SUPPLEMENTARY INFORMATION

Table of Contents

Building and testing substitution probability framework in non-coding intergenic region

Data Access

Sourcing population samples

Samples were obtained from phase 1 of the 1000 Genomes Project. Further details about sample collection, sequencing, and variant calling are available in the original publication[1]. We considered only the variants from African (n= 246 individuals), European (n = 379), and East Asian (n = 286) ancestries.

Selection of intergenic non-coding sequences

Intergenic sequences were defined as the full set of genomic sequences that are not annotated in ENSEMBL Biomart[2] (Ensembl Genes 75 and Homo sapiens genes GRCh37.p13) and RefSeq Genes[3]. We initially removed centromeric, telomeric, and repetitive regions from these non-coding sequences by filtering out the contiguous sequences at the ends of the chromosomes and “gene deserts” of length greater than 2 MB. We also filtered away the sequences that were not present in the combined accessibility mask (version 20120824) of the 1000 genomes project. As a result, we were left with ~1100 Mb of autosomal intergenic regions and ~90 Mb on the X chromosome. Within these intergenic regions, we found 10,809,273 variants in the African populations, 7,051,667 variants in the European populations, and 6,024,240 variants in the East Asian populations.

Selection of HapMap variants

Single nucleotide polymorphic variants were obtained from 2010-8 phase 3 release of the HapMap project. We considered the variants from African ancestry only, belonging to populations YRI (Yoruba), LWK (Luhya), MKK (Maasai). We also filtered for variants occurring in our intergenic non-coding sequences, resulting in a total of 1,659,929 variants.

Basis of substitution probability framework

Statistical framework to model substitution probabilities

To explain our approach for modeling nucleotide substitution probabilities observed in a given population, we will first describe a simple model that does not take into account local sequence context, then build upon this simple framework by incorporating additional features to model nucleotide substitution probabilities in a way that considers the impact of local sequence contexts of varying lengths. 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 nucleotide change C-to-A has occurred. Similarly, n_{CG} is the number of sites where a change C-to-G has occurred and n_{CT} is the number of sites where a change C-to-T has occurred. Then the probability of nucleotide substitution or polymorphism within the population genome-wide can be described at a given genomic site using a multinomial distribution:

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

where the probabilities of observing a substitution from C-to-A, C-to-G, and C-to-T are expressed as $\alpha_{CA,}$ $\alpha_{CG,}$ and $\alpha_{CT,}$ respectively. After iterating over all possible substitutions (*i.e.*, 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, G-to-T), we merged the reverse-complementary pairs (*e.g.*, A-to-C was merged with T-to-G, etc.) to yield 6 "substitution classes" as parameters for the simple model, which we refer to as the "1-mer model". This model can be naturally extended to consider the effects of local sequence context by replacing the count of $\mathrm{n_x}$ occurrences of nucleotide *X* with the count of occurrences of a particular nucleotide sequence context. For example, if we want to consider the local sequence context ACA, then we count the number times n_{ACA} that this 3-mer sequence occurs in the reference genome. A subset of n_{ACA} will be polymorphic at the middle position C within a given population. Thus, let $n_{ACA \rightarrow AAA}$ represent the number of sites where a nucleotide change C-to-A has occurred at the middle position, $n_{ACA \rightarrow AGA}$ represent the number of sites where a nucleotide change C-to-G has occurred at the middle position, and $n_{ACA \rightarrow ATA}$ represent the number of sites where a nucleotide change C-to-T has occurred at the middle position. All of these combinations represent a 3-mer sequence context in which the polymorphic middle position is flanked by fixed nucleotides A on both sides. After merging reverse complementary sequences, there are 16 unique sequence contexts (e.g. four possibilities (A, C, G, or T) for the single fixed nucleotide located 5′ of the polymorphic site, and four possibilities for the single fixed nucleotide located 3′ of the polymorphic site) per substitution class. Across all six substitution classes, there are a total of 96 parameters estimated under this “3-mer model”. We analogously extend the size of the sequence context window to evaluate the “5-mer model” and the “7-mer model” by considering additional fixed nucleotides (2 and 3, respectively) on either side of the polymorphic site, thereby estimating a total of 1536 parameters for the 5-mer model and 24,576 parameters for the 7-mer model. For sake of comparison, we also considered a very simplistic null model that completely ignores sequence context and merges substitution classes into a single group, such that Equation 1 simplifies to a binomial distribution with a single estimated parameter.

