for a specific amino acid encoded by a gene by performing simulations from the standard multinomial distribution using our coding substitution probability estimates. For a given gene, we then normalize the difference between the observed numbers of changes at a specific amino acid versus the number of changes expected from simulation using the equation

\[
\frac{\left( \mu_{AA} - n_{AA} \right)}{\sigma_{AA}}
\]

(9)

where \(\mu_{AA}\) and \(\sigma_{AA}\) represent the mean and standard deviation of the specific amino acid replacement polymorphisms generated from simulations based on our model and \(n_{AA}\) is the empirical number of amino acid replacement polymorphisms observed in the data. We consider the normalized value in equation (9) as the final tolerance score for that amino acid within the given gene. We interpret a positive amino acid tolerance score to indicate that the observed number of changes for that specific amino acid for the given gene was even fewer than expected. Thus, the amino acid tolerance score serves to identify amino acids experiencing stronger than average purifying selection.

Sourcing information about pathogenic variants. We used the HGMD (professional 2014.4) to identify pathogenic variants for our autosomal genes of interest, which supplied 60,504 variants distributed over 3,647 genes for 5,359 putative human disorders.

Application of gene and amino acid score on autism spectrum de novo sequencing data. We used the de novo sequencing data for autism spectrum disorder\(^{42}\) to test the efficacy of our gene and amino acid score approach in identifying and prioritizing genes and variants newly associated with autism. We found the de novo mutations belonging to cases and controls separately for each of our genic sequences of interest and considered a total of 2,171 mutations in 2,508 cases and 1,421 mutations in 1,911 controls. For a uniform comparison of gene scores across different approaches\(^{12,32}\), we only considered the top 752 intolerant genes identified from each approach. We chose 752 genes because this was the number of intolerant genes identified in ref. 12, which mapped to our autosomal genic sequences of interest (which pass the stringent criteria of sequencing quality in the 1000 Genomes Project). We used the odds ratio to find the burden of de novo mutations in cases as opposed to controls in the set of intolerant genes. Fisher’s exact test was used to compare the significance of burden. For amino acid score, all statistical comparisons were performed using the Wilcoxon rank-sum test. More details are provided in the Supplementary Note.

Code availability. The computational pipelines used for probability estimation for the noncoding and coding genomes and for forward regression and feature selection are available upon request.