

CLAIMS

What is claimed is:

1. A method for estimating a selection coefficient for a missense variant of a gene, the method comprising the steps of:
 - calculating a depletion metric (380) for the missense variant of the gene based on a pathogenicity percentile score of the missense variant and a relationship (436) between the percentile pathogenicity scores and depletion metrics for the gene; and
 - estimating the selection coefficient (390) of the missense variant based on the depletion metric (380) for the missense variant and a selection-depletion relationship derived for the gene.
2. The method of claim 1, further comprising:
 - verifying the depletion metric (380) is within the range bound by 0 and 1.
3. The method of claim 2, further comprising:
 - assigning the depletion metric (380) a value of 0 if less than 0 or a value of 1 if greater than 1.
4. The method of claim 1, further comprising the steps of:
 - determining, using a pathogenicity scoring neural network (102), a set of pathogenicity scores for possible missense variants within the gene;
 - deriving a corresponding percentile pathogenicity score (420) for each variant of the set;
 - binning (424) the percentile pathogenicity scores (420) into a plurality of bins;
 - calculating (430) a depletion metric (380) for each bin, wherein each depletion metric (380) quantitatively characterizes the proportion of missense mutations removed by selection for the corresponding bin; and
 - deriving (434) the relationship (436) between the percentile pathogenicity scores (420) and the depletion metrics (380).
5. The method of claim 4, wherein the pathogenicity scores of the set of pathogenicity scores are respectively generated based on an amino acid length sequence processed by one or

more of a neural network, statistical model, or machine learning technique trained or parameterized to generate the pathogenicity score from amino acid sequences.

6. The method of claim 5, wherein the neural network is trained using both human sequences and non-human sequences.

7. The method of claim 1, wherein the selection-depletion relationship derived for the gene comprises a selection-depletion curve (312).

8. The method of claim 1, wherein the selection-depletion relationship derived for the gene is determined using possible missense variants within the gene and a simulated allele frequency spectrum (304).

9. The method of claim 8, wherein the simulated allele frequency spectrum (304) is modeled for the gene under neutral selection over time by performing steps comprising:

simulating a forward time population model for the gene using model parameters comprising at least:

one or more growth rates (284) estimated for a population for an overall time simulated, wherein each growth rate (284) corresponds to a different sub-interval of time within the overall time simulated; and

one or more *de novo* mutation rates (280);

sampling a set of simulated chromosomes (294) for a target generation (290) simulated by the forward time population model for the gene; and

generating the simulated allele frequency spectrum (304) by averaging (292) across the set of simulated chromosomes (294).

10. The method of claim 9, wherein the model parameters are derived by performing steps comprising:

generating a plurality of simulated allele frequency spectra (304) using plurality of growth rates (284) and *de novo* mutation rates (280);

generating a synonymous allele frequency spectrum (308);

testing the fit of each simulated allele frequency spectrum (304) of the plurality to the synonymous allele frequency spectrum (308); and

based upon the fit of each simulated allele frequency spectrum (304) to the synonymous allele frequency spectrum (308), determining the model parameters.

11. The method of claim 9, wherein the *de novo* mutation rates (280) correspond to one or more of a genome-wide mutation rate; a transition mutation rate at CpG sites with high methylation, a transition mutation rate at CpG sites with low methylation, a transition mutation rate at non-CpG sites, or a transversion mutation rate.

12. The method of claim 1, wherein the selection-depletion relationship for the gene is derived by modeling frequency of variants for the gene under selection over time using a forward time simulation, the modeling being performed by performing steps comprising:

simulating a forward time population model for the gene using model parameters comprising at least:

one or more growth rates (284) estimated for the population for an overall time simulated, wherein each growth rate (284) corresponds to a different sub-interval of time within the overall time simulated;

one or more *de novo* mutation rates (280); and

a plurality of selection coefficients (320);

generating (306) at least one simulated allele frequency spectrum (304) for each selection coefficient (320); and

deriving the selection-depletion relationship for the gene, wherein depletion measures the proportion of variants removed by selection.

13. The method of claim 12, wherein depletion is derived based on the ratio of the number of variants with selection over the number of variants without selection.

14. A non-transitory computer-readable medium storing processor-executable instructions that, when executed by one or more processors, cause the one or more processors to perform steps comprising:

calculating a depletion metric (380) for a missense variant of a gene based on a pathogenicity percentile score of the missense variant and a relationship (436) between the percentile pathogenicity scores and depletion metrics for the gene; and

estimating a selection coefficient (390) of the missense variant based on the depletion metric (380) for the missense variant and a selection-depletion relationship derived for the gene.

15. The non-transitory computer-readable medium of claim 14, wherein the processor-executable instructions, when executed by one or more processors, cause the one or more processors to perform further steps comprising:

determining, using a pathogenicity scoring neural network (102), a set of pathogenicity scores for possible missense variants within the gene;

deriving a corresponding percentile pathogenicity score (420) for each variant of the set;

binning (424) the percentile pathogenicity scores (420) into a plurality of bins;

calculating (430) a depletion metric (380) for each bin, wherein each depletion metric (380) quantitatively characterizes the proportion of missense mutations removed by selection for the corresponding bin; and

deriving (434) the relationship (436) between the percentile pathogenicity scores (420) and the depletion metrics (380).

16. The non-transitory computer-readable medium of claim 14, wherein the selection-depletion relationship derived for the gene is determined using possible missense variants within the gene and a simulated allele frequency spectrum (304).

17. The non-transitory computer-readable medium of claim 16, wherein the processor-executable instructions, when executed by one or more processors, cause the one or more processors to perform further steps comprising:

simulating a forward time population model for the gene using model parameters comprising at least:

one or more growth rates (284) estimated for a population for an overall time simulated, wherein each growth rate (284) corresponds to a different sub-interval of time within the overall time simulated; and

one or more *de novo* mutation rates (280);

sampling a set of simulated chromosomes (294) for a target generation (290) simulated by the forward time population model for the gene; and

generating the simulated allele frequency spectrum (304) by averaging (292) across the set of simulated chromosomes (294).

18. The non-transitory computer-readable medium of claim 14, wherein the processor-executable instructions, when executed by one or more processors, cause the one or more processors to perform further steps comprising:

simulating a forward time population model for the gene using model parameters comprising at least:

one or more growth rates (284) estimated for the population for an overall time simulated, wherein each growth rate (284) corresponds to a different sub-interval of time within the overall time simulated;

one or more *de novo* mutation rates (280); and

a plurality of selection coefficients (320);

generating (306) at least one simulated allele frequency spectrum (304) for each selection coefficient (320); and

deriving the selection-depletion relationship for the gene, wherein depletion measures the proportion of variants removed by selection.

19. A method for determining a selection coefficient (390) associated with one or more mutations, the method comprising:

determining a number of observed loss-of-function (LOF) mutations (360) within an allele frequency dataset for the gene;

calculating (378) a depletion metric (380) using the number of observed LOF mutations (360) and an expected number of LOF mutations (364), wherein the depletion metric (380) characterizes the proportion of LOF mutations removed by selection; and

determining (388) a selection coefficient (390) for LOF mutations of the gene using the depletion metric (380).

20. The method of claim 19, wherein the depletion metric (380) is based on the ratio of the number of observed LOF mutations (360) over the number of expected LOF mutations (364).

21. The method of claim 19, wherein determining the selection coefficient (390) for LOF mutations comprises:

comparing the depletion metric (380) to a pre-determined relationship between selection and depletion for the gene to derive the selection coefficient (390).