Incorporating prior information into the nucleotide context models

We may have some existing “prior” beliefs regarding probabilities of nucleotide substitution that can be incorporated into our framework using Bayesian statistics. For example, rates of nucleotide substitution in the coding genome should be proportional to, but not exactly the same as, the rates that are observed in the non-coding genome. This prior information can be incorporated into our model as follows. Because the likelihood of our framework is based on a multinomial distribution, we utilize its conjugate prior, *i.e.*, the dirichlet distribution, for models that incorporate sequence context. For the null model, we can analogously utilize its conjugate prior, *i.e.*, the beta distribution. For inference in the intergenic, non-coding genome, we selected the objective version of the prior for our analysis, with all concentration parameters (or shape parameters for the analogous beta prior) of the dirichlet prior as 1.

Testing framework

Log-likelihood ratio testing for model comparison

To evaluate how increasing the length of the context sequence affects our competing models' fit to empirical data, we utilized a log-likelihood ratio testing procedure. First, the likelihood of the observed distribution of polymorphic sites given a specific sequence context model (null, 1-mer, 3-mer, 5-mer, or 7-mer) was calculated using the substitution rate parameters estimated using all of the data. We calculate the likelihood ratio test statistic as:

$$-2\ln(L[Data|Context\ S_1) + 2\ln(L[Data|Context\ S_2) \qquad (2)$$

where S_1 and S_2 represent parameters estimate from two competing sequence context models. The test is chi-squared distributed, with degrees of freedom equal to the difference in the number of parameters between the two models (*e.g.*, comparing the 1-mer model versus the null model requires 5 degrees of freedom; comparing the 7-mer model versus the 3-mer model requires 24,480 degrees of freedom). Reported P-values are approximated analytically from the appropriate chi-square distribution using the R package (version 3.0.3).

Bayes Factor analysis for model comparison

We utilized the Bayes Factor approach, the Bayesian alternative to likelihood ratio testing, to contrast competing sequence context models against each other. We calculated the approximate posterior likelihood, using the Chib's method, on the overall data using the maximum *a posteriori* (MAP) estimates of the substitution probabilities for a specific sequence context model (null, 1-mer, 3-mer, 5-mer, or 7-mer) found before. We then calculate the approximate Bayes factor as:

$$\frac{Posterior\ likelihood\ under\ Model_2}{Posterior\ likelihood\ under\ Model_1} = \frac{Prob(Data|Context\ S_2) \times Prob(Context\ S_2)}{Prob(Data|Context\ S_1) \times Prob(Context\ S_1)} \qquad (3)$$

where S_1 and S_2 represent parameters estimate from two competing sequence context models. Since we use flat objective priors in the noncoding region and the MAP and MLE estimates are similar, the approximate Bayes factor reduces to the ratio of likelihood estimates under the two models. We use the Jefferey's scale for interpreting the approximate Bayes Factors, where the ratio if greater than 100 is considered to be decisive evidence against the Model$_1$.

Regression modeling and feature selection

We hypothesized that, within a substitution class (described above), the probability of polymorphism could be predicted using a linear combination of features based on the nucleotides at flanking positions within the 7-mer context. For the analysis below, we considered the posterior probabilities generated using data from the African group (1KG). We considered each substitution class separately and created an additional substitution class for each of the three possible changes within a CpG context (*i.e.*, where the polymorphic 4th position nucleotide may change C-to-A, C-to-G, or C-to-T, but the 5th position in the 7-mer context sequence is fixed as nucleotide G), resulting in nine substitution classes that are taken into regression modeling. For each substitution class, we considered the initial regression model:

$$Pr[X_1 \rightarrow X_2|S] = \alpha + \beta_1 p_1^C + \beta_2 p_1^G + \beta_3 p_1^T + \cdots + \beta_n p_7^T + \varepsilon \quad (4)$$

