Supplementary Note for Measuring Intolerance to Mutation in Human Genetics

Zachary L. Fuller*[1], Jeremy J. Berg[1], Hakhamanesh Mostafavi[1], Guy Sella[1,2,3], and Molly Przeworski[1,2,3]

[1] *Department of Biological Sciences, Columbia University*
[2] *Department of Systems Biology, Columbia University*
[3] *Program for Mathematical Genomics, Columbia University*

December 17, 2018

1 Simulating PTV Counts

1.1 Background

The basis of our forward simulations was developed in Simons *et al.* (2014)[1] and further modified in Amorim *et al.* (2017)[2] and Simons *et al.* (2018)[3]. A key modification here, is that instead of simulating the frequencies of deleterious alleles, we are interested in recording the distinct number of segregating sites in a population for a finite number of mutational opportunities M.

Our model assumes that each mutational opportunity is a biallelic site in a diploid individual. Following our notation from the main manuscript, we assume that there are two possible alleles at a site: wild-type (A) and deleterious (D). Here, protein truncating variants (PTVs) are modeled as the D allele. At each site, mutations from $A \rightarrow D$ arise at a rate u per-gamete per-generation and can only arise on a background currently free of other PTV mutations (as is highly likely). Generations are formed by Wright-Fisher sampling with selection, modeled by choosing parents for each generation according to their fitness. The fitnesses of individuals with genotypes AA, AD, and DD are 1, $1 - hs$, and $1 - s$, respectively, where s is the selection coefficient and h is the dominance coefficient. We assume no intragenic recombination.

1.2 Constant Population Size Model

We considered a human gene of typical length *i.e.,* 225 PTV mutational opportunities —the average number in the human genome. Mutations arise at rate $u = 1.5 \times 10^{-8}$ per mutational opportunity M. While this value of u is only approximate, it yields realistic numbers of PTVs; the qualitative conclusions are the same for other choices. We first simulated PTVs in a constant population of diploid individuals of size $N = 100,000$, reflective of the more recent time period relevant to the dynamics of deleterious mutations[2], and ran each simulation for $10N$ generations. The number of segregating PTVs are estimated from a sample size of diploid 33,370 individuals drawn at present, to match the number of non-Finnish Europeans (NFE) in ExAC[4].

1.3 Plausible Demographic Model

We simulated the dynamics of PTVs under a plausible model of changes in the effective population size of Europeans inferred by Schiffels & Durbin (2014)[5]. Again, we considered a human gene with M=225 and $u = 1.5 \times 10^{-8}$. Each simulation begins with a constant population size N of 14,448 (the ancestral size inferred by[5]) and a burn-in of $10N$ generations to create an equilibrium distribution of segregating sites. The first population size change occurs 55,490 generations ago. Following[3], the population size changes and the generations they occur at are determined by piecing together the multiple sequentially Markovian coalescent (MSMC) inferences from European (CEU) HapMap individuals[5] of four haplotypes for times corresponding to $<$ 170 Kya and two haplotypes for more ancient times

[*] Corresponding author: zlf2101@columbia.edu

> 170 Kya. At the last generation, corresponding to the present, the number of PTVs segregating in the population are estimated from a sample size matching the number of NFE individuals in ExAC[4].

2 Calculating pLI

Lek *et al.* consider that a gene can belong to one of three categories: null, recessive, and haploinsufficient ($c \in \{Null, Rec, HI\}$)[4]. For a gene i, the probability of being loss-of-function intolerant, pLI, is defined as:

$$pLI_i = \frac{p(Z_i = HI|\pi_{HI}, PTV_i)}{\sum_c p(Z_i = c|\pi_c, PTV_i)} \tag{1}$$

where Z is the unobserved class label, PTV_i represents the observed number of PTVs in gene i, and π_c represents the proportion of all genes that belong to category c. Using an expectation-maximization algorithm, Lek *et al.* find the maximum-likelihood estimate (MLE) for π_{HI} used in the calculation of pLI. Here, we used the MLE of π, *i.e.,* the final mixing weights of each category, obtained from ExAC ($\pi_{Null} = 0.208$, $\pi_{Rec} = 0.489$, and $\pi_{HI} = 0.304$)[4]. The probability of a gene ($p(Z_i = c|\pi_c, PTV_i)$) belonging to each category c is found as:

$$p(Z_i = c|\pi_c, PTV_i) = \frac{Pois(PTV_i|N\lambda_c)\pi_c}{\sum_c Pois(PTV_i|N\lambda_c)\pi_c} \tag{2}$$

where $Pois$ is the Poisson likelihood, N is the sample size, and λ_c is the expected amount of depletion of PTVs for a category c ($\lambda_{Null} = 1$, $\lambda_{Rec} = 0.463$, $\lambda_{HI} = 0.089$). Thus, $N\lambda_c$ is the expected number of PTVs in a gene for a category c. Lek *et al.* find the expected number of PTVs under neutrality ($N\lambda_{Null}$) in each gene using a method introduced by[6]. Here, in our simulations for a given gene we determined the expected number of PTVs under neutrality by averaging over 10^6 replicates with $h = 0$ and $s = 0$. Then, using this number as $N\lambda_{Null}$ we calculated pLI for any observed number of PTVs (*i.e., PTV_i*) generated in a given simulation replicate of that gene with various h and s parameter combinations. Since we used the true expected number of PTVs under neutrality, rather than an estimate (as is the case in practice[4]), we are somewhat under-estimating the variability in pLI scores.

3 Data Availability

The `C++` code and accompanying scripts used for analysis and visualization are available online.

References

[1] Y. B. Simons, M. C. Turchin, J. K. Pritchard, and G. Sella, The deleterious mutation load is insensitive to recent population history, *Nature Genetics* **46**, 220 (2014), ISSN 1061-4036.

[2] C. E. G. Amorim, Z. Gao, Z. Baker, J. F. Diesel, Y. B. Simons, I. S. Haque, J. Pickrell, and M. Przeworski, The population genetics of human disease: The case of recessive, lethal mutations, *PLOS Genetics* **13**, e1006915 (2017), ISSN 1553-7404.

[3] Y. B. Simons, K. Bullaughey, R. R. Hudson, and G. Sella, A population genetic interpretation of GWAS findings for human quantitative traits, *PLOS Biology* **16**, e2002985 (2018), ISSN 1545-7885.

[4] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, et al., Analysis of protein-coding genetic variation in 60,706 humans, *Nature* **536**, 285 (2016), ISSN 0028-0836.

[5] S. Schiffels and R. Durbin, Inferring human population size and separation history from multiple genome sequences, *Nature genetics* **46**, 919 (2014), ISSN 1061-4036.

[6] K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnström, S. Mallick, A. Kirby, et al., A framework for the interpretation of de novo mutation in human disease, *Nature Genetics* **46**, 944 (2014), ISSN 1061-4036.