where the probability that a nucleotide changes from X_1 to X_2 is modeled using a position-base variable *p*, a set of bases (*e.g.*, {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 α, and error term ε. 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 (*e.g.*, 0,0,0 = A; 0,0,1 = C; 0,1,0 = G; 1,0,0 = T). Position 5 is fixed as G for substitution classes within a CpG context, enabling us to remove position 5 terms from those models. Similarly, models of non-CpG classes considered only C and T bases at position 5. Next, we examined non-additivity (*i.e.*, interactions) between nucleotides at sequence context positions. Rather than including all possible interaction terms, we employed feature selection (*i.e.*, model training and testing to select the most informative features) and incorporated these terms into the final model. We considered 2-way, 3-way, and 4-way interactions across positions within the 7-mer as:

$$Pr[X_1 \rightarrow X_2|S] = \alpha + \beta_1 p_1^C + \beta_2 p_1^G + \beta_3 p_1^T + \cdots + \beta_n p_7^T +$$
$$\beta_a p_i^w \times p_j^x + \cdots + \beta_b p_i^w \times p_j^x \times p_k^y + \cdots + \beta_c p_i^w \times p_j^x \times p_k^y \times p_l^z + \cdots + \varepsilon \quad (5)$$

where the probability that a nucleotide changes from X_1 to X_2 is modeled as described in Equation 4, and a set of additional terms related to interactions is also incorporated. Interaction terms are obtained from the product of thermometer variables *p* for bases *w*, *x*, *y*, or *z* (*e.g.*, {C, G, or T} where A is the reference base) at positions *i*, *j*, *k*, or *l* (= 1, 2, 3, 5, 6, or 7). The effect of the interaction is represented by terms β_a for 2-way interactions, β_b for 3-way interactions, and β_c for 4-way interactions. We only considered interaction terms that involved nucleotides located at different positions within the sequence context (*i.e.*, *i* not equal to *j*, *j* not equal to *k*, and *k* not equal to *l*). We divided the genome into two distinct sets for feature selection, using all even-numbered chromosomes for training and all odd-numbered chromosomes for model testing. During training, we performed stepwise forward regression for each level of interaction in order of increasing complexity (*i.e.*, first 2-way, then 3-way, and finally 4-way). For each level of interaction, we further trained the model by sequentially incorporating interaction terms, one at a time, and evaluating whether each term improved the model using the ANOVA F-test. The most informative interaction term was added to the model at each step. We repeated this process until no additional features further improved the model (*i.e.*, all proposed features were $P > 0.001$ by the F-test). For higher-order (3-way and 4-way) interactions, we ensured that a proposed feature maintained the hierarchy constraint (*i.e.*, a selected 4-way term must bring with it all of its associated 3-way and 2-way terms). As a result of this constraint, when considering higher-order terms, we simultaneously considered any associated lower-order terms that had not been selected during prior lower-order training, thereby adding degrees of freedom to our F-test assessment. As our final model, we selected the trained model with the lowest mean-squared error, calculated via 8-fold cross-validation within each substitution class. We report Akaike Information Criteria and adjusted-R^2 values for the final model using the testing data set. Regression analysis was performed using R (version 3.0.3) using lm() for regression modeling, and the packages: leaps (v2.9), DAAG (1.20), lattice (v0.20-29), grid (v3.0.3), latticeExtra (v0.6-26), and RColorBrewer (v1.0.5).

Sourcing CpG methylation data

We obtained CpG methylation data for our intergenic regions of interest from a published whole genome bisulphite sequencing study performed on germline (sperm, oocyte)[4], blastocyst[4], blood[4] and brain[5] tissues. For each tissue, we divided the CpG sites into three bins: (i) sites that were methylated in all samples, (ii) sites that were methylated in some but unmethylated in other samples and (iii) sites that were unmethylated in all samples. Very few sites fell into the second bin, so we excluded sites where methylation signal was inconsistent among the samples. 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 in all sperm samples. The same procedure was followed for samples from other tissues. We summarized the methylation signal across all samples for a tissue by calculating the mean intensity.

Sequence Motif Identification

We examined the top and bottom 10 sequences for each substitution class, and manually identified a total of 6 motifs that we tested in each substitution class, stratified by CpG context. This results in a total of (9 substitution classes) * (2 tails, high and low) * (6 motifs) = 108 total tests. Note that we required a nominal P = 4.6 x 10^{-4} (Bonferroni correction for multiple testing). We used Fisher's exact test to find the P-value associated with the enrichment of specific sequence motif using the fisher.test function in the R package (version 3.0.3). The contingency tables for the test were populated by considering the enrichment of sequence motifs in the top or bottom 1% of substitution probabilities for that specific class of change. We report in Table 1 those sequence motifs for each category of substitution that pass a Bonferroni corrected threshold.

Sourcing recombination data

We obtained recombination rate map of the YRI population from the phase 1 release of the 1000 Genomes project

(ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates/YRI_o mni_recombination_20130507.tar), and segregated our intergenic non-coding regions of interest into high (recombination rate >3 cM/Mb) and low recombination rate (rate < 0.05 cM/Mb) regions. As a result we considered ~203 Mb of intergenic non-coding sequence as belonging to high recombination rate region and ~494 Mb of intergenic non-coding sequence as belonging to low recombination rate region.

Human and primate divergence

We obtained human-chimpanzee and human-macaque chain and netted alignments from the golden path directories in the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsPanTro4/axtNet/, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/vsRheMac3/axtNet/) and found divergence between the human-primate pair by calculating fixed differences between the aligned intergenic non-coding sequences at each 7-mer sequence context. We were able to align 1.06 Gb of intergenic non-coding sequences between human-chimpanzee and 0.88 Gb between human-macaque. For each 7-mer sequence context, we calculated the divergence as the ratio of total number of fixed differences between the human-primate pair, and the total number of sequence context occurrences in the aligned region.

Variants across the frequency spectrum

We defined the rare variants as those occurring only once or twice in the population, and low or high frequency variants as those with MAF greater than 1%. We only considered the variants present in 1000 genomes project belonging to the African ancestry and occurring in the intergenic non-coding sequences, and found 2,789,383 rare and 8,019,893 low/high frequency variants.

De novo mutations

We only considered the *de novo* mutations from the high quality pedigree sequencing dataset of DECODE Genetics[6], that occurred in the accessible regions of the 1000 genomes project. This filtering was necessary because the original study did not describe the genome-wide regions that were "sequenceable". We make an implicit assumption that atleast the accessible regions in the 1000 genomes project were sequenced in the original high quality pedigree sequencing study. We then found the observed *de* novo mutations for each motif class. The expected number of mutations occurring in each class was simulated under a normalized 1-mer sequence context model, such that the overall *de novo* mutation rate was fixed at 1.2 x 10^{-8} *de novo* mutations per generation and per sample.

Building and testing substitution probability framework in the coding region

Extension of the statistical framework for coding regions

To model substitution probabilities for the coding genome, we utilized the statistical model developed for intergenic regions with the following modifications: First, we accounted for codon position-effects (*i.e.*, a given sequence context around a polymorphic site may occur at three different positions on a codon), which can lead to amino acid changes that may be subject to different levels of selective constraint. To model this phenomenon, we considered the probabilities for each of the three possible codon positions separately, resulting in a total of 73,728 (3 * 24,576) parameters for the 7-mer context model. Second, we utilized probabilities learned from the intergenic non-coding region model as our Bayesian prior for the coding model. The parameters for this prior include the baseline probabilities from the intergenic noncoding region as shape parameters for the dirichlet distribution, multiplied by an additional normalizing weighted constant, per the following:

$$\left(p_{S_1\to S_2} * \frac{10}{1+e^{-n}}\right), \qquad \left(p_{S_1\to S_3} * \frac{10}{1+e^{-n}}\right), \qquad \left(p_{S_1\to S_4} * \frac{10}{1+e^{-n}}\right),$$

$$\left(\frac{10}{1+e^{-n}} - \left(p_{S_1\to S_2} * \frac{10}{1+e^{-n}}\right) - \left(p_{S_1\to S_3} * \frac{10}{1+e^{-n}}\right) - \left(p_{S_1\to S_4} * \frac{10}{1+e^{-n}}\right)\right) \tag{6}$$

where *p* represents the intergenic noncoding substitution probability from sequence context *S* to each possible polymorphic change (1, 2, 3, or 4 represent each possible nucleotide base at the site), *n* is the number of occurrence of the context *S* in the coding region. This choice of shape parameter in the prior allowed for inference of coding substitution probabilities, while utilizing the intergenic substitution probabilities, and without the prior overwhelming the evidence observed in the coding region.

Data Access

Sourcing of coding sequences

We selected exonic coordinates of the longest transcript for each gene annotated in ENSEMBL Biomart (Ensembl Genes 75 and Homo sapiens genes GRCh37.p13). We only considered those transcripts where

(i) the total exonic region length was a multiple of 3, and (ii) 90% or larger of it was present in the combined accessibility mask (version 20120824) filter of the 1000 Genomes project. For all genes of interest, we used phase information to map each genomic coordinate to a specific position on a codon, yielding 16,386 autosomal transcripts and 679 transcripts from the X chromosome.

To test our model in a different data set, SNP sites for ~4300 individuals of European ancestry were obtained from large-scale independent exome sequencing studies generated by the NHLBI GO Exome Sequencing Project, from the Exome Variant Server (EVS[7], http://evs.gs.washington.edu/EVS/, downloaded on August 26th 2013).

Annotation of SNP variants in the autosomal coding genome

For 1KG data, we manually annotated the type of codon change caused by each variant, yielding 92,893 synonymous, 110,645 missense, and 1,639 nonsense variants (total n = 205,282) for the African group. We repeated the same strategy for the non-Africans, resulting 64,756 synonymous, 89,863 missense, and 1,591 nonsense variants (total n = 156,298) within the European group and 58,304 synonymous, 80,689 missense, and 1,378 nonsense variants (total n = 140,450) within the Asian group. For the EVS data (European ancestry), we also manually annotated the type of codon change, yielding a total of 226,833 synonymous, 388,149 missense, and 15,287 nonsense variants (total n = 636,122) distributed over the coding regions of interest. To obtain a representative spectrum of allele frequencies (and impact of background selection) observed from the smaller set of individuals found in the 1KG data, we considered only EVS variants with frequency greater than 0.03% resulting in a total of 169,659 variants.

Sourcing information about pathogenic variants

We used the Human Gene Mutation Database[8] (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.

Testing framework

Scaling the model for use with a larger data sample

As the number of individuals sequenced increases, the observed number of polymorphic sites segregating within the dataset will also increase. To calibrate our model (built using the 1KG dataset) for use with the larger EVS dataset, we rescaled the substitution probabilities estimated using 1KG data to make them proportional to the EVS dataset. We used a constant scaling factor defined as:

$$\frac{Overall\ Substitution\ probability\ in\ the\ new\ dataset}{Overall\ substitution\ probability\ in\ the\ 1000\ genomes\ dataset} \tag{7}$$

on all substitution probabilities in the new dataset.

Bayes Factor analysis for model comparison in the coding region

We utilized the Bayes Factor approach, the Bayesian alternative to likelihood ratio testing, to contrast competing coding sequence context models against each other. We compared the 7-mer model with codon position effects and priors from noncoding region as described before, against the basic 3-mer model with no codon position effects and with a flat objective prior. The approximate posterior likelihood, using the Chib's method, on the overall coding data was then calculated using the maximum *a posteriori* (MAP) estimates of the substitution probabilities for the two coding sequence context models as found before. We then calculate the approximate Bayes factor using Equation 3, above. For the 7-mer model the probability of parameters is found using the dirichlet distribution function in the gtools (v3.4.1) package in R (v3.0.3). Since we use flat objective priors in for the 3-mer model so the probability of parameters reduces to calculating the normalizing beta function in the dirichlet distributions. We use the Jefferey's scale for interpreting the approximate Bayes Factors, where the ratio if greater than 100 is considered to be decisive evidence against the Model$_1$.

Simulating variability in substitution probabilities within all types of amino acid replacement

To simulate the distribution in variability for substitution probabilities within different amino acid substitution type, we randomly distributed the number of observed substitutions within the type using a fixed rate model. We then calculate the respective 7-mer probabilities using our multinomial distribution model for the randomization, and use those to tabulate the variance across different amino acid substitution types. 10^6 simulations are used to generate the distribution of substitution probabilities.

Measuring the effects of selection on the probability of polymorphism

To minimize the effects of selection on initial estimates of substitution probabilities, we selected intergenic non-coding intervals for model development. Assuming that the mechanisms that introduce new mutations into coding regions are similar to those at work in the non-coding genome, we inferred that the relative ratio of coding-to-non-coding substitution probabilities could indicate natural selection occurring in the coding genome. Furthermore, we expected that the rates of certain types of amino acid change should be less frequent than others (*e.g.*, on average, we expect to observe non-synonymous changes less frequently than synonymous changes) as a result of background selection. To quantify the effect of selection on substitution probabilities, we measured the $\log_{10}$ ratio of coding-to-non-coding substitution probabilities using all coding variants (n = 205,282) observed in the 1KG African group. Estimates for coding substitution probabilities were uncertain under certain conditions, owing to a limited number of a given variant type for a particular 7-mer context. Thus, rather than use our MAP estimates for these sequences contexts, we simulated the substitution probabilities from the beta distribution using a 3-mer context model extended to the coding region. We then calculated the log-ratio of the intergenic non-coding substitution probability to the mean obtained from simulation.

Gene Scores

Calculating tolerance scores for genes

Using our estimates for substitution probabilities in the coding genome, we performed simulations using the standard multinomial distribution for each sequence context to define the distribution of polymorphism levels expected for each gene based on our model. We then normalized the difference between the observed levels of polymorphism and those generated from our simulations, to obtain gene tolerance score defined as:

$$\frac{(\mu_{NS} - n_{NS})}{\sigma_{NS}} \tag{8}$$

where μ_{NS} and σ_{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 in Equation 8 indicates that the number of observed substitutions is fewer than expected, and serves to identify genes experiencing stronger than average purifying selection. In our analysis, we determined gene scores for the African, European, and EVS populations.

Categorizing genes

We subdivided genes into various categories – *i.e.*, essential genes (where the mouse homolog knock-out is lethal), ubiquitously expressed genes, genes with known phenotypes described in OMIM, immune-related genes, keratin genes, olfactory genes and those belonging to several neuropsychiatric diseases. The dataset from[9] was used to find the first two categories, while[10] was used to classify OMIM genes. OMIM sub-categorizes 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[11] (version 6.7) to classify immune-related, keratin, and olfactory genes. We considered the gene list published in the latest *de novo* sequencing analysis papers of Autism[12], Epilepsy[13], Intellectual disability[14–16] and Developmental disorder[17], as the gene set belonging to these diseases. We merged the gene lists of the aforementioned diseases, treating them as 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[10,18] for classifying genes into the gene sets we described above. Since, the Petrovski et. al. scores[10] were originally released for HGNC gene ids, we were only able to convert 16,910 genes out of a total of 16,957 to corresponding ids in ENSEMBL format. Similarly the Samocha et. al. approach[18], only identified 1,003 genes to be intolerant and released their scores for Refseq gene ids, so we were able to map 997 genes only to corresponding ids in ENSEMBL format. Moreover, for a uniform comparison between different approaches, we only considered the previously annotated scores for autosomal genes that we identified before (i.e., which passed the stringent quality criteria of sequencing in the 1000 genomes project). 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. The linear model was fitted using the glm function (with binomial family parameter) in R (v3.0.3). The performance of the models on different gene sets was evaluated using the pred and performance functions (with auc as a parameter) using the ROCR (v1.0-5) package.

Amino Acid Scores

Calculating tolerance scores for amino acids

Using our estimates for substitution probabilities in the coding genome, we performed simulations using the standard multinomial distribution for each sequence context to determine the expected number of changes for a specific amino acid within a given gene. Within a given gene, we normalized the difference between the observed numbers of amino acid changes at a specific codon versus the number of changes expected from simulation using the equation:

$$\frac{(\mu_{AA} - n_{AA})}{\sigma_{AA}} \qquad (9)$$

where μ_{AA} and σ_{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 (AA) tolerance score to indicate that the observed number of changes for that specific amino acid within the given gene was *even fewer* than expected. Thus, the AA tolerance score serves to identify amino acids experiencing stronger than average purifying selection. Moreover, since the AA scores measure the tolerance of a gene at an amino acid level, they further improve the resolution of the gene scores, which measure the overall tolerance in a gene. In our analysis, we determined AA tolerance scores for the African population.

Application of Gene and Amino acid scores on Autism spectrum disorder *de novo* sequencing data

We used the *de novo* sequencing data for Autism spectrum disorder[19], to test the efficacy of our gene and amino acid score approach in identifying and prioritizing novel genes and variants associated with Autism. We found the *de novo* mutations belonging to cases and controls separately for each of our genic sequences of interest and further classified them into synonymous, missense, nonsense, splice and indel categories only. As a result, we considered a total of 2,171 mutations in 2,508 cases and 1,421 mutations in 1,911 controls, belonging to our genic sequences of interest.

For a uniform comparison of gene scores across different approaches[10,18], we only considered the top 752 intolerant genes identified from each approach. We choose 752 genes because this was the number of intolerant genes identified in[18], which mapped to our autosomal genic sequences of interest (i.e., 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.

The amino acid scores were found on known Autism genes identified in the latest *de novo* sequencing paper[12], and compared with (a) all mutations in controls or with (b) all mutations in cases belonging to non-Autism genes. All statistical comparisons were performed using the Wilcoxon sum ranked test. Similar analysis was also performed on genes with a higher burden of functional (missense, nonsense changes for which amino acid scores are generated) *de novo* mutations in cases as opposed to controls.

Supplementary Table Legends

Supplementary Table 1: Summary statistics evaluating proposed framework and model for sequence context. (**A**) Pearson's correlation and Root Mean-Squared Error (RMSE) for substitution probabilities estimated from the training (all but the two listed chromosomes) and testing (the two listed chromosomes) sets from the intergenic non-coding genome. We present measurement for null (i.e., fixed rate), 1-mer, 3-mer, 5-mer and 7-mer models. (**B**) P-values for the each likelihood ratio test comparing competing sequence context models (null, 1-mer, 3-mer, 5-mer and 7-mer), using all data from the intergenic non-coding genome. The matrix is symmetric, so "-" is presented where appropriate. (**C**) Natural logarithm of the approximate Bayes Factor comparing competing sequence context models (null, 1-mer, 3-mer, 5-mer and 7-mer), using all data from the intergenic non-coding genome. The matrix is symmetric, so “-” is presented where appropriate.

Supplementary Table 2: R^2 and correlation between the substitution probabilities estimated using HapMap and 1000 Genomes variant data from the intergenic non-coding genome, for different sequence context models (3-mer model with randomized sequence context beyond adjacent nucleotides, 7-mer model). Also shown is the comparison specific for CpG and nonCpG sequence contexts.

Supplementary Table 3: P-values and Bayes factor comparing sequence context models (1-mer, 3-mer, 5-mer and 7-mer) using all HapMap variant data from the intergenic non-coding genome.

Supplementary Table 4: Average nucleotide substitution probabilities for different population groups (African, European, and Asian) on different types of regions (coding versus intergenic non-coding) and on different chromosomes (All autosomal versus X chromosome).

Supplementary Table 5: Stepwise regression model analysis for each substitution class various models consider for in the intergenic non-coding region. Data for the training phase was based on the collection even numbered chromosomes; data for the testing phase was based on odd numbered chromosomes. “# Features” denotes the features selected for that model. “AIC” is the Akaike Information Criterion, “MSE” represents Mean-squared Error; “adj-R2” is the adjusted R^2 from the model. The best performing model (lowest MSE after 8-fold cross validation) are highlighted in red, and reflect the models presented in Table 1.

Supplementary Table 6: Aggregated sequence context features and their effect on the substitution probabilities for all classes of substitutions in the intergenic non-coding region. Order denotes the number of interacting nucleotides in the context. “BETA” indicates the regression coefficient for the sequence context for the given substitution class (i.e., A-to-C, A-to-G, etc.). “All_DIRXN” denotes the direction of effect for the feature on the substitution probability (+ indicates increase higher substitution probability, – indicates lower substitution probability). We present estimated values using (I) all data from 1KG, (II) data used in the training phase (all even-numbered chromosomes), and (III) data used for the testing phase (all odd-numbered chromosomes). (**A**) Substitution classes for sequence contexts outside of CpG sites (**B**) Substitution classes for sequences context including CpG sites (polymorphic 4th position is C and 5th position fixed at G).

Supplementary Table 7: Posterior probabilities of nucleotide substitution for all substitution classes within all 7-mer sequence contexts in the intergenic non-coding region for African, European and Asian populations groups (1KG project). The forward and reverse complementary sequences are presented for each probability.

Supplementary Table 8: Enrichment of motifs identified in nucleotide substitution probabilities inferred from HapMap variant data in the intergenic non-coding genome. CpG+ indicates the distribution of sequence contexts which include a CpG site (4th position polymorphic site is C, 5th position fixed as G). Enrichment P-value is based 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, “Lower” implies enrichment in the lower 1% tail. Odds ratio and [95% CI] denotes the odds ratio (and 95% confidence interval) of enrichment of motif in the upper or lower 1% tail of the sequence context probability distribution. Fold change in substitution rate denotes the fold increase or decrease in substitution rates for the motif relative to its substitution class.

Supplementary Table 9: R^2 and correlation between the substitution probabilities in intergenic non-coding genome and human primate divergence at a context, for different sequence models (3-mer model with randomized sequence context beyond adjacent nucleotides, 7-mer model). Also shown is the comparison specific for CpG and nonCpG sequence contexts.

Supplementary Table 10: P-values and Bayes factor comparing sequence context models (1-mer, 3-mer, 5-mer and 7-mer) using all data from the intergenic non-coding genome. (**A**) Sequence context rates inferred from low frequency (1% and above MAF) variants from the 1000 genomes project (**B**) Sequence context rates inferred from rare (singletons and doubletons) variants from the 1000 genomes project.

Supplementary Table 11: Expected (95% CI) and observed *de novo* mutations for each class of change calculated on high quality pedigree sequencing data from 78 trios[6]. If number of observed mutations fall in the expected confidence interval then we denote it "As expected" otherwise as "Higher than expected".

Supplementary Table 12: Natural logarithm of the approximate Bayes Factor comparing the posterior likelihoods of the 3-mer context model with and without accounting for codon context, and our proposed 7-mer context model which does include codon context on multiple data sets. We present data from the African and European groups (1KG) and an analysis of the EVS dataset (individuals of European ancestry) after filtering out variants with minor allele frequencies less than 0.03%.

Supplementary Table 13: Posterior probabilities of nucleotide substitution for all substitution classes within all 7-mer sequence contexts in coding region for African, European and Asian populations groups (1KG project). The forward and reverse complementary sequences are presented for each probability. The corresponding amino acid changes associated with each substitution class within the 7-mer sequence context, as well as their reverse complements, are also listed in the table.

Supplementary Table 14: Estimates of the variability in amino acid substitution probabilities. (**A**) Simulated and observed variance in nucleotide substitution probabilities grouped by type of amino acid replacement class. (**B**) Simulated and observed variance in nucleotide substitution probabilities, stratified for each possible types of amino acid replacement. Reported simulated values are based on 1,000,000 repetitions, based on a fixed rate model for each class of substitution.

Supplementary Table 15: Gene scores and annotations for >16,000 transcripts in humans. We annotate each gene using Ensembl, attached to a specific transcript identifier. Columns 3 through 14 refer to the annotation attached to set membership (Essential, Ubiquitous, Immune, Olfactory, Keratin, Omim de novo, dominant, and haploinsufficient). Details and citations describing how each gene set was identified are presented in the Methods. The last three columns are gene scores calculated by our approach (for the African population), and various published methods[10,18].

Supplementary Table 16: Prediction accuracy of gene tolerance scores to classify membership in various gene sets analyzed in our study. Area under the curve (AUC) calculations for gene scores of [10,18] and our 7-mer codon context gene scores.

Supplementary Table 17: Amino acid tolerance scores for >16,000 transcripts in humans. These scores quantify the number of excess substitutions for each type of amino acid change relative to expected, with larger scores indicating fewer substitutions (intolerance) for that specific amino acid. Scores were developed using 1KG project data using the African group.

References

1. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

2. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–55 (2014).

3. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756–63 (2014).

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

5. Hovestadt, V. *et al.* Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature* **510,** 537–41 (2014).

6. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–5 (2012).

7. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–20 (2013).

8. 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).

9. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* **9,** e1003484 (2013).

10. 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).

11. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).

12. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).

13. Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501,** 217–21 (2013).

14. Hamdan, F. F. *et al.* De Novo Mutations in Moderate or Severe Intellectual Disability. *PLoS Genet.* **10,** e1004772 (2014).

15. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* **380,** 1674–82 (2012).

16. De Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).

17. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519,** 223–8 (2015).

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

19. